Bill Shock: Inattention and Price-Posting Regulation

Michael D. Grubb†

May 31, 2011

Abstract

For many goods and services, such as cellular-phone service and debit-card transactions, the marginal price of the next unit of service depends on past usage. As a result, consumers who are inattentive to their past usage may be aware of contract terms and yet remain uncertain about the marginal price of the next unit. I develop a model of inattentive consumption, derive equilibrium pricing when consumers are inattentive, and evaluate price-posting regulation requiring firms to publish marginal price at the time of each transaction. Inattention leads firms to charge surprise penalty fees for high usage when consumers are heterogeneous ex ante or have biased beliefs. When consumers are homogeneous ex ante and have unbiased beliefs, inattention and penalty fees increase welfare in fairly-competitive markets, and price-posting regulation is socially harmful. Under these conditions, cellular-phone usage-alerts under consideration by the FCC could reduce welfare and harm consumers. If consumers are homogeneous ex ante but underestimate their demand, then price-posting regulation has an ambiguous impact on total welfare but may have large distributional benefits by increasing price competition and protecting consumers from exploitation. Hence the Federal Reserve’s new opt-in rule for debit-card overdraft protection could substantially benefit consumers.

---

*A previous version of this paper circulated under the title "Penalty Pricing: Optimal Price-Posting Regulation with Inattentive Consumers". I thank Glenn Ellison, Bob Gibbons, Alesandro Bonatti, Tavneet Suri, Ginger Jin, and Bob Pindyck for helpful comments and suggestions.

†MIT Sloan School of Management. mgrubb@mit.edu, www.mit.edu/~mgrubb.
1 Introduction

Marginal prices vary with usage for a wide variety of products and services including electricity, cellular-phone service, health insurance, and debit and credit-card transactions. Moreover, consumers are often uncertain about the marginal price at the point of sale because they are inattentive and do not keep track of past usage.\(^1\) For example, a cellular-phone customer may know that the first 500 minutes are free while later minutes are charged a penalty (or “overage”) rate of 45 cents a minute. However, he may be uncertain whether the next call will cost zero cents or 45 cents per minute, because he does not know how many minutes he has already used the phone. Similarly, a checking account holder may know that overdraft fees are $35 per transaction, but be unaware whether her next debit transaction will be free or incur a $35 penalty because she is uncertain about her checking balance.

In each example, firms have the ability to disclose whether a penalty fee is applicable at the point of sale. (Absent such disclosure I refer to the penalties as surprise penalty fees.) A mobile phone screen could flash “overage rate applies” before a call is made and a debit-card terminal could ask “Overdraft fee applies. Continue - Yes/No?” before processing transactions on an overdrawn account. That firms do not to make these disclosures and oppose regulation requiring such disclosure suggests that firms benefit from consumer uncertainty about marginal price (Altschul, Guttman-McCabe and Josef 2011). Recent regulation of overdraft fees by the Fed and consideration of “bill shock” regulation by the FCC suggest that regulators believe that the lack of transparency is bad for consumers and bad for welfare.\(^2\)

I develop a model to answer the following questions: First, if consumers are inattentive to their own past consumption, do firms profit by charging surprise penalty fees for high usage? Second, if so, does price-posting regulation requiring disclosure of marginal price at the point of sale benefit consumers more than it harms firms and thus increase welfare? Third, how do the conclusions depend on the level of competition between firms and the presence of consumer heterogeneity or consumer biases?

I find that firms profit by charging surprise penalty fees for high usage when price discriminating between low- and high-average demand market segments or when exploiting consumer bias

---

\(^1\)The same price uncertainty also arises without inattention when multiple family members consume from the same family-talk plan or joint checking-account but do not continually update each other about purchases. Shared usage will be an alternative interpretation to inattention throughout the paper.

\(^2\)FCC Chairman Julius Genachowski said, “something is clearly wrong with a system that makes it possible for consumers to run up big bills without knowing it,” and a variety of consumer advocacy groups agree (Genachowski 2010, Deloney, Sherry, Grant, Desai, Riley, Wood, Breyault, Gonzalez and Lennett 2011).
but not when consumers are vertically homogeneous and unbiased. The first of two main results is that price-posting regulation can lower social-welfare when firms use surprise penalty fees to price discriminate between unbiased consumers and that this is always true in fairly-competitive markets. The second main result is that, when consumers underestimate demand, price-posting regulation can eliminate consumer exploitation, increase consumer welfare by more than first-best social surplus, and do so even in fairly-competitive markets by forcing firms to compete on fees to which consumers are more price-sensitive.

I begin by modeling the consumption behavior of inattentive consumers. I assume that once an inattentive consumer signs a cellular-phone contract or opens a bank account, two consumption opportunities arise sequentially and each decision to make an additional phone call or debit-card transaction is made without any recollection of prior usage. Moreover, I assume that consumers are aware of their own inattention when forecasting their own future consumption choices. In Section 3, I show that for any price schedule, an inattentive consumer’s optimal strategy is to use a threshold rule and consume only those units valued above the endogenous expected marginal price.

In Section 3 I develop a benchmark model which assumes that at the time of contracting consumers are (vertically) homogeneous (so there is no scope for price discrimination) and consumers have correct beliefs (so there are no biases to exploit). I analyze a disclosure requirement that is sufficient to make inattentive consumers attentive. I therefore solve for equilibrium prices under two conditions: first with attentive consumers and second with inattentive consumers. The benchmark result is equivalence. Regardless of the level of market competition, neither consumer inattention nor price-posting regulation affect substantive market outcomes including allocations, firm profits, and consumer surplus. The only effect of price-posting regulation is to restrict the set of feasible equilibrium prices: Regulation causes firms to eliminate penalty fees but compensate with other charges. This captures an argument of some critics of price-posting regulation: that it would only cause firms to recoup lost penalty fees through fixed fees and other charges (Federal Reserve Board 2009a).

The benchmark result relies heavily on consumer homogeneity. Moreover, it does not explain firms’ widespread use of surprise penalty fees or firms’ opposition to disclosure regulation (Altschul et al. 2011). One reason that surprise penalty fees are used in practice may be that they are useful for price discrimination. For instance, cellular-phone overage fees encourage consumers who anticipate

---

3This provides a micro-foundation for the threshold labor supply rule used by Saez (2002) and the consumption rules used by Borenstein (2009) and Grubb and Osborne (2010). These papers use the threshold rules in demand or labor supply estimation, while I explore the supply-side ramifications of such behavior.

4Jamie Dimon, CEO of JPMorgan Chase said, “If you’re a restaurant and you can’t charge for the soda, you’re going to charge more for the burger. Over time, it will all be repriced into the business” (Dash and Schwartz 2010).
high demand to self-select into larger calling-plans. Section 4 enriches the benchmark model by incorporating two ex ante types, with low and high expectations of future demand. Given such heterogeneity, I find that if consumers are inattentive, surprise penalty fees and the resulting price uncertainty can strictly increase not only firm profits but also welfare. The intuition is that price uncertainty relaxes incentive constraints which otherwise limit a firm’s ability to price discriminate. This allows firms with market power to extract more information rents from consumers and increase profits - which can explain firm aversion to price-posting regulation. More surprisingly, inattention may help some consumers and increase overall welfare. It can allow firms to price discriminate effectively while imposing smaller allocative distortions than they would otherwise. This is not always the case (sometimes inattention can increase firm profits but also cause them to increase distortions and reduce welfare), but it is always true when markets are fairly competitive. Thus, the first of two main results is that inattention and surprise penalty fees can be socially valuable when firms price discriminate between unbiased consumers, and this is always true in fairly-competitive markets.

Section 5 enriches the benchmark model in a second direction by assuming that consumers underestimate their own future demand. There is substantial evidence that consumers often have biased beliefs at the time of contracting (e.g. Ausubel and Shui (2005), DellaVigna and Malmendier (2006), Grubb (2009)). Given biased consumers, the total welfare effects of price-posting regulation are ambiguous, but may be second-order relative to the distributional effects. This follows because the degree to which firms can profit from consumer bias is limited by incentive constraints and inattention relaxes these constraints. As a result, attentive consumers who underestimate their own value for a service cannot be exploited in that they can never be induced to pay more than their average value for a product or service. In contrast, consumers who are both inattentive and underestimate their own value for a service can be exploited, and firms can extract profits greater than total social surplus using surprise penalty fees. Thus, the second main result is that price-posting regulation can mitigate consumer welfare losses due to biased beliefs and stop consumer exploitation, possibly lowering profits and increasing consumer surplus by more than total welfare to do so. This is true even in fairly-competitive markets, as the combination of penalty fees and consumer inattention can significantly soften price competition.

Turning to policy applications, discussed in Section 6, the results suggest that regulators should require price-posting for services such as overdraft protection that are not differentially priced to sort consumers into different contracts. However, regulators should be more cautious for services such as cellular-phone calls for which penalty fees help sort consumers across contracts. In particular, the Fed’s opt-in rule for overdraft fees on debit transactions could strongly benefit consumers, but bill-
shock regulation under consideration by the FCC could be socially harmful. Moreover, consumer advocacy for bill-shock regulation may be misplaced because of a failure to anticipate endogenous price-changes.

The paper proceeds as follows. Section 2 discusses related literature. Section 3 introduces the benchmark model, derives an inattentive consumer’s consumption rule, and shows the benchmark equivalence result. Section 4 analyzes the model enriched with ex ante heterogeneity, which explores the role of inattention, penalty fees, and price-posting regulation in price discrimination. Section 5 makes the alternative extension to biased consumer beliefs, for which inattention can increase the scope for exploitation. Section 6 discusses policy implications for proposed FCC bill-shock regulation and Fed overdraft opt-in regulation, and Section 7 concludes. Proofs are in the online appendix (www.mit.edu/~mgrubb).

2 Related Literature

Standard models of consumer choice from multi-part tariffs are static and assume that individuals make a single quantity choice, tailored to the ex post marginal price relevant at the chosen quantity. This implicitly assumes perfect consumer foresight and is empirically rejected by the lack of bunching at tariff kink points in electricity (Borenstein 2009) and cellular-phone-service (Grubb and Osborne 2010) consumption. Relaxing the perfect foresight assumption, attentive consumers will reduce consumption following periods of high usage that increase the chance of triggering penalty fees. Using call-level data, Grubb and Osborne (2010) find no evidence of this behavior among cellular-phone users, suggesting that they are inattentive to their own past usage. Stango and Zinman (2009) find other evidence of inattention: the median checking-account holder could avoid more than 60% of overdraft charges by using alternative cards with available liquidity. Using different data, Stango and Zinman (2010) find that at least 30 percent of overdraft fees are avoidable and that in survey responses “60% of overdrafters reported overdrafting because they ‘thought there was enough money in my account.’”

Liebman and Zeckhauser (2004) analyze optimal nonlinear pricing given ironing or spotlighting decision errors. Consumers who iron have perfect foresight but confuse average price with marginal price. Consumers who spotlight myopically consider only on the marginal price of the current unit. In contrast, I assume consumers make choices optimally conditional on their limited memory.

5Stango and Zinman (2010) also show that individuals who are reminded about overdraft fees by answering an online survey with related (but uninformative) questions such as “Do you have overdraft protection?” are substantially less likely to overdraft. This is similar to Agarwal, Driscoll, Gabaix and Laibson’s (2008) finding that accruing one credit card late penalty fee reduces the likelihood of incurring one in the following month.
In this paper, inattentive consumers are aware of prices when signing a contract, but are uncertain about marginal prices at the point of sale. Many models of add-on pricing examine the opposite situation, by assuming that consumers are aware of marginal prices at the time of purchase, but are unaware of marginal prices or hidden fees at the time they make an ex ante decision to visit a store or purchase a base product (Diamond 1971, Ellison 2005, Gabaix and Laibson 2006, Bubb and Kaufman 2009). As a result, add-ons are sold at monopoly prices in spite of competition or the use of two-part tariffs, either of which would normally lead to marginal cost pricing.

The price-discrimination model with attentive consumers is a competitive sequential-screening model and hence closely related to the literatures on monopoly sequential-screening (surveyed by Rochet and Stole ((2003), Section 8)) and competitive static-screening (surveyed by Stole (2007)), as discussed in Section 4.

A common finding between the biased belief model with attentive consumers and the existing literature is that demand underestimation, due either to biased beliefs (Eliaz and Spiegler 2008, Grubb 2009), myopia (Gabaix and Laibson 2006, Miao 2010), or naive quasi-hyperbolic-discounting (DellaVigna and Malmendier 2004, Eliaz and Spiegler 2006), leads to high marginal prices above marginal cost. In competitive markets, models typically predict that firms offset high marginal fees with lower fixed fees (Shapiro 1995, Gabaix and Laibson 2006, Grubb 2009). Miao (2010) shows that profits from aftermarket sales are not necessarily competed away in primary market competition because firms cannot set negative prices for primary goods. Requiring total prices be nonnegative (the no-free-lunch constraint), I also find that biased beliefs soften price competition. Moreover, I show that inattention can exacerbate the effect by shifting competition from base marginal charges to less-salient penalty fees. Ellison (2005) shows that shrouded add-on fees can soften price competition without biased beliefs, if the consumers most price-sensitive to cuts in fixed fees are those least likely to purchase add-ons.

Gabaix and Laibson (2006) and Bubb and Kaufman (2009) focus on the cross-subsidization of unbiased consumers by biased consumers. Despite cross-subsidization, biased consumers who are attentive can never be exploited in that they always achieve at least their outside options. In contrast, I show that inattention allows consumers to be exploited and can exacerbate the cost of biased beliefs to consumers, even in fairly-competitive markets.

---

6In Bubb and Kaufman’s (2009) model, biased consumers correctly predict their value for the bundle of the base good and the add-on, but overestimate their value of the base good without the add-on. Since they are over-estimating the value of the base good, they can be induced to over-pay and be exploited.
3 Benchmark Model

This section develops the model structure used throughout the paper. The benchmark assumptions that are relaxed later are that consumers have correct beliefs and are (vertically) homogeneous at the time of contracting. After describing the model, I derive optimal strategies of attentive and inattentive consumers. Attentive consumers solve a dynamic programming problem and buy all units valued above a critical threshold which is a function of the date and past consumption. Inattentive consumers cannot condition on past usage, so implement a constant threshold. I define price-posting regulation formally, which in the context of the model is equivalent to making inattentive consumers attentive. Comparing equilibrium pricing with inattentive consumers to that with attentive consumers thus illuminates the effect of price-posting regulation. I show an equivalence result: neither inattention nor price-posting regulation affects substantive market outcomes.

3.1 Model

There are mass 1 of consumers and $N \geq 1$ firms. Each consumer privately learns a vector $x$ of $N$ firm-specific mean-zero brand-tastes, which could for instance capture location on a Hotelling line. At the contracting stage, $t = 0$, firms simultaneously offer contracts, and each consumer either signs a contract or receives her outside option (normalized to zero). At each later period, $t \in \{1, 2\}$, consumers privately learn a taste shock $v_t$ that measures a consumer’s value for a unit of add-on service. Taste shocks $v_t$ are drawn independently with cumulative distribution $F$ that is atomless and has full support on $[0, 1]$. Then consumers (who have accepted a contract) make a binary quantity choice, $q_t \in \{0, 1\}$, by choosing whether or not to consume a unit of service. In the final period, consumers contracted with firm $i$ make a payment $P(q, p_i)$ to firm $i$, as a function of past quantity choices $q = (q_1, q_2)$. Firm $i$’s offered contract can be any deterministic price schedule.\(^7\) A contract is characterized by a vector of prices $p_i = (p_i^0, p_i^1, p_i^2, p_i^3)$ that include a fixed fee $p_0$, base marginal charges $p_1$ and $p_2$ charged for purchasing a unit in either period 1 or 2 respectively, and an additional penalty fee $p_3$ charged if both units are purchased:

$$P(q, p_i) = p_i^0 + p_i^1 q_1 + p_i^2 q_2 + p_i^3 q_1 q_2. \quad (1)$$

A consumer’s total payoff is the sum of a base payoff and a brand taste. A consumer’s base payoff $u$ from contracting with firm $i$ is a function of the value of the base good $v_0$, add-on quantity

\(^7\)See Rochet and Stole (2002) for an insightful discussion of this assumption.
choices $q_t$, private taste shocks $v_t$, and payment to the firm:

$$u(q, v, p^i) = v_0 + q_1 v_1 + q_2 v_2 - P(q, p^i).$$  \hspace{1cm} (2)$$

Conditional on contract prices $p$, a consumer’s optimal consumption strategy can be described by a function mapping valuations to quantity choices: $q(v; p)$. A consumer’s base expected payoff from contracting with firm $i$ and making optimal consumption choices thereafter is $U^i = E[u(q(v; p^i), v, p^i)]$. Similarly, let $S^i = v_0 + E[\sum_{t=1}^{2} (v_t - c) q_t(v; p^i)]$ be the expected social surplus generated by a consumer contracting with firm $i$ and making optimal consumption choices.

A consumer’s total expected payoff, $U^i + x^i$, includes brand taste $x^i$. Thus, fraction $G(U^i; U^{-i})$ of consumers buy from firm $i$ if $i$ offers base expected utility of $U^i$, while competitors offer $U^{-i}$:

$$G(U^i; U^{-i}) = \Pr(U^i + x^i \geq \max_{j \neq i} \{U^j + x^j\}).$$

Firm profits per consumer equal payments less fixed costs (normalized to zero) and marginal cost $c \geq 0$ per unit served. Thus firm $i$’s expected profits are

$$\Pi^i = G(U^i; U^{-i}) E\left[P(q(v; p^i), p^i) - c(q_1(v; p^i) + q_2(v; p^i))\right],$$

which can always be rewritten in terms of total surplus and consumer utility:

$$\Pi^i = G(U^i; U^{-i}) (S^i - U^i).$$

### 3.2 Consumer Strategies

The optimal consumption rule for an attentive consumer who signs a contract would be to consume a unit of service at time $t$ if and only if her value for the unit, $v_t$, exceeds a threshold $v^*(q^{t-1}, t)$ that is a function of the date $t$ and the vector of past usage choices $q^{t-1}$. Let the period one and two thresholds be $v_1^*$ and $v_2^*(q_1)$ respectively. Then, suppressing firm $i$ superscripts from prices,

$$v_2^*(q_1) = p_2 + p_3 q_1,$$ \hspace{1cm} (3)

and $v_1^*$ depends on the distribution of taste shocks:

$$v_1^* = p_1 + (1 - F(p_2 + p_3)) p_3 + \int_{p_2}^{p_2+q_3} (v - p_2) f(v) dv.$$ \hspace{1cm} (4)
The intuition is that $v^*_t$ equals the expected marginal price conditional on purchase, $p_1 + (1 - F (p_2 + p_3)) p_3$, plus the expected opportunity cost of foregone second-period purchases, $\int_{p_2}^{p_2 + p_3} (v_2 - p_2) f (v_2) dv_2$.

Integrating by parts, equation (4) simplifies to

$$v^*_1 = p_1 + \int_{p_2}^{p_2 + p_3} (1 - F (v)) dv.$$  (5)

An inattentive consumer cannot condition her strategy on the date $t$ or on past usage $q^{t-1}_t$ because she does not keep track of these variables. She exhibits imperfect recall. Otherwise, I assume that inattentive consumers are entirely rational and, in particular, are aware of their own inattention and plan accordingly.\(^8\) Formally, the consumer’s decision problem exhibits Piccione and Rubinstein’s (1997) absentmindedness. Unlike optimal strategies for Piccione and Rubinstein’s (1997) absent-minded driver, I show that an inattentive consumer’s optimal strategy is time consistent and hence Bayesian Nash Equilibrium is an appropriate solution concept.\(^9\) Proposition 1 describes an inattentive consumer’s optimal strategy.

**Proposition 1** An inattentive consumer’s optimal strategy is a constant threshold strategy, to buy if and only if $v_t$ exceeds $v^*$. The optimal threshold $v^*$ is equal to the expected marginal price conditional on purchasing in the current period and satisfies:

$$v^* = \frac{p_1 + p_2}{2} + (1 - F (v^*)) p_3.$$  (6)

Equation (6) is necessary up to the fact that all thresholds above one are equivalent and all thresholds below zero are equivalent. For all $p_3 \geq 0$, equation (6) has a unique solution and is sufficient as well as necessary for $v^*$ to be the optimal threshold. A consumer’s choice of $v^*$ is time consistent, she will find it optimal to follow through and implement her chosen $v^*$ in periods one and two.

Note that given fixed prices and a positive penalty fee, equation (6) implies that $v^*$ and $(1 - F (v^*))$ both increase as the distribution of values $F$ increases in a first-order stochastic domi-
nance sense. Thus as anticipated demand increases, the likelihood of incurring a penalty fee and the expected marginal price both increase, leading consumers to be more selective in their consumption choices.

### 3.3 Price-Posting Regulation

Suppose that a firm faced some inattentive consumers and had the option either to disclose nothing or to make inattentive consumers attentive by disclosing the pair \( \{t, q^{t-1}\} \) at the point of sale. I refer to the joint disclosure of \( \{t, q^{t-1}\} \) as price-posting, since in this model it is equivalent to disclosing the date and the marginal price of the current unit.\(^{10}\)

**Definition 1** Price-Posting Regulation (PPR) is the requirement that firms disclose \( \{t, q^{t-1}\} \) at the point of sale.

Note that in a richer model with more than two purchase opportunities, reporting the full purchase history \( q^{t-1} \) would require more than posting transaction prices at point of sale. However, firms commonly set prices only as a function of total purchases \( \sum_{t=1}^{T} q_t \) rather than the full vector \( q^T \).\(^{11}\) In this case, disclosing total purchases to date rather than \( q^{t-1} \) is sufficient to make inattentive consumers attentive.

An alternative regulation that could be considered would prohibit the use of penalty fees:

**Definition 2** Constant-Marginal-Price Regulation (CMPR) is the requirement that firms charge a constant marginal price as a function of usage: \( p_1 = p_2 \) and \( p_3 = 0 \).

In the benchmark model (as well as the first model of biased beliefs in Section 5.1) it will be a result that firms optimally offer attentive consumers two-part tariffs with zero penalty fees. In this case, the two forms of regulation have the same effect on market outcomes, since inattentive consumers behave as attentive consumers do when penalty fees are zero. Moreover, although the formal results in Sections 4 and 5.2 are shown only for PPR, the two regulations would have

\(^{10}\)I do not consider the possibility that firms might disclose \( t \) but conceal \( q^{t-1} \) or vice versa. This is purely for simplicity. Disclosing \( q^{t-1} \) without \( t \) leads consumers make inferences about \( t \) from \( q^{t-1} \). A model in which consumers know \( t \) and are only inattentive to \( q^{t-1} \) has the following feature: When penalty fees are sufficiently high, a consumer who knew \( t \) but not \( q^{t-1} \) would choose different thresholds \( v_1^* \neq v_2^* \) in each period. This would endogenously limit the size of penalty fees but would not qualitatively affect the primary pricing or welfare predictions.

\(^{11}\)For instance, cellular bills are typically only a function of total calling within each calling category (peak, off-peak, etc.), and do not depend on when during the billing cycle calls occurred. Note that I find that it is optimal for firms to deviate from such simple pricing when consumers are attentive. However, it is reasonable to believe that in practice firms are restricted to price as a function only of total usage because contract complexity is inherently expensive. Adding such a restriction to the model would not qualitatively change the main predictions about the consequences of regulation in Propositions 2 and 7 or Corollary 2.
qualitatively similar effects in both the price-discrimination and biased-beliefs models (see Online Appendix C).

3.4 Benchmark Result

When consumers have homogeneous unbiased beliefs ex ante, firms do best by setting marginal charges to implement the first-best allocation and extracting surplus through the fixed fee $p_0$ (balancing the trade-off between mark-up and volume in the standard way). As a result, neither inattention nor PPR have any substantive effect on market outcomes.

**Proposition 2** If consumers have homogeneous unbiased beliefs, $v_t \sim F(v_t)$, then there is a unique equilibrium outcome in which equilibrium allocations are efficient. If at least some consumers are attentive, then equilibrium contracts must offer marginal cost pricing ($p_1 = p_2 = c$ and $p_3 = 0$). If all consumers are inattentive, the set of possible equilibrium prices is larger and includes all three-part tariffs with $p_1 = p_2 = p$ and $p_3 = \frac{c-p}{1-F(c)}$ for $p \in [0, c]$. Price-posting and CMPR would both restrict equilibrium prices but have no effect on allocations, firm profits, or consumer surplus.

The equivalence result in Proposition 2 captures an argument of some critics of PPR: that it would only cause firms to recoup lost penalty fees through fixed fees and other charges (Federal Reserve Board 2009a). However, the result relies heavily on the joint assumptions of (vertical) homogeneity and correct beliefs. Further, Proposition 2’s prediction that firms are indifferent to the use of penalty fees and disclosing marginal price at the point of sale appears inconsistent with firm behavior. In particular, Proposition 2 does not explain firms’ choices not to post transaction prices at the point of sale or their expressed aversion to PPR requiring such disclosure (see Section 6).

4 Price-Discrimination Model

In this section, I relax the assumption of ex ante (vertical) homogeneity imposed in the benchmark model and show that (vertical) heterogeneity and the resulting incentive for firms to price discriminate can explain why consumer inattention is strictly profitable for firms. In this alternative setting, the equivalence result fails and PPR does affect substantive market-outcomes. In particular, PPR will be socially harmful in fairly-competitive markets.
4.1 Model

The model is the same as the Section 3 benchmark, except that there are two types of consumers. Prior to choosing a contract, each consumer privately receives one of two private signals $s \in \{L, H\}$, where $Pr(s = H) = \beta$. As a result, each firm $i$ simultaneously offers a menu with a choice of two contracts, $s \in \{L, H\}$. Each contract is characterized by the vector of prices $p_i^s = (p_{0i}^s, p_{1i}^s, p_{2i}^s, p_{3i}^s)$ and equation (1). Each consumer either signs a contract, $\hat{s} \in \{L, H\}$, from one of the firms or receives her outside option (normalized to zero).

As before, at each later period, $t \in \{1, 2\}$, a consumer privately learns her value $v_t$ for a unit of add-on service. Conditional on receiving signal $s$, a consumer’s values $v_t$ are drawn independently with cumulative conditional distribution $F_s$, which is atomless and has full support on $[0, 1]$. The conditional value distributions are ranked by first-order-stochastic-dominance (FOSD), $F_L(v) \geq F_H(v)$, and the ranking is strict at $v = c$. Marginal cost is assumed to be less than 1 so that the service is socially valuable: $c \in [0, 1)$.

As before, a consumer’s total payoff is the sum of a base payoff and a brand taste, where the base payoff $u(q, v, p_s^i)$ is given by equation (2). The expected base payoff of a consumer of type $s$ who chooses contract $\hat{s}$ from firm $i$ at time zero and makes optimal consumption choices thereafter is $U_{ss}^i = E[u(q_s(v; p_s^i), v, p_s^i) | s]$, where $q_s(v; p)$ is the optimal consumption rule for type $s$ given prices $p$. Define $U_s^i \equiv U_{ss}^i$ to be the expected base payoff of a consumer who chooses the intended contract from firm $i$. Similarly, let $S_s = v_0 + E\left[\sum_{t=1}^{2} (v_t - c) q_{t,s}(v; p_s^i) | s\right]$ be the expected social surplus from a consumer of type $s$ who chooses contract $s$ and makes optimal consumption choices at $t \in \{1, 2\}$. A consumer’s total expected payoff, $U_s^i + x^i$, includes brand taste $x^i$. Fraction $G_s(U_s^i; U_s^{-i})$ of consumers of type $s$ buy from firm $i$ if firm $i$ offers contract $s$ with base expected utility of $U_s^i$, while competitors offer $U_s^{-i}$:

$$G_s(U_s^i; U_s^{-i}) = Pr(U_s^i + x^i \geq \max_{j \neq i} \{U_j^i + x^j\}).$$

Suppressing competitors’ offers $U_s^{-i}$ and firm $i$ superscripts from the notation, the firm’s expected profit maximization problem is:

$$\max_{p_L, p_H} \left( (1 - \beta) G_L(U_L) (S_L - U_L) + \beta G_H(U_H) (S_H - U_H) \right)$$

s.t. $U_s \geq U_{ss} \ \forall s, \hat{s} \in \{L, H\}.$

This initial statement of the firm’s problem encompasses both attentive and inattentive cases. They vary only by the consumers’ optimal consumption rule $q_s(v; p)$, which is given as a function
of prices by equations (3) and (5) in the attentive case but by Proposition 1 in the inattentive case. The constraint that type $H$ not choose contract $L$ ($U_H \geq U_{HL}$) is the downward incentive constraint. The constraint that type $L$ not choose contract $H$ ($U_L \geq U_{LH}$) is the upward incentive constraint.

Conceptually, the firm’s pricing problem can be broken into two parts. First, the firm’s choice of marginal prices determines contract allocations and hence expected social surpluses from serving each type, $S_L$ and $S_H$. Second, the firm’s choice of fixed fees then determines the utilities offered to each type, $U_L$ and $U_H$. The differences $\mu_s \equiv (S_s - U_s)$ are the firm’s markup on each contract and the profit per customer served. Absent ex ante incentive constraints, the choice of markup would be a standard monopoly pricing problem.

I make one of two assumptions: (1) Zero outside option monopoly (ZOOM): $G_s (U_s)$ is one if $U_s \geq 0$ and zero otherwise, which captures a monopolist serving horizontally-homogeneous customers with no brand preference.\footnote{I assume there are $T = 2$ sub-periods when quantity choices are made after a contract is signed. Given attentive consumers and $T = 1$, ZOOM coincides with Courty and Li (2000), which models airline-ticket refund-contracts. When consumers are attentive and $T \geq 1$, ZOOM is nearly a special case of the problem studied by Pavan, Segal and Toikka (2009). However, because I assume period-zero types are discrete rather than continuous, Pavan et al.’s (2009) results do not apply, and conditional independence of values does not lead to a repetition of the Courty and Li (2000) solution. Moreover, I allow for heterogeneous outside-options so that I can move beyond monopoly pricing and analyze imperfect competition.} (2) Heterogeneous outside options (HOO): $G_s (U_s)$ is differentiable and $U_s + \frac{G_s (U_s)}{g_s (U_s)}$ is strictly increasing, which corresponds to a decreasing marginal revenue assumption, guaranteeing the simple monopoly pricing problem has a uniquely optimal markup.

Definition 3 Unconstrained optimal markup $\mu_s^*$ is the optimal markup for type $s$ given first-best allocations and ignoring ex ante incentive constraints: $\mu_s^* = S_s^{FB} - \hat{U}_s$ where $\hat{U}_s \equiv \arg \max_U G_s (U) (S_s^{FB} - U)$.

Unconstrained optimal markups are those that would be charged under third-degree price discrimination. Given ZOOM, $\hat{U}_s = 0$ and $\mu_s^* = S_s^{FB}$. Given HOO, $\mu_s^* = \frac{G_s (U_s)}{g_s (U_s)}$ where $\hat{U}_s$ uniquely satisfies $S_s^{FB} = \hat{U}_s + \frac{G_s (U_s)}{g_s (U_s)}$. Under ZOOM, $\mu_H^* > \mu_L^*$, and under HOO I will often focus on the case in which $\mu_H^* \geq \mu_L^*$. This is a natural assumption if high-average-value customers are high-income customers who have a lower marginal-value of money.

4.2 Attentive Case

I first characterize equilibrium pricing when consumers are attentive because this will be the outcome if PPR is imposed. (Recall that disclosures mandated by PPR compensate for inattentive consumers’ limited memory, enabling them to operate like attentive consumers.) The main result will be that equilibrium contracts induce inefficient allocations except in knife-edge circumstances.
Let $v_{s\hat{s}}$ be the optimal first-period consumption-threshold of an attentive consumer of type $s$ who chooses contract $\hat{s}$ and let $v_s = v_{s\hat{s}}$. The expression for $v_{s\hat{s}}$ is an extension of equation (5):

$$v_{s\hat{s}} = p_{1\hat{s}} + \int_{p_{2\hat{s}}}^{p_{2\hat{s}}+p_{3\hat{s}}} (1 - F_s(v)) \, dv.$$  \hfill (7)

An attentive consumer $s$ who chooses contract $\hat{s}$ earns base expected utility

$$U_{s\hat{s}} = v_0 - p_{0\hat{s}} + \int_{v_{s\hat{s}}}^{1} (v - p_{1\hat{s}}) \, dF_s(v)$$

$$+ F_s(v_{s\hat{s}}) \int_{p_{2\hat{s}}}^{1} (v - p_{2\hat{s}}) \, dF_s(v) + (1 - F_s(v_{s\hat{s}})) \int_{p_{2\hat{s}}+p_{3\hat{s}}}^{1} (v - p_{2\hat{s}} - p_{3\hat{s}}) \, dF_s(v),$$

and for $\hat{s} = s$ earns $U_s = U_{s\hat{s}}$ and generates expected social surplus

$$S_s = v_0 + \int_{v_s}^{1} (v - c) \, dF_s(v) + \int_{p_{2s}+p_{3s}}^{1} (v - c) \, dF_s(v) + F_s(v_s) \int_{p_{2s}}^{p_{2s}+p_{3s}} (v - c) \, dF_s(v).$$  \hfill (9)

It is useful to reframe the firm’s problem in two ways. First, think of the firm choosing offered utility levels $U_s$ so that fixed fees $p_{0s}$ are determined by equation (8) evaluated at $\hat{s} = s$ as function of $U_s$. Second, think of the firm choosing a consumer’s first-period threshold $v_s$ rather than marginal price $p_{1s}$. Given a choice of $v_s$, it is necessary for $p_{1s}$ to satisfy equation (7) evaluated at $\hat{s} = s$.

The firm’s problem can be written as:

$$\max_{U_L, v_L, p_{2L}, p_{3L}} \left\{ (1 - \beta) G_L(U_L) (S_L(v_L, p_{2L}, p_{3L}) - U_L) + \beta G_H(U_H) (S_H(v_H, p_{2H}, p_{3H}) - U_H) \right\}$$

s.t. $U_s \geq U_{s\hat{s}} \forall s, \hat{s} \in \{L, H\}$,

where $U_{s\hat{s}}$ and $S_s$ are given by equations (8) and (9) and $p_{1s}$ and $p_{0s}$ are given by equations (7) and (8) evaluated at $\hat{s} = s$.

Proposition 3 characterizes the solution to a single firm’s problem, treating residual demand $G_s(U_s)$ as exogenous. Proposition 4 applies the result to a Hotelling duopoly, where firm $i$’s residual demand $G_s(U_s^i, U_s^j)$ depends endogenously on firm $j$’s equilibrium offer $U_s^j$.

The solution to the firm’s problem varies depending on which incentive constraints bind. When there is no reason to price discriminate ($\mu_L^* = \mu_H^*$) neither ex ante incentive constraint binds and the firm offers a single first-best contract. When market segment $L$ would receive a discounted markup under third-degree price discrimination ($\mu_L^* < \mu_H^*$) the downward incentive constraint is binding, contract $H$ is first best, and first-order conditions for marginal prices on contract $L$ are:

$$v_L = c + \int_{p_{2L}}^{p_{2L}+p_{3L}} (v - c) f_L(v) \, dv + \frac{\beta}{1 - \beta} \frac{-\partial \Pi / \partial U_H}{\beta G_L(U_L)} \frac{F_L(v_L) - F_H(v_{HL})}{f_L(v_L)},$$  \hfill (10)
\[ p_{2L} = c + \frac{\beta}{1 - \beta} \frac{-\partial \Pi/\partial U_H}{f_H (v_{HL})} F_L (v_{HL}) F_L (p_{2L}) - F_H (p_{2L}) f_H (p_{2L}), \]

\[ p_{2L} + p_{3L} = c + \frac{\beta}{1 - \beta} \frac{-\partial \Pi/\partial U_H}{f_H (v_L)} (1 - F_H (v_{HL})) F_L (p_{2L} + p_{3L}) - F_H (p_{2L} + p_{3L}), \]

where \( v_{HL} = v_L + \int_{p_{2L}}^{p_{2L} + p_{3L}} (F_L (v) - F_H (v)) \, dv \). Equations (10)-(12) imply that allocations are distorted downwards on contract \( L \). When market segment \( H \) would receive a discounted markup under third-degree price discrimination (\( \mu^*_L > \mu^*_H \)) the reverse is true: the upward incentive constraint is binding, contract \( L \) is first best, and first-order conditions for marginal prices on contract \( H \) are:

\[ v_H = c + \int_{p_{2H}}^{p_{2H} + p_{3H}} (v - c) f_H (v) \, dv - \frac{1 - \beta}{\beta} \frac{-\partial \Pi/\partial U_L}{f_L (v_{LH})} F_L (v_{LH}) - F_H (v_H) f_H (v_H), \]

\[ p_{2H} = c - \frac{1 - \beta}{\beta} \frac{-\partial \Pi/\partial U_L}{f_H (v_{LH})} F_L (v_{LH}) F_L (p_{2H}) - F_H (p_{2H}), \]

\[ p_{2H} + p_{3H} = c - \frac{1 - \beta}{\beta} \frac{-\partial \Pi/\partial U_L}{f_H (v_{LH})} (1 - F_H (v_{LH})) F_L (p_{2H} + p_{3H}) - F_H (p_{2H} + p_{3H}), \]

where \( v_{LH} = v_L + \int_{p_{2L} + p_{3L}}^{p_{2L} + p_{3L}} (F_L (v) - F_H (v)) \, dv \). Equations (13)-(15) imply that allocations are distorted upwards on contract \( H \). This characterization is summarized by Proposition 3, which also shows that penalty fees are strictly positive on distortionary contracts, irrespective of the direction of the distortion.

**Proposition 3** Assume a monopolist faces demand curves \( \{G_L (U_L), G_H (U_H)\} \) that satisfy ZOOM or HOO. Demand will fall into one of three categories, depending on how unconstrained optimal markups (Definition 3) are ranked across low and high market-segments:

1. If \( \mu^*_L = \mu^*_H \), then a single marginal-cost contract with markup \( \mu^*_L \) gives both types first-best allocations.

2. If \( \mu^*_H > \mu^*_L \), then \( H \)'s allocation is first best via marginal-cost pricing but \( L \)'s allocation is distorted downwards: \( v_L, p_{2L}, p_{2L} + p_{3L} > c \). Penalty fee \( p_{3L} \) is strictly positive. The triple \( \{v_L, p_{2L}, p_{3L}\} \) satisfies equations (10)-(12).

3. If \( \mu^*_H < \mu^*_L \), then \( L \)'s allocation is first best via marginal-cost pricing but \( H \)'s allocation is distorted upwards: \( v_H, p_{2H}, p_{2H} + p_{3H} < c \). Penalty fee \( p_{3H} \) is strictly positive. The triple \( \{v_H, p_{2H}, p_{3H}\} \) satisfies equations (13)-(15).

Proposition 3 shows that optimal pricing for attentive consumers requires a positive penalty fee and is not only a function of total usage because \( p_{1s} \) differs from \( p_{2s} \) for one of the two contracts.
if \( \mu_L^* \neq \mu_H^* \). To understand optimal pricing it is helpful to compare equations (10)-(12) to the optimal marginal price for ZOOM and \( T = 1 \) characterized by Courty and Li (2000):

\[
p_{CL}^* = c + \frac{\beta}{1-\beta} \frac{F_L(p_{CL}^*) - F_H(p_{CL}^*)}{f_L(p_{CL}^*)}.
\] (16)

Equations (10)-(12) all involve the additional term

\[
\frac{-\partial \Pi/\partial U_H}{\beta G_L(U_L)} > 0
\]

(equal to one given ZOOM) which results from allowing for heterogeneous outside-options. First-period marginal cost in equation (10) is adjusted by \( \int p^L \left( v - c \right) f_L(v) \, dv \), which is the second-period surplus lost when a first-period purchase triggers the penalty fee in period two. Second-period distortions away from marginal cost in equations (11)-(12) are adjusted by the additional terms \( F_H(v_{HL})/F_L(v_L) < 1 \) and \((1 - F_H(v_{HL}))/ (1 - F_L(v_L)) > 1 \) respectively. This implies that penalty fee \( p_{3L} \) is positive.\(^{13}\) Marginal prices are distorted upwards to discourage the high type from choosing the low contract. A positive penalty fee makes the second-period distortion larger after an initial purchase. This is optimal because a deviating high type is more likely to purchase in the first period than a low type. (In fact, the additional terms simply adjust the likelihood ratio \( \beta/(1-\beta) \) to condition on first-period purchase information.)

Proposition 3 can explain the use of penalty fees but not the use of surprise penalty fees. Importantly, Proposition 3 shows that allocations are first best only when unconstrained optimal markups are identical for both types. As Proposition 3 shows, this implies that allocations are only efficient in a Hotelling duopoly when both market segments have identical transportation costs.

**Proposition 4** Let duopolists with marginal costs \( c > 0 \) compete on a uniform Hotelling line with transport costs \( \tau_H \) and \( \tau_L > 0 \) for high and low types respectively that are sufficiently small for strict\(^{14}\) full-market-coverage. (1) If \( \tau_H = \tau_L = \tau \), then the unique equilibrium is for firms to split the market and each offer a single marginal-cost contract with fixed-fee markup of \( \tau \). (2) If \( \tau_H \neq \tau_L \), then all equilibria are inefficient. (3) If \( \tau_H > \tau_L \), then in all symmetric equilibria, high types receive first-best allocations, while low types’ allocation is distorted downwards. For \( \tau_H < \tau_L \), low types receive first best, while high types’ allocation is distorted upwards.

---

\(^{13}\)A final difference is that \( v_{HL} \neq v_L \). This is the reason that Pavan et al.’s (2009) trick of relaxing multi-step-deviation incentive-constraints does not work. The positive penalty-fee ensures that the multi-step deviation to \( \{L,v_{HL}\} \) is more tempting for the high type than the single step deviation to \( \{L,v_L\} \).

\(^{14}\)Strict full-market-coverage requires that every consumer strictly prefer the best offer to her outside option.
The knife-edge efficiency-result in Proposition 3 and Proposition 4 is analogous to findings by Armstrong and Vickers (2001) and Rochet and Stole (2002) in a static rather than sequential screening context. Moreover it is very intuitive: If unconstrained optimal markups are equal, firms can implement first-best allocations with marginal-cost pricing and charge both groups the same fixed fee. If \( \mu_L^* < \mu_H^* \), however, a firm would like to maintain first-best allocations but offer low types a discount relative to high types. This is not incentive compatible, as high types would always pool with low types and choose the discount. As a result, firms are forced to distort the allocation of the low type downwards to maintain incentive compatibility.

### 4.3 Inattentive case

I now characterize equilibrium pricing when consumers are inattentive and respond only to the expected marginal price because they do not keep track of past usage. I first solve the firm’s problem assuming that the firm does not disclose \( \{ t, q_1^{t-1} \} \) to consumers (thereby keeping them inattentive and ensuring penalty fees are a surprise) and then later show that this nondisclosure is optimal in fairly-competitive markets. It is striking that, in contrast to the attentive case, firms can charge different markups to different market segments without distorting allocations. This leads to the result that PPR will reduce welfare in fairly-competitive markets.

Let \( v_{s\hat{s}} \) be the optimal consumption threshold of an inattentive consumer of type \( s \) who chooses contract \( \hat{s} \), and let \( v_s = v_{s\hat{s}} \). The first-order condition for \( v_{s\hat{s}} \) is a natural extension of equation (6):

\[
v_{s\hat{s}} = \frac{p_{1s} + p_{2\hat{s}}}{2} + p_{3\hat{s}} (1 - F_s(v_{s\hat{s}})).
\]  

An inattentive consumer \( s \) who chooses contract \( \hat{s} \) earns base expected utility

\[
U_{s\hat{s}} = v_0 - p_{0\hat{s}} + 2 \int_{v_{s\hat{s}}}^{1} v dF_s(v) - (p_{1\hat{s}} + p_{2\hat{s}}) (1 - F_s(v_{s\hat{s}})) - p_{3\hat{s}} (1 - F_s(v_{s\hat{s}}))^2,
\]

and for \( \hat{s} = s \) earns \( U_s = U_{ss} \) and generates expected surplus

\[
S_s = v_0 + 2 \int_{v_s}^{1} (v - c) dF_s(v).
\]

Define \( \bar{p}_s = (p_{1s} + p_{2s}) / 2 \). When consumers are inattentive, any pair \( \{ p_{1s}, p_{2s} \} \) which have the same average are equivalent, both in terms of allocations and surplus division. I focus on symmetric pricing, \( p_{1s} = p_{2s} \), for which the firm’s problem reduces to the choice of \( p_{0s}, \bar{p}_s, \) and \( p_{3s} \) for \( s \in \{ L, H \} \). It is useful to reframe the firm’s problem in two ways. First, think of the firm
choosing offered utility levels $U_s$ so that fixed fees $p_{0s}$ are determined by equation (20):

\[ p_{0s} = -U_s + v_0 + 2 \int_{v_s}^1 v df_s(v) - 2 \bar{p}_s (1 - F_s(v_s)) - p_{3s} (1 - F_s(v_s))^2. \] (20)

Second, think of the firm first choosing consumer threshold $v_s$ and then choosing the best marginal prices $\bar{p}_s$ and $p_{3s}$ which implement $v_s^*$. Given any fixed choice of offered utility $U_s$ and consumer threshold $v_s$, by Proposition 1 it is necessary for $\bar{p}_s$ to satisfy

\[ \bar{p}_s = v_s - p_{3s} (1 - F_s(v_s)). \] (21)

The firm’s problem can be written as:

\[
\max_{U_L,v_L,p_{3L}} \max_{U_H,v_H,p_{3H}} \left( (1 - \beta) G_L (U_L) (S_L(v_L) - U_L) + \beta G_H (U_H) (S_H(v_H) - U_H) \right) \\
\text{s.t. } U_s \geq U_{s\hat{s}} \forall s, \hat{s} \in \{L, H\}, \\
v_s \in \arg \max_x \left\{ 2 \int_x^1 v f_s(v) dv - 2 \bar{p}_s (1 - F_s(x)) - p_{3s} (1 - F_s(x))^2 \right\},
\]

where $U_{s\hat{s}}$, $S_s$, $p_{0s}$, and $\bar{p}_s$ are given by equations (18) through (21).

Notice that only offered utilities $U_s$ and consumer thresholds $v_s$ enter the objective function directly. Penalty fee $p_{3s}$ only affects profits via the incentive constraints. The first-order condition in equation (21) is sufficient for $v_s$ to be incentive compatible for all $p_{3s} \geq 0$. Moreover, for any $v_s > 0$, increasing $p_{3s}$ weakly relaxes both ex ante incentive constraints, from which it follows that it is weakly optimal to set $p_{3s}$ as large as possible.

**Proposition 5** Increasing $p_{3L}$ weakly relaxes the downward incentive constraint without affecting the upward incentive constraint. Increasing $p_{3H}$ weakly relaxes the upward incentive constraint without affecting the downward incentive constraint. It is weakly optimal to choose nonnegative penalties $p_{3s}$ as large as possible.

For intuition behind Proposition 5 consider what happens if penalty $p_{3s}$ is increased by one dollar. First, base marginal charge $\bar{p}_s$ must be reduced by $(1 - F_s(v_s))$ to keep expected marginal price $v_s$ constant (equation (21)). This reduces expected variable payments by $(1 - F_s(v_s))^2$ because the extra dollar in the penalty fee is paid with probability $(1 - F_s(v_s))^2$ but the $(1 - F_s(v_s))$ discount on the base marginal charge is paid with probability $2(1 - F_s(v_s))$. Thus a second change is that the fixed fee is increased by $(1 - F_s(v_s))^2$ (equation (20)). By construction, these changes

\[\uparrow Up to the fact that all thresholds above one are equivalent, and all thresholds below zero are equivalent.\]
leave type $s$ indifferent. Any other type $\hat{s}$ that chooses contract $s$ would pay the same increase in the fixed fee but receive a smaller reduction in expected variable payments. If type $\hat{s}$ buys with probability $\pi = 1 - F_\hat{s}(v_{\hat{s}s})$, her expected variable payments are reduced by $2\pi (1 - F_s(v_s)) - \pi^2$. Notice that this reduction is maximized at $\pi = (1 - F_s(v_s))$. Thus regardless of whether type $\hat{s}$ is a higher type and $\pi > (1 - F_s(v_s))$ or a lower type and $\pi < (1 - F_s(v_s))$, increasing the penalty $p_{3s}$ increases total expected payments for her on contract $s$.

Proposition 5 suggests that optimal penalty fees could be unreasonably high. In practice, however, they would be restricted by a variety of forces.

Remark 1 As penalty fees grow large, the remaining profit increase from increasing them all the way to infinity becomes arbitrarily small because profits are bounded (strictly) below first-best surplus. Hence an arbitrarily small cost of raising penalty fees would be sufficient to endogenously limit penalty fees to be finite. Economic forces that would endogenously restrict penalty fees include: (1) limited liability, (2) mild risk aversion, (3) regulatory threat, (4) a small fraction of consumers who are attentive, (5) rationally inattentive consumers who could invest effort $k > 0$ to be attentive if it were worth doing so, and (6) consumers who attend to the date.

For simplicity, I exogenously impose the upper bound $p_{3s} \leq h_s(v_s)$ stated in Condition 1. This upper bound corresponds either to a cap on penalty fees in the case $h_s(v_s) = p_{\text{max}} > 0$, or to the restriction to nonnegative marginal prices in the case $h_s(v_s) = v_s / (1 - F_s(v_s))$. Notice that all prior results and statements remain true with this addition to the problem.\footnote{In particular, the constraint is symmetric so that any $\{p_{1s}, p_{2s}\}$ which have the same mean are still equivalent.}

**Condition 1** Penalty fees are bounded by $p_{3s} \leq h_s(v_s) \in \{p_{\text{max}}, v_s / (1 - F_s(v_s))\}$.

Proposition 6 characterizes the solution to a single firm’s problem, treating residual demand $G_s(U_s)$ as exogenous. Proposition 7 applies the result to a fairly-competitive Hotelling duopoly, where firm $i$’s residual demand $G_s(U_{is}, U_{js})$ depends endogenously on firm $j$’s equilibrium offer $U_{is}$ and draws conclusions about PPR.

As in the attentive case, the solution to a single firm’s problem varies depending on which incentive constraints bind. When the downward incentive constraint is binding, contract $H$ is first best ($v_H = c$) but contract $L$ allocations are distorted downwards ($v_L > c$) and threshold $v_L$ satisfies the first-order condition:

$$v_L = c + \frac{\beta}{1 - \beta} \left( F_L(v_L) - F_H(v_{HL}) - \frac{\partial \Pi / \partial U_H}{\beta G_L(U_L)} \left( (1 + p_{3L} f_L(v_L)) + \frac{1}{2} (F_L(v_L) - F_H(v_{HL})) h'_L(v_L) \right) \right),$$

(22)
where \( v_{HL} = v_L + p_{3L} (F_L(v_L) - F_H(v_{HL})) \). When the upward incentive constraint is binding, contract \( L \) is first best \((v_L = c)\) but contract \( H \) allocations are distorted upwards \((v_H < c)\) and Threshold \( v_H \) satisfies the first-order condition:

\[
v_H = c - \frac{1 - \beta}{\beta} F_L(v_{HL}) - F_H(v_H) \frac{-\partial \Pi / \partial U_L}{(1 - \beta) G_H(U_H)} \left[ (1 + p_{3H} f_H(v_H)) - \frac{1}{2} (F_L(v_{HL}) - F_H(v_H)) h_H'(v_H) \right],
\]

where \( v_{LH} = v_H - p_{3H} (F_L(v_{LH}) - F_H(v_H)) \). The crucial difference relative to the attentive case is that both constraints may be slack even when unconstrained optimal markups differ.

To state Proposition 6 first define \( X_H \equiv 2 \int_c^{v_{HL}} (v - c) dF_H(v) + (v_{HL} - c) (F_L(c) - F_H(v_{HL})) \) where \( v_{HL} \) uniquely satisfies \( v_{HL} = c + h_L(c) (F_L(c) - F_H(v_{HL})) \) and \( X_L \equiv 2 \int_{v_{LH}}^{c} (c - v) dF_H(v) - (c - v_{LH}) (F_L(v_{LH}) - F_H(c)) \) where \( v_{LH} \) uniquely satisfies \( v_{LH} = c - h_H(c) (F_L(v_{LH}) - F_H(c)) \).

Note that both \( X_L \) and \( X_H \) are strictly positive.

**Proposition 6** Assume (1) a monopolist faces demand curves \( \{G_L(U_L), G_H(U_H)\} \) that satisfy ZOOM or HOO, (2) the firm chooses not to disclose \( \{t, q^{t-1}\} \), and (3) penalty fees \( p_{3s} \) are restricted by Condition 7. Demand will fall into one of three categories, depending on how unconstrained optimal markups (Definition 3) differ across market-segments:

1. If \( \mu^*_L - \mu^*_H \in [-X_L, X_H] \), then both types receive first-best allocations \((v_L = v_H = c)\) and contract mark-ups are \( \mu^*_L \) and \( \mu^*_H \) respectively.

2. If \( \mu^*_L - \mu^*_H > X_H \), then \( H \)’s allocation is first best \((v_H = c)\) but \( L \)’s allocation is distorted downwards \((v_L > c)\). Threshold \( v_L \) satisfies equation (22). \( L \) pays the maximum penalty fee.

3. If \( \mu^*_L - \mu^*_H < -X_L \), then \( L \)’s allocation is first best \((v_L = c)\) but \( H \)’s allocation is distorted upwards \((v_H < c)\). Threshold \( v_H \) satisfies equation (23). \( H \) pays the maximum penalty fee.

Note that while Proposition 6 exogenously assumes nondisclosure of penalty fees, it is clear by comparison to Proposition 3 that if unconstrained optimal markups differ but satisfy \((\mu^*_H - \mu^*_L) \in [-X_L, X_H]\) then nondisclosure is strictly optimal. Comparing Propositions 3 and 6 shows the underlying insight of the first main result: that the combination of surprise penalty fees and consumer inattention can be both profitable and socially valuable by reducing allocative distortions due to price discrimination when unconstrained optimal markups across different consumer segments are different, but not too different. In the attentive problem, contracts implement first-best allocations only for the knife-edge case \( \mu^*_L = \mu^*_H \). With inattentive consumers this is no longer true. Slack ex ante incentive constraints and first-best allocations are a feature for \((\mu^*_H - \mu^*_L) \) in an interval around zero because penalty fees relax incentive constraints when consumers are inattentive.
For intuition, suppose that $\mu^*_H > \mu^*_L$. If consumers are attentive, firms cannot induce first-best allocations and charge low types a discounted markup. First-best allocations require marginal-cost pricing for every unit on every contract. With identical marginal prices on all contracts, all consumers choose the lowest fixed fee and pay the same markup. To discount low types’ markup, firms must combine a discounted fixed fee with higher marginal prices that distort quantity choices. The striking result for inattentive consumers is that this is no longer the case for small discounts. First-best allocations only require that expected marginal prices equal marginal cost and could, for instance, be implemented by offering $\bar{p}_s = 0$ and $p_{3s} = c/(1 - F_s(c))$. As high types pay penalty fees more often than low types, these contracts involve lower penalty fees on the high contract to achieve the same expected marginal price. Moreover, high types are willing to pay a higher fixed-fee premium for the reduction in penalty fees than are low types. As a result, the low-type contract can offer a discounted markup without distorting allocations or attracting high types.

Building on these insights from comparing Propositions 3 and 6, Proposition 7 completes the first main result. The combination of penalty fees and consumer inattention are socially valuable and PPR is counterproductive whenever markets are fairly competitive.

**Proposition 7** Let duopolists compete on a uniform Hotelling line, high types have transportation costs $\tau_H = \tau_H$ strictly higher than low types $\tau_L = \tau_L$, and marginal cost $c$ be strictly positive. If $\tau > 0$ is sufficiently small, then: (1) In the unique (up to penalty fees) symmetric equilibrium, all customers are served, allocations are first best, and mark-ups are $\mu_s = \tau_s$. Moreover, surprise penalty fees are charged but not disclosed at the point-of-sale and the set of equilibrium prices includes $p_{1s} = p_{2s} = 0$ and $p_{3s} = c/(1 - F_s(c))$. (2) PPR would strictly decrease welfare and firm profits. Low types would lose while high types would win.

The intuition behind the PPR welfare result in Proposition 7 is as follows. In a symmetric equilibrium, unconstrained optimal markups equal transportation costs. Thus, in fairly-competitive markets when $\tau$ is small, and hence the difference between $\tau_H$ and $\tau_L$ is also small, unconstrained optimal markups will satisfy $(\mu^*_H - \mu^*_L) \in [-X_L, X_H]$. Equilibrium without PPR is efficient because competition ensures firms only want to charge high types a slightly higher markup and this can be achieved using only surprise penalty fees. PPR eliminates surprise penalty fees from firms’ price-discrimination toolbox and forces firms to introduce quantity distortions on contract $L$, thereby lowering social welfare. Because the distortion makes price discrimination costly for firms, they also respond by charging more similar markups: increasing the markup on contract $L$ and reducing

---

17Note that part (2) of Proposition 7 is also true for regulation banning penalty fees. See Online Appendix C.
the markup on contract $H$. Low types are hurt both by the quantity distortion and the higher markup on contract $L$ while high types benefit from the markup reduction on contract $H$. Firm market shares are unaffected in equilibrium, but profits are reduced because the loss from reducing markups on contract $H$ exceed the gains from raising markups on contract $L$ by a factor of $H/L$. This is because $L$ types are more price-sensitive, so on the margin it is more expensive to raise markups on contract $L$ in terms of market share.\footnote{Shifts in markups in each segment are already inversely weighted by shares of each segment $\beta$ and $(1 - \beta)$ since the shares reflect the cost of distorting that segment. Thus the difference in price sensitivity drives the difference in relative profit changes, rather than relative segment sizes.}

An important caveat to Proposition\footnote{Shifts in markups in each segment are already inversely weighted by shares of each segment $\beta$ and $(1 - \beta)$ since the shares reflect the cost of distorting that segment. Thus the difference in price sensitivity drives the difference in relative profit changes, rather than relative segment sizes.} is that it implicitly assumes firms’ only available tools for price discrimination are nonlinear pricing and surprise penalty fees. Thus the only substitute for surprise penalties following PPR is allocative distortion via nonlinear pricing. If firms have other alternatives, such as a student discount or other third-degree price-discrimination schemes, firms might substitute towards these pricing tactics and need to distort allocations less. Thus the welfare loss from PPR identified by Proposition\footnote{Shifts in markups in each segment are already inversely weighted by shares of each segment $\beta$ and $(1 - \beta)$ since the shares reflect the cost of distorting that segment. Thus the difference in price sensitivity drives the difference in relative profit changes, rather than relative segment sizes.} may only be an upper bound on the true loss.

In contrast with fairly-competitive markets, sufficient market power implies that inattention and surprise penalty fees do not produce efficient outcomes. This is illustrated by Corollary\footnote{Shifts in markups in each segment are already inversely weighted by shares of each segment $\beta$ and $(1 - \beta)$ since the shares reflect the cost of distorting that segment. Thus the difference in price sensitivity drives the difference in relative profit changes, rather than relative segment sizes.}

**Corollary 1** If a monopolist serves consumers with zero outside option, the upward ex ante incentive constraint binds and the low type’s allocation is distorted below first best.

When there is sufficient market power, the impact of regulation becomes ambiguous. Both surprise penalty fees and quantity distortions are useful tools for price discrimination. In some cases (including fairly-competitive markets) they are substitutes and regulation that eliminates surprise increases quantity distortions. In other cases they are complementary and the reverse is true. This is illustrated by the following example:

**Example 1** Consider a monopolist serving consumers with zero outside option. Let values of high types be uniformly distributed between 0 and 1. Let values of low types be uniformly distributed between 0 and $1/2$ with probability $3/4$ and be uniformly distributed between $1/2$ and 1 with probability $1/4$. For any $c \in (0, 1)$, the monopolist finds it strictly profitable to make penalty fees a surprise. For $c = 1/4$, PPR would lower firm profits but raise total welfare. However, for $c = 1/2$, PPR would reduce total welfare.
5 Biased Beliefs Model

Section 4 showed that consumer inattention makes surprise penalty fees an efficient tool for price discrimination. Proposition 7 concludes that, in fairly-competitive markets, PPR reduces social welfare because it forces price-discriminating firms to substitute towards quantity distortions. This section explores an alternative role for surprise penalty fees: a tool to exploit consumer bias. When consumers underestimate their consumption of the add-on good or service, welfare implications for PPR differ substantially. PPR may exacerbate or ameliorate allocative distortions created by biased beliefs depending on the size of marginal costs. However, the effect of first-order importance may be on surplus distribution rather than total welfare. PPR prevents exploitation of inattentive consumers:

Definition 4 A consumer is exploited if on average she receives less than her outside option.

5.1 Continuous taste shocks and welfare

Return to the assumption in the benchmark model that consumers all have the same distribution of taste shocks $F$. Now, however, assume that consumers believe that the distribution is $F^*$, which is first-order-stochastically-dominated by $F$ so that consumers underestimate their demand for the add-on service. Both $F$ and $F^*$ are continuous and strictly increasing on $[0, 1]$ and the FOSD relationship is strict for some $v \in (0, 1)$.

A consumer’s true base-expected-payoff from contracting with firm $i$ at the contracting stage and making optimal consumption choices thereafter remains $U^i = E[u(q(v;p^i), v) | F]$. However, a consumer’s perceived base-expected-payoff differs because expectations are taken with respect to consumer beliefs: $U^{s^i} = E[u(q(v;p^i), v) | F^*]$. The fraction of consumers of type $s$ who buy from firm $i$ depends on the perceived base-expected-utility offers of firms rather than the true expected-utilities: $G(U^{s^i}; U^{s^-i})$. Thus firm $i$’s expected profits are

$$\Pi^i = G(U^{s^i}; U^{s^-i}) E[P(q(v;p^i), p^i) - c(q_1(v;p^i) + q_2(v;p^i)) | F],$$

or rewritten in terms of true social surplus and consumers’ true and perceived expected-utilities:

$$\Pi^i = G(U^{s^i}; U^{s^-i})(S^i - U^i).$$

19Overconfidence with only two subperiods would require underestimation of the correlation in $v_t$ across periods.
5.1.1 Attentive Case

If attentive consumers underestimate their demand for the service ex ante, firms have an incentive to set marginal charges above marginal cost, irrespective of competition (e.g. Grubb (2009)). This is reflected in Proposition 8 which characterizes pricing in the attentive case.\(^\text{20}\)

**Proposition 8** If attentive consumers underestimate demand, then the optimal contract is a two-part tariff \((p_3 = 0, p_1 = p_2 = p)\) with marginal price \(p = c + (F^* (p) - F (p)) / f (p)\) and profits

\[
\Pi = G (U^*) \left( v_0 - U^* + 2 \int_{p}^{1} \left( v - c - \frac{F^* (v) - F (v)}{f (v)} \right) f (v) dv \right). \tag{24}
\]

No consumers are exploited and all transactions generate positive surplus. If \(F (c) < F^* (c)\) (bias is strict at \(p = c\)) then \(p > c\) and allocations are inefficiently low.

In the absence of inattention, bias distorts consumption downwards because a firm is limited in how much surplus it can extract ex ante through fixed fees by consumers’ low estimate of their value for the service. The firm must wait until consumers draw high values and extract surplus through distortionary marginal charges. Nevertheless, there is no exploitation or surplus-reducing trade. Note that Proposition 8 implies that CMPR is equivalent to PPR because two-part tariffs are optimal when consumers are attentive.

5.1.2 Inattentive Case

The consumption threshold chosen by an inattentive consumer with biased beliefs satisfies

\[
v^* = \frac{p_1 + p_2}{2} + p_3 \left( 1 - F^* (v^*) \right), \tag{25}
\]

which substitutes consumer beliefs in place of the true distribution of tastes in equation \([6]\). As before, I focus on symmetric pricing \(p_1 = p_2 = \bar{p}\) and reframe the firm’s problem in two ways. First, think of the firm choosing perceived expected-utility \(U^*\) so that the fixed fee \(p_0\) is given by

\[
p_0 = -U^* + v_0 + 2 \int_{v^*}^{1} vdF^* (v) - 2\bar{p} \left( 1 - F^* (v^*) \right) - p_3 \left( 1 - F^* (v^*) \right)^2. \tag{26}
\]

Second, think of the firm first choosing consumer threshold \(v^*\) and then choosing the best marginal prices \(\bar{p}\) and \(p_3\) which implement \(v^*\). Using equations \((25)\) and \((26)\), firm profits can be written as

\(^{20}\)Marginal pricing is the unit-demand analog of that characterized by Grubb (2009) for continuous demand and \(T = 1\), repeated in each subperiod \(t \in \{1, 2\}\).
a function of perceived expected-utility $U^*$, consumer threshold $v^*$, and penalty $p_3$:

$$
\Pi = G(U^*) \left( v_0 - U^* + 2 \int_{v^*}^{1} \left( v - c - \frac{F^*(v) - F(v)}{f(v)} \right) f(v) \, dv \right) + p_3 (F^*(v^*) - F(v^*))^2 .
$$

Comparing equations (24) and (27) shows that the firm can make strictly higher profits charging a surprise penalty fee to inattentive consumers than by selling to attentive consumers. Profits increase linearly in the penalty fee, and for simplicity I impose a maximum penalty fee $p_{\text{max}} > 0$. (Many of the economic forces identified in Remark 1 would endogenously restrict penalty fees, although the cost of raising penalty fees must be substantial rather than arbitrarily small to have an effect.) As the optimal penalty fee is $p_{\text{max}}$ and equation (25) is sufficient for $v^*$ to be incentive compatible given a positive penalty fee, the firm’s problem can be written as:

$$
\max_{U^*, v^*} G(U^*) \left( v_0 - U^* + 2 \int_{v^*}^{1} \left( v - c - \frac{F^*(v) - F(v)}{f(v)} \right) f(v) \, dv \right) + p_{\text{max}} (F^*(v^*) - F(v^*))^2 .
$$

Equation (28) shows that profits are increasing in the size of the disagreement between consumer and firm about the consumer’s per-period purchase probability $|F^*(v^*) - F(v^*)|$. Firms therefore have an incentive to adjust consumers’ threshold choice $v^*$ to increase this disagreement, which could increase or decrease distortions relative to the attentive case. For some intermediate marginal costs, maximizing disagreement ameliorates inefficiency and PPR reduces welfare. However, if marginal costs are close to zero or one, then maximizing disagreement entails increased inefficiency and PPR increases welfare (Proposition 9). For example, if marginal cost is zero, then $F^*(c) = F(c) = 0$ and at marginal-cost pricing both firm and consumers agree that the consumer always purchases. Hence attentive pricing is at marginal cost, but with inattentive consumers firms raise the expected marginal price $v^*$ above zero to create exploitable disagreement.

To state the next result, I parameterize consumers’ degree of bias. Let $F$ and $\hat{F}$ have full support with continuous densities on $[0,1]$ that cross finitely many times, $F < \hat{F}$ for all $v \in (0,1)$ (a strong form of strict FOSD), and $F^* = \gamma \hat{F} + (1 - \gamma) F$ for some $\gamma \in (0,1]$. Consumers underestimate demand for any $\gamma > 0$ but consumers’ bias goes to zero as $\gamma$ goes to zero.

**Proposition 9** Assume penalty fees are bounded by $p_{\text{max}} > 0$. If bias is sufficiently small ($\gamma$ is sufficiently close to zero) then: (1) When marginal cost $c \geq 0$ is close to zero, inattention exacerbates underconsumption. (2) There exists an intermediate marginal cost $c \in (0,1)$ for which inattention ameliorates underconsumption. In addition, if the maximum penalty fee $p_{\text{max}}$ is sufficiently large then: (3) When marginal cost is close to one, inattention creates overconsumption worse than the attentive underconsumption. In cases (1) and (3) PPR strictly improves welfare, while in case (2)
it strictly reduces welfare. CMPR would have an identical effect.

Note that Proposition 9 points out that when consumers are inattentive, consumers who under-estimate their demand for a product or service may be induced to overconsume. For instance when \( c \) is slightly above one, all product sales are inefficient. Yet because inattentive consumers underestimate their likely values for the product, sales still take place. PPR would increase consumer surplus and total welfare by ending sales of these products.

Although PPR’s effect on total welfare is ambiguous, clearer predictions can be made about its effect on the distribution of surplus between firms and consumers:

**Proposition 10** (1) Assume the firm is a zero-outside-option monopolist. If the upper bound on penalty fees, \( p^{\text{max}} \), is sufficiently large then inattentive consumers are exploited. PPR eliminates exploitation by shifting surplus from the firm to consumers. (2) Assume duopolists compete on a uniform Hotelling line with transport cost \( \tau \). If the add-on is socially valuable \((c < 1)\) and the market is sufficiently competitive \((\tau < (2/3)(v_0 + \int_1^{\infty} (v - c) f^*(v) dv))\) then there is full-market-coverage and firm profits equal \( \tau \) independent of PPR.

Proposition 10 suggests that PPR could be beneficial for preventing consumer exploitation despite ambiguous effects on social welfare. However, Proposition 10 suggests that this is only true for a monopoly. Even if large revenues are earned on surprise penalty fees when consumers are inattentive, Proposition 10 shows that, under imperfect competition, these are rebated back to consumers through lower fixed fees so that firms earn identical markups to those charged under PPR. Therefore consumers are residual claimants of social surplus with respect to PPR.

### 5.2 No-free-lunch and surplus distribution

Proposition 10 suggests that, under monopoly, the effects of inattention and PPR on total welfare may be less important than their effects on the distribution of surplus. However, Proposition 10 also shows that duopoly profits are invariant to PPR: While PPR limits penalty-fee revenue, in equilibrium these revenue losses are exactly off-set by increases in fixed fees. While this competitive result is intuitive, it is not robust because it relies on the unrealistic assumption that firms can charge substantially negative fixed fees.

To better understand the effect of PPR on surplus distribution under imperfect competition, I endogenously restrict penalty fees by imposing a no-arbitrage condition that I call the no-free-lunch (NFL) constraint. The NFL constraint arises endogenously if there exists a large pool of attentive potential-customers (or potential customers with a very low cost \( k \) of paying attention) with zero
value for the service. Such consumers restrict the payment function to be nonnegative: If the payment function were negative at some allocation, such consumers would buy exactly the right quantities to earn the subsidy.

**Definition 5** The no-free-lunch (NFL) constraint restricts consumer payments to be nonnegative at all allocations: $p_0s \geq 0$, $p_0s + p_1s \geq 0$, $p_0s + p_2s \geq 0$, and $p_0s + p_1s + p_2s + p_3s \geq 0$.

Finally, to focus on distributional issues, assume taste shocks have a Bernoulli distribution: $v_t$ are independent and are equal to one with probability $\alpha$ and zero otherwise. Consumers underestimate their demand and believe that $v_t$ equals one with probability $\alpha' < \alpha$. Also assume $c \in (0,1)$. By introducing Bernoulli taste shocks, I ensure that firms induce first-best allocations with or without PPR so that regulation only effects the distribution of surplus:

**Lemma 1** Given Bernoulli taste shocks, attentive or inattentive consumers who underestimate demand ($\alpha' < \alpha$), $c \in (0,1)$, and the NFL constraint, firms set prices which induce the efficient allocation: consumers buy if and only if $v_t = 1$.

As before, the first step to solve for equilibrium duopoly pricing is to write down and solve a monopolist’s problem. This initial step, for both attentive and inattentive cases, is made in Online Appendix A where Propositions 13 and 14 characterize optimal monopoly pricing. In this section, I move directly to describing equilibrium duopoly pricing.

### 5.2.1 Attentive case

Absent the NFL constraint, Hotelling duopolists would each offer a two-part tariff with marginal price of one to maximally exploit consumer bias. They would compete on fixed fees, which would be set to earn a markup of $\tau$ in equilibrium. For $\tau \geq 2\alpha (1-c)$, this is exactly the outcome with the NFL constraint. However, in more competitive markets with $\tau < 2\alpha (1-c)$, this would require charging a negative fixed fee. When $\tau$ falls below this threshold and the fixed fee has already been reduced to zero, firms must lower marginal fees if they wish to lower their markups. In more competitive markets competition shifts first to base marginal charges and finally to penalty fees. This progressively softens price competition and firms charge markups above $\tau$. Thus increasing competition is partially mitigated by reduced consumer price sensitivity.

---

21Similar results would follow from a negative lower-bound rather than zero lower-bound on payments. In fact, because fixed costs have been normalized to zero, the normalized lower bound should be equal to the negative of fixed costs. (The presence of high fixed-costs allows firms to subsidize consumers without making payments to consumers. Hardware discounts with cellular-phone-service contracts are an example.)

22Consumers may or may not want an extra unit but always have the same value when they do want one.
Consider a firm’s choice of marginal fees \( \{p_1, p_2, p_4\} \), where \( p_4 = p_2 + p_3 \) is the marginal price for a second-period purchase conditional on a first-period purchase. When lowering prices, a firm prefers to first cut those fees to which consumers are most price-sensitive. Consumers are less price-sensitive to \( p_1 \) than to the fixed fee because they underestimate the chance of paying \( p_1 \) by a factor \( \alpha'/\alpha \). The reduced price-sensitivity is compounded for \( p_4 \), since consumers underestimate the chance of making two purchases by \( (\alpha'/\alpha)^2 \). However, the reduce price-sensitivity is mitigated for \( p_2 \). While consumers underestimate the chance of demanding a unit in the second period, they overestimate the chance that \( p_2 \) is the relevant second-period price because they underestimate the likelihood of an initial purchase triggering a penalty fee. Thus, once fixed fees are reduce to zero, a firm reducing prices would like to first cut \( p_2 \), second cut \( p_1 \), and lastly cut \( p_4 \). (In fact a firm cutting \( p_2 \) must simultaneously reduce \( p_1 \) a proportion \( \alpha' \) as much to satisfy the incentive constraint \( v^*_1 \leq 1 \).)

This program of price reduction in response to increasing competition leads to four qualitative pricing regions depicted in the top panel of Figure 1, which plots the equilibrium markup as a function of the transportation cost \( \tau \).\(^{23}\) Four dashed lines show the markups relevant for the four possible pricing regions. The solid bold line shows the equilibrium markup, which is increasing in \( \tau \) within pricing regions but constant between regions. Starting at the right side of the figure, and working leftward as \( \tau \) falls and competition increases, equilibrium begins in region 1 where firms compete on fixed fees and markups equal \( \tau \). After fixed fees have reached zero, consumers become discontinuously less price-sensitive and markups are temporarily flat until equilibrium transitions to region 2 where firms begin competing on \( p_1 \) and \( p_2 \) and markups equal \( (\alpha/\alpha') (1 - \alpha + \alpha') \tau \). As transportation costs fall, equilibrium continues to transition through the four competitive regions so that markups are weakly decreasing in absolute levels but weakly increasing as a proportion of transportation costs as competition shifts towards fees to which consumers are less and less price-sensitive. In region 3 firms compete on \( p_1 \) and markups equal \( (\alpha/\alpha') \tau \) while in region 4 firms compete on the penalty fee and markups equal \( (\alpha/\alpha')^2 \tau \).

**Proposition 11** Assume duopoly on a uniform Hotelling line, the NFL constraint, Bernoulli taste shocks, attentive consumers who underestimate demand \( (\alpha' < \alpha) \) and \( c \in [0,1) \). Let base-good value \( v_0 \) be sufficiently large for strict full-market-coverage.\(^{24}\) There are four competitive regions

\(^{23}\)Note that the figure is shown for full market-coverage and \( c = 0 \). If marginal cost is strictly positive then some of the four competitive regions may not be relevant. For instance, if \( c > \alpha/2 \) then region 4 is never reached. Under perfect competition \( (\tau = 0) \) expected markups will always be zero, but for \( c > \alpha/2 \) this means the penalty fee alone is insufficient for the firm to break even so other fees must remain positive.

\(^{24}\)See footnote 14.
over which markups are proportional to $\tau$. Markups are constant between regions.

<table>
<thead>
<tr>
<th>Region</th>
<th>$\tau_{\min}$</th>
<th>$\tau_{\max}$</th>
<th>markup $\mu$</th>
<th>Competition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$2\alpha (1 - c)$</td>
<td>$2\alpha' (1 - c) / (1 - \alpha' + \alpha')$</td>
<td>$\tau$</td>
<td>fixed fees</td>
</tr>
<tr>
<td>2</td>
<td>$\alpha' (\alpha - 2c)$</td>
<td>$\alpha' (\alpha - 2c) + \alpha' (1 - \alpha')$</td>
<td>$(\alpha/\alpha') \tau$</td>
<td>base marginal charges</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>$(\alpha^2 / \alpha) (\alpha - 2c)$</td>
<td>$(\alpha/\alpha')^2 \tau$</td>
<td>penalty fees</td>
</tr>
</tbody>
</table>

Duopoly profits equal the markup and consumers’ true expected utility is $U = S^{FB} - \mu \geq 0$. No consumers are exploited.

An interesting feature is that although consumers are always made better off by increased competition, the cost of their bias is not decreasing monotonically with competition. Between pricing regions where markups are constant, increasing competition increases the gap between the markup and $\tau$. As the markup would be $\tau$ if consumers were unbiased, this means the cost of their bias is locally increasing in competition. This contradicts a common intuition that increased competition reduces the importance of policy interventions to address consumer biases.

5.2.2 Inattentive case

Proposition 11 shows that the NFL constraint softens competition when consumers are biased, leading to markups above $\tau$ by forcing firms to compete on marginal fees rather than the fixed fee. Proposition 12 shows the same is true with inattentive consumers, but the magnitude of the effect is typically higher because inattention allows firms to raise penalty fees. As a result, Corollary 2 shows that PPR typically helps consumers by intensifying competition.

To state the proposition, define $Y \equiv \frac{(\alpha - \alpha')^2}{\alpha (1 - \alpha')}$. 

**Proposition 12** Assume duopoly on a uniform Hotelling line, the NFL constraint, Bernoulli taste shocks, consumers who underestimate demand ($\alpha' < \alpha$) and $c \in [0, 1)$. Let base-good value $v_0$ be sufficiently large for strict full-market-coverage. Firms charge surprise penalty fees preferring not to disclose $\{q^{t-1}, t\}$, a strict preference for $\tau > (\alpha')^2 (\alpha - 2c) / \alpha$. There are two competitive regions over which markups are proportional to $\tau$. Markups are constant between regions.

<table>
<thead>
<tr>
<th>Region</th>
<th>$\tau_{\min}$</th>
<th>$\tau_{\max}$</th>
<th>markup $\mu$</th>
<th>Competition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$(2\alpha (1 - c) + Y) / (1 + Y) - \alpha'$</td>
<td>$(1 + Y) \tau$</td>
<td>all fees</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>$(\alpha' / \alpha) (\alpha - 2ca')$</td>
<td>$(\alpha/\alpha')^2 \tau$</td>
<td>penalty fees</td>
</tr>
</tbody>
</table>
Duopoly profits equal the markup and consumers’ true expected utility is \( U = S_{FB} - \mu \).

The bottom panel of Figure 1 illustrates Proposition 12 by plotting the equilibrium markup as a function of the transportation cost \( \tau \). Dashed lines show the markups relevant for the two possible pricing-regions as well as the markup \( \tau \) that would prevail in the absence of bias or the NFL constraint. The solid bold line shows the equilibrium markup, which is increasing in \( \tau \) within pricing regions but constant between regions. Starting at the right side of the figure, and working leftward as \( \tau \) falls and competition increases, equilibrium begins in region 1 where firms compete on a mixture of fees and markups equal \((1 + Y) \tau\). In this region, penalty fees exceed \(1/\alpha'\), which requires negative base marginal-charges to ensure expected marginal price does not exceed 1. Negative base marginal-charges in turn require positive fixed fees to satisfy NFL. Once penalty fees fall to \(1/\alpha'\), all other fees are zero and firms compete on penalty fees alone. As a result, consumers become discontinuously less price-sensitive and markups are temporarily flat until equilibrium transitions to region 2 where markups equal \((\alpha/\alpha')^2 \tau\).

A sufficient condition for the full-market-coverage assumption in Propositions 11 and 12 is \( \tau < 2v_0/3 \). Comparing Propositions 11 and 12 for \( \tau \in (0, 2v_0/3) \) uncovers the effects of PPR under competition. Without PPR, expected marginal-prices must be no-higher than one which allows penalty fees to be as high as \(1/\alpha'\) when base marginal-charges are zero and higher when base marginal-charges are negative. The important effect of PPR is that it means implementing the efficient allocation (as is optimal) requires every marginal price be at most one, and hence penalty fees be no higher than one. Holding the level of bias fixed (and \( c < \alpha/2 \)), sufficient competition \((\tau \leq (\alpha')^2 (\alpha - 2c)/\alpha)\) implies that this does not matter because firms choose to offer sufficiently high perceived-expected-utility levels that penalty fees must be less than one. As a result, firms offer the same contract and markup regardless of whether or not PPR is implemented. For any higher level of market power \((\tau > (\alpha')^2 (\alpha - 2c)/\alpha)\), however, PPR does constrain firms’ use of penalty fees. Typically this shifts competition towards fees to which consumers are more price-sensitive, thereby intensifying competition and lowering firm markups. This is always the case for severe bias\(^{25}\) \((\alpha'/\alpha < \max \{1/2, (2\alpha - 1)/\alpha^2\})\) but the reverse can be true for intermediate values of \( \tau \) given mild bias \((\alpha'/\alpha > \max \{1/2, (2\alpha - 1)/\alpha^2\})\) as made precise in parts 2 and 3 of Corollary 2.\(^{26}\) The comparison is illustrated for severe and mild biases in top and bottom panels of Figure 2. Part 1 of Corollary 2 holds \( \tau > 0 \) fixed and shows that sufficiently large bias leads

\(^{25}\) The term severe bias is somewhat misleading. As \( \alpha \) approaches 1, a belief \( \alpha' < \alpha \) arbitrarily close to \( \alpha \) will satisfy the condition for severe bias.

\(^{26}\) When bias is mild and \( \tau \in (\tau_1, \tau_2) \), PPR strictly increases equilibrium markups. In this case although keeping penalty fees a surprise is individually optimal for each firm, as an industry group firms would favor PPR regulation.
to arbitrarily high markups and consumer exploitation. Thus while markups are unambiguously reduced for severe bias, for sufficiently high bias this reduction in markups also means an end to consumer exploitation.

Corollary 2 Assume duopoly on a uniform Hotelling line, the NFL constraint, Bernoulli taste shocks, inattentive consumers who underestimate demand ($\alpha' < \alpha$), and $c \in [0, 1)$. Let $\tau < (2/3) v_0$. The market will be fully covered and allocations will be first best with or without PPR.

1. For fixed $\tau > 0$, if bias is sufficiently large ($\alpha'/\alpha$ is sufficiently small) then all consumers are exploited. PPR increases competition, strictly reduces markups, and eliminates consumer exploitation.

2. If bias is severe ($\alpha'/\alpha < \max\{1/2, (2\alpha - 1)/\alpha^2\}$) then PPR weakly reduces markups for all $\tau \geq 0$, and strictly reduces markups for all $\tau > \max\{(\alpha')^2 (\alpha - 2c)/\alpha, 0\}$ (which is for all $\tau > 0$ if $c \geq \alpha/2$).

3. If bias is mild ($\alpha'/\alpha > \max\{1/2, (2\alpha - 1)/\alpha^2\}$) then PPR effects markups as described for severe bias except for intermediate $\tau \in [\tau_1, \tau_2]$, where $\tau_1 = (1 - \alpha + \alpha') / (\alpha - 2c)\alpha'$ and $\tau_2 = 2\alpha (1 - c) / (1 + Y)$. For $\tau \in (\tau_1, \tau_2)$, PPR strictly increases markups.

Proposition 10 and Corollary 2 capture the second main result in the paper: combined with biased beliefs, inattention can cause consumer exploitation which is eliminated by PPR. In the monopoly setting this is a direct result of the fact that PPR constrains the size of penalty fees - precisely those fees which consumers most underestimate the chance of paying. In a competitive setting the result is more indirect. Absent additional constraints on prices, total markups would equal $\tau$ independent of the fraction earned from penalty fees. However, the NFL constraint ensures that fairly-competitive firms compete on marginal charges rather than fixed fees. This leads to markups above $\tau$ because consumers are less price-sensitive to marginal charges, which they underestimate the likelihood of paying. PPR limits the extent that this competition is over penalty fees, and typically forces firms to compete on more salient base marginal charges intensifying competition and protecting consumers. The online appendix shows that CMPR has a similar effect. (Note that eliminating the underlying bias would be even better, making all fees equally salient and lowering markups to $\tau$. However, debiasing consumers completely is likely to be difficult and costly relative to implementing PPR.)
6 Policy Applications

6.1 FCC’s Proposed Bill-Shock Regulation

US cellular-phone customers are typically charged steep penalty fees for exceeding usage allowances, and the variation in usage allowances across calling plans is an essential instrument for encouraging consumers to self-select into different calling plans. On October 14th, 2010, the FCC proposed bill-shock regulation that would require carriers to notify customers, via voice or text alerts, when they are about to exceed plan limits and begin incurring overage charges (FCC 2010). The FCC’s proposed bill-shock regulation has strong support from consumer groups but is opposed by major cellular carriers (Genachowski 2010, Deloney et al. 2011, Wyatt 2010, Stross 2011).

The price-discrimination model in Section 4 provides an explanation both for carriers’ use of surprise penalty fees on typical cellular contracts and for carriers’ opposition to proposed bill-shock regulation. If one believes that cellular-phone customers are unbiased and the cellular market is sufficiently competitive, then Proposition 7 implies that the FCC’s proposed bill-shock regulation would be counterproductive and lower social welfare. Moreover, consumer groups’ strong support for the regulation would be misplaced, since it would harm some consumers. (Their support is nevertheless understandable since the regulation would be unambiguously beneficial to consumers but for the resulting endogenous price changes predicted by the model.)

There are two caveats to this criticism of the FCC’s proposal. First, the US cellular service market may not be fairly competitive as assumed in Proposition 7. As illustrated by Example 1, sufficient market power means that price-posting regulation could increase or decrease social surplus. Second, Grubb (2009) and Grubb and Osborne (2011) show that cellular-phone customers have biased beliefs about their likely usage and in particular are overconfident. While overconfidence is not the same bias modeled in Section 5, it would also cause consumers to underestimate the likelihood of paying penalties and have similar effects. Proposition 9 shows that such biases in beliefs provide a second reason (in addition to market power) that price-posting regulation could raise rather than lower social surplus. (The analysis in Section 5.2 is likely not relevant because cellular-contract fixed fees are typically positive and hence a NFL constraint cannot be binding.)

In sum, the welfare impact of price-posting regulation is ambiguous but there is a clear reason that it could be socially harmful and caution should be applied in adopting the FCC’s proposal.

---

27 The wireless industry trade group, C.T.I.A. - The Wireless Association, argues that proposed bill-shock regulation “violates carriers’ First Amendment protections... against government compelled speech” (Altschul et al. 2011).

28 A third caveat is that the regulation would apply to fees beyond overage charges such as roaming fees which are typically the same across calling plans, and hence not used for price-discrimination purposes or relevant to this theoretical argument. Roaming charges were the target of recently adopted bill-shock regulation in the EU.
Ongoing complementary work by Grubb and Osborne (2011) will try to resolve the ambiguity empirically and make predictions about the effect of proposed bill-shock regulation using counterfactual simulations based on estimated parameters of a structural demand-model.

6.2 Overdraft Fees

Turning to a second application, consider overdraft fees: In 2009, US bank overdraft fee revenues from ATM and one-time debit-card transactions were $20 billion (Martin 2010). Prior to the Fed’s adoption of an opt-in rule, Bank of America and other banks charged high (often $35) overdraft fees on debit and ATM transactions without notifying customers at the point of sale. When the Fed proposed opt-in regulation, banks opposed it.\(^{29}\) Nevertheless, effective August 15, 2010 (July 1, 2010 for new accounts) new Federal Reserve Board rules “prohibit financial institutions from charging consumers fees for paying overdrafts on automated teller machine (ATM) and one-time debit-card transactions, unless a consumer consents, or opts-in, to the overdraft service for those types of transactions” (Federal Reserve Board 2009b). In response, Bank of America chose to stop offering overdraft protection on debit-card transactions, despite the fact that Bank of America is estimated to have earned $2.2 Billion from ATM and debit-card-transaction overdraft fees in 2009 (Sidel and Fitzpatrick 2010). Other major banks have been accused of responding with deceptive marketing campaigns to induce opt-in. For instance, customers filed a federal class-action lawsuit against JPMorgan Chase for such bad behavior in August 2010 (Dinzeo 2010, McCune, Wright, Arevalo and Kim 2010).

The model used throughout the paper is stylized and fits the overdraft application imperfectly. In particular, while quantity is one-dimensional in the model, overdraft fees depend on both dollars spent and the number of transactions. Overdraft fees are only triggered once dollars spent exceed the account balance, but are then charged on a per-transaction basis. Moreover, marginal costs are likely increasing rather than constant. Nevertheless, the model captures important features of the setting. For instance, consider the model with Bernoulli taste-shocks in Section 5.2. We can interpret \(\alpha\) as the probability a consumer wishes to make a purchase with her debit card. Then \(v = 1\) is her value for making the purchase with her debit card, rather than an alternative such as a credit card. Consumers underestimate the likelihood of purchases to be \(\alpha'\) because they are partially-naive beta-delta discounters who not only undersave and overspend due to time

---

\(^{29}\)Prior to regulating overdraft fees, the Federal Reserve solicited public comment. Industry commenters sought to undermine the regulation in every possible way. For instance “industry commenters . . . urged the Board to permit institutions to vary the account terms . . . for consumers who do not opt in [to overdraft protection]” (Federal Reserve Board 2009a). Clearly banks wanted to be able to make declining overdraft protection an expensive account feature.
inconsistency (Laibson 1997) but also underestimate how much they spend due to partial naivete (O’Donoghue and Rabin 2001).

Does Section 4’s model of price discrimination and Proposition 7 imply that the combination of consumer inattention and overdraft fees could be socially valuable by making price discrimination by banks less distortionary? In fact they do not apply. While banks offer different types of checking accounts, prior to the regulation, banks typically charged the same overdraft fees on all accounts (e.g. Bank of America (2010)). Thus heterogeneity in expectations of overdraft usage is typically not an important dimension of self-selection across checking accounts.

Neither the benchmark model nor Section 4’s model of price discrimination explain banks’ widespread use of overdraft fees, failure to notify consumers at the point-of-sale, or aversion to opt-in regulation. A more compelling explanation for these facts is that consumers underestimate the incidence of overdraft fees. Interestingly, Proposition 9 shows that for products with no social value (c > 1) price-posting regulation would end consumer exploitation and increases social surplus by shutting down the market for the add-on. This possibility is an intriguing explanation for Bank of America’s choice to stop a service which had been earning an estimated $2.2 billion per year and JPMorgan Chase’s choice to resort to deceptive marketing tactics to encourage opt-in.30 Proposition 9 also shows that price-posting regulation could lower rather than raise social surplus if overdraft coverage for ATM and debit-card transactions is socially valuable. However, the more important effect of regulation may be to shift surplus from banks to consumers. The recent rise in the use of overdraft fees has coincided with the rise of free checking accounts, which suggest an NFL constraint has been binding in the industry (Burhouse, Cashman, Cordeiro, Critchfield, Lee, Pawelski and Samolyk 2008, Stango and Zinman 2010). Thus Corollary 2 implies that price-posting regulation could force banks to compete on more salient fees, intensify competition, and lower prices to consumers.

The opt-in regulation adopted by the Fed is weaker than price-posting regulation. If (as this paper assumes) consumers are aware of their own inattention, then opt-in regulation should also limit overdraft fees because consumers will opt-out if overdraft fees are too high. Thus opt-in regulation could also limit the degree to which consumers’ bias can be exploited. This is consistent with early analyst predictions that opt-in regulation would dramatically reduce overdraft-fee revenue (Campbell 2009). In fact, opt-in rates have been high (75%) and overdraft-fee revenue has been relatively stable (Benoit 2010). This suggests either that deceptive marketing practices have been successful, or perhaps that consumers are unaware of their own inattention. A consumer who

30 An alternative explanation for Bank of America’s choice is regulatory threat.
believes himself to be attentive should always opt-in. Either way, this experience suggests that opt-in regulation is a poor substitute for price-posting regulation.

7 Conclusion

Price-posting regulation can help inattentive consumers avoid surprise penalty fees. While this is good for consumers holding prices fixed (and hence may attract the support of consumer groups), it is essential for policy evaluation to incorporate firms’ pricing response.

If unbiased consumers with heterogeneous forecasts of their future demand are inattentive, surprise penalty fees become a useful tool for price discrimination. The combination of inattention and surprise penalty fees are socially valuable when firms view them to be a substitute for inefficient quantity distortions. This is always the case in fairly-competitive markets, but may not be true when firms have sufficient market power. Thus, in fairly-competitive markets, price-posting regulation will be socially harmful because firms will continue to price discriminate but they will be forced to impose greater allocative inefficiencies to do so.

When consumers underestimate their future demand, the combination of consumer inattention and surprise penalty fees can be highly profitable for firms. The bias in beliefs make high penalty fees an attractive way for firms and consumers to take opposing sides of a bet. Inattention relaxes incentive constraints that would otherwise limit the size of penalty fees and the resulting bet. As a result, inattention and surprise penalty fees can enable firms to earn more in profit than the entire social surplus from a transaction and even profit from selling a product with negative social value. In these cases consumers are exploited in the sense that doing business with the firm makes them worse off. It is ambiguous whether price-posting regulation will increase or decrease social welfare, but for severely biased beliefs it will transfer surplus from firms to consumers and eliminate consumer exploitation. Regulation is effective even in fairly-competitive markets because (when fixed fees cannot be negative) it shifts competition away from penalty fees towards other more salient fees to which consumers are more price-sensitive, thereby intensifying competition.

The results suggest that regulators should require price-posting for services such as overdraft protection that are not differentially priced to sort consumers across contracts. However regulators should be more cautious about the FCC’s proposed bill-shock regulation and other applications for which penalty fees do help sort consumers across contracts. The limited impact of the Fed’s opt-in regulation on overdraft-fee revenues suggests it is a poor substitute for price-posting regulation, perhaps because consumers are unaware of their own inattention. Such naivete presents an interesting avenue for future work.
Figure 1: Firm markup as a function of transportation cost $\tau$ in a Hotelling duopoly with the no-free-lunch constraint and consumers who receive Bernoulli taste shocks and underestimate demand ($\alpha' < \alpha$). Top panel: attentive consumers. Bottom panel: inattentive consumers. The figure is plotted for $c = 0$, $\alpha = 3/4$, $\alpha' = 1/4$, and $v_0$ sufficiently high for strict full-market-coverage.
Figure 2: Firm markup as a function of transportation cost $\tau$ in a Hotelling duopoly with the no-free-lunch constraint and consumers who receive Bernoulli taste shocks and underestimate demand ($\alpha' < \alpha$). Solid line: attentive consumers. Dashed line: inattentive consumers. Top panel depicts severe bias: $\alpha = 3/4$ and $\alpha' = 1/4$. Bottom panel depicts mild bias: $\alpha = 1/2$ and $\alpha' = 1/3$. In both cases, $c = 0$ and $v_0$ is sufficiently high for strict full-market-coverage.
References


