

# Penalty Pricing: Optimal Price-Posting Regulation with Inattentive Consumers

Michael D. Grubb  
MIT Sloan School of Management  
Cambridge, MA 02142  
mgrubb@mit.edu  
www.mit.edu/~mgrubb

October 27, 2010

## **Abstract**

For many goods and services, such as cellular phone service, electricity, health-care, and debit or credit 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 still uncertain about the marginal price of the next unit. I develop a model of inattentive consumption, derive optimal pricing-policies when consumers are inattentive, and evaluate price-posting regulation requiring firms to publish marginal price at the time of each transaction. When consumers are homogeneous and have unbiased beliefs, inattention has no substantive effect on market outcomes. Otherwise, inattention leads firms to charge penalty fees for high usage. When consumers are heterogeneous ex ante and have unbiased beliefs, inattention and penalty fees increase welfare in sufficiently competitive markets, and price-posting regulation is counterproductive. Under these conditions, cellular phone usage alerts under consideration by the FCC could reduce welfare and harm consumers. When 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.

# 1 Introduction

In many important situations, consumers may be fully aware of the full schedule of marginal charges when making an ex ante decision to sign a contract, but nevertheless, ex post are uncertain about the marginal price of any given transaction. This occurs whenever marginal prices vary with the level of consumption (as they do when firms levy penalty fees for excessive usage) and, due to inattention, consumers are unaware of their past consumption when making additional consumption choices. Note that 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. In each case, inattention would create uncertainty about marginal price at the point of sale.

For example, a cellular phone customer may be fully aware that the first 500 minutes are billed at zero cents a minute and later minutes at a penalty (or "overage") rate of 45 cents a minute. However, he may be uncertain whether the next call will be billed at zero cents or 45 cents per minute, because he does not know how much he has already used the phone. Similarly, a new checking account enrollee may be fully aware that overdraft penalty 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. (Stango and Zinman (2010) find in survey data "that 60% of overdrafters reported overdrafting because they 'thought there was enough money in my account'".)

In each example, firms have the ability to disclose to consumers whether or not a penalty fee is applicable at the point of sale. A mobile phone screen could flash "overage rate applies" before making any call once included minutes are used up. A debit-card processing terminal could ask "Overdraft fee applies. Continue - Yes/No?" before processing transactions on an overdrawn account. Firms' choices not to make this information so readily available suggest that firms benefit from consumer uncertainty about marginal price. Recent regulation of overdraft fees by the Federal Reserve Board 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.

In this paper, I develop a model to answer the following three questions: First, if consumers are inattentive to their own past consumption, do firms profit by charging penalty fees for excessive usage? Second, if so, does price-posting regulation requiring firms to disclose 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? In the process I also answer a fourth question: how do the conclusions depend on consumer heterogeneity and consumer biases?

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, 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 making plans. 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. This 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.)

In Section 3, I develop a base model which serves as a benchmark for the rest of the paper. The base model assumes that at the time of contracting consumers are homogeneous (so there is no scope for price discrimination) and consumers have correct beliefs (so there are no biases to exploit). For simplicity, I assume throughout the paper that there are only two consumption opportunities. As a result, the effect of price-posting regulation is to make inattentive consumers attentive. (With more consumption opportunities, greater disclosure would be needed to make inattentive consumers attentive.) To analyze the effect of price-posting regulation, I therefore solve for equilibrium prices under two conditions: first with attentive consumers and second with inattentive consumers.

Under the base model assumptions, the primary 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. Firms would find it optimal to set marginal price equal to marginal cost and not charge any penalty fees for excessive usage, regardless of whether consumers are attentive or inattentive. However these prices are uniquely optimal only with price-posting regulation. Thus price-posting regulation could induce firms to eliminate penalty fees but compensate with other charges. This captures the 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).<sup>1</sup> However, the result relies heavily on the assumption of homogeneity. Moreover, it clearly does not explain the widespread use of penalty fees, choices by firms not to disclose marginal price or alert consumers to penalty fees at the point of sale, or firms' expressed aversion to regulation which would help consumers avoid purchases that trigger penalties (Federal Reserve Board 2009a).

One reason that penalty fees are used in practice may be that they are useful for discriminating between consumers with heterogeneous expectations about their own future demand for the

---

<sup>1</sup>Jamie 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).

product or service. For instance, cellular phone overage fees are not only designed to generate revenue directly (Grubb (2009) finds 22 percent of revenues were from overage charges), but also to encourage consumers who anticipate high demand to self select into larger calling plans. Section 4 enriches the base 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, 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. Perhaps more surprising is the fact that inattention may 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 in fairly-competitive markets with heterogeneous inattentive-consumers who have correct beliefs, penalty fees are socially valuable and price-posting regulation is counter productive.

The paper's first main result could suggest caution in adopting bill-shock regulation under consideration by the FCC, which would require carriers to alert customers of rapidly accumulating fees by text message (FCC 2010). A fundamental part of cellular-phone-service pricing is separating consumers with different expectations of usage among different contracts with different allowances of included minutes. If one believes that cellular phone customers have correct beliefs and the cellular market is sufficiently competitive, then inattention is good for welfare - and price-posting regulation would be counter-productive. But note that these assumptions about beliefs and competition may not be valid. In fact, evidence shows that cellular customers have biased beliefs (Grubb 2009, Grubb and Osborne 2010) and it is not obvious that the industry is highly competitive. As a result the welfare impact of price-posting regulation is ambiguous.<sup>2</sup>

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). Effective July 1, 2010 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"

---

<sup>2</sup>Moreover, 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.

(Federal Reserve Board 2009b). Does Section 4’s model of heterogeneous consumers with correct beliefs suggest this regulation is welfare reducing? In fact it does not apply. Prior to the regulation, banks typically did not differentiate checking accounts by varying overdraft fees. For instance, before ending overdraft protection on ATM and debit-card transactions, Bank of America offered a variety of checking accounts, but offered the same overdraft fee schedule on all of them (Bank of America 2010). Thus heterogeneity in expectations of overdraft usage is typically not an important dimension of self-selection across checking accounts.

Since neither the base model nor Section 4’s model of price discrimination explain banks’ widespread use of overdraft fees, I explore a more compelling alternative: that consumers underestimated the incidence of overdraft fees. There is substantial evidence that consumers often have biased beliefs at the time of contracting (Ausubel and Shui 2005, DellaVigna and Malmendier 2006, Grubb 2009). Section 5 enriches the base model by assuming that consumers underestimate their own future demand. Firms can profit from this bias by raising marginal prices that consumers underestimate the likelihood of paying. However, attentive consumers who underestimate their own value for a service cannot be exploited in the sense that they can never be induced to pay more than their average value for a product or service. In contrast, the paper’s second main result is that if consumers are both inattentive and underestimate their own values for a service, they can be grossly exploited and firms can extract profits orders of magnitude higher than the total surplus using penalty fees. This is true even in fairly competitive markets, as the combination of penalty fees and consumer inattention can significantly soften price competition.

The total welfare effects of price-posting regulation are ambiguous, but may be second order relative to the distributional effects. Regulation requiring the posting of the marginal price at the point of each transaction would mitigate the consumer welfare losses due to biased beliefs and ensure that consumers are not exploited. The redistribution of surplus from firms to consumers involved in ending exploitation could be orders of magnitude larger than the total surplus generated by the market. In fact, it is possible for a service with zero or negative social value to be sold at high profit (and high consumer loss) prior to regulation, but be efficiently shut down by price-posting regulation. This may explain the fact that Bank of America, the bank with the largest overdraft fee revenue in 2009 (estimated to be \$2.2 billion per year by a Sandler O’Neill + Partners report (Sidel and Fitzpatrick 2010)), responded to the Fed’s new ”opt-in” requirement by ending overdraft protection for one-time debit-card transactions (Martin 2010).<sup>3</sup>

---

<sup>3</sup>Bank of America still offers an alternative overdraft protection service that transfers money from a linked savings or credit card account for a \$10 fee. This opt-in service has always been available, but was typically unused by customers who failed to opt-in and ended up paying the higher \$35 overdraft fees that are now subject to regulation.

This paper considers settings where consumers are inattentive to their own past consumption and shows that firms optimally charge penalty fees for excessive usage to take advantage of such inattention. In such settings, the results suggest that regulators should require price-posting for products such as overdraft protection that are not differentially priced to sort consumers into different contracts. However, regulators should be more cautious for products such as cellular-phone calls that are an important dimension of consumers' self selection across contracts. In particular, it predicts that the Federal Reserve Board's opt-in rule for overdraft fees on debit transactions could strongly benefit consumers, but that the bill shock regulation under consideration by the FCC has the potential to be counter productive.

The paper proceeds as follows. Section 2 discusses related literature. Section 3 introduces the base 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. Finally Section 6 concludes. All proofs not included in the text are provided in the appendix.

## 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 assumption is made in both empirical work (Cardon and Hendel 2001, Reiss and White 2005, Gaynor, Shi, Telang and Vogt 2005, Lambrecht, Seim and Skiera 2007, Huang 2008) and throughout the theoretical literatures on nonlinear pricing (Wilson 1993) and two-period sequential screening (Baron and Besanko 1984, Riordan and Sappington 1987, Miravete 1996, Courty and Li 2000, Miravete 2005, Grubb 2009). When applied to settings in which consumers make many separate consumption decisions within in a billing period, the implicit assumption is that consumers have perfect foresight to predict all these individual choices at the start of the billing period. This is usually implausible 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, if firms charge penalty fees for excessive consumption, attentive consumers must solve a dynamic programming problem similar to the airline revenue management problem surveyed by McAfee and te Velde (2007). A key feature of the solution is that attentive consumers reduce consumption after penalty fees are triggered (equation (1)). Using

detailed call-level data, Grubb and Osborne (2010) find no evidence of this behavior among cellular phone subscribers, suggesting that they are in fact inattentive to their own past usage within the billing cycle. In the context of checking-account overdraft-fees, Stango and Zinman (2009) find even more direct evidence of inattention: the median consumer could avoid more than 60% of overdraft charges by using alternative cards (checking or credit) with available liquidity. Using a different data set, 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'".<sup>4</sup>

Formally, the inattentive consumer's decision problem analyzed in Section 3 exhibits Piccione and Rubinstein's (1997b) *absentmindedness*. Subject to the information constraint imposed by absentmindedness, consumers behave optimally. Psychology experiments demonstrate that attention is a limited resource (Broadbent 1958). DellaVigna (2009) surveys recent work in economics which examines inattention to shipping costs, nontransparent taxes, financial news, and other information. I show that inattentive consumers purchase all units valued above the endogenous expected marginal price.

Liebman and Zeckhauser (2004) analyze optimal pricing given alternative deviations from unbounded rationality by consumers faced with multi-part tariffs. Liebman and Zeckhauser's (2004) deviations, which they dub "ironing" and "spotlighting", are based on decision errors rather than an information limitation. Liebman and Zeckhauser's (2004) first model (ironing) is static. It assumes that consumers make a single quantity choice and confuse the average price with the marginal price. Liebman and Zeckhauser's (2004) second model (spotlighting) is dynamic. It assumes consumers make consumption decisions one unit at a time and myopically base their consumption choices on the marginal price of the current unit.

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 (Diamond 1971), purchase a base product such as a printer (Ellison 2005), select a hotel (Gabaix and Laibson 2006), or open a checking account (Bubb and Kaufman 2009). As a result, marginal fees for add-on products or services are set at monopoly levels in spite of competition or the use of two-part tariffs, either of which would normally lead to marginal cost pricing.

---

<sup>4</sup>Stango 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 Agrawal's finding that accruing one credit card late penalty fee reduces the likelihood of incurring one in the following month.

Section 4's model of price discrimination is related to the literature on sequential screening (Baron and Besanko 1984, Riordan and Sappington 1987, Miravete 1996, Courty and Li 2000, Miravete 2005, Grubb 2009, Pavan, Segal and Toikka 2009), in which consumers first choose from a menu of contracts and then make quantity choices after the arrival of more information. Both Courty and Li (2000) and Pavan et al. (2009) model monopoly pricing when consumers have zero outside options. Under this market condition, the solution to my benchmark model with attentive consumers corresponds to a repetition of the Courty and Li (2000) solution and is nearly a special case of Pavan et al.'s (2009) model, although I assume two ex ante types at the contracting stage rather than a continuum. I go further, however, by solving my attentive model under more general market conditions: monopoly with heterogeneous outside options and duopoly.

Although I am unaware of other work on competitive sequential-screening, there is related work on competitive static-nonlinear-pricing, for which Stole (2007) provides an excellent survey. In particular, I incorporate competition following a similar approach to that taken by Armstrong and Vickers (2001) and Rochet and Stole (2002). Armstrong and Vickers (2001) and Rochet and Stole (2002) both contain versions of the same result: that sufficient competition in nonlinear price-schedules leads to two-part-tariff pricing at marginal cost and first-best allocations. This is a knife-edge result, which depends on the assumption that the optimal markup (ignoring incentive constraints) is exactly the same for all customer segments. I find an analogous result in my attentive model with competitive sequential screening. The first-best-allocation result (although not the two-part-tariff-pricing result) also extends to competitive sequential-screening with inattentive consumers, but in this case is more general as it holds even if optimal markups differ across customer segments.

The model explored in Section 5 assumes that at the time of contracting consumers underestimate their demand for the good or service for sale. Such consumers exhibit similar behavior to naive quasi-hyperbolic-discounters (DellaVigna and Malmendier 2004, Eliaz and Spiegler 2006). There is a small related literature on optimal pricing when attentive consumers have biased beliefs (Eliaz and Spiegler 2006, Gabaix and Laibson 2006, Sandroni and Squintani 2007, Eliaz and Spiegler 2008, Grubb 2009, Bubb and Kaufman 2009). A common finding is that demand underestimation, due either to biased beliefs (Gabaix and Laibson 2006, Eliaz and Spiegler 2006, Grubb 2009), or naive quasi-hyperbolic-discounting (DellaVigna and Malmendier 2004), leads to high marginal prices above marginal cost.

In competitive markets, firms offset high marginal fees with lower fixed fees (Gabaix and Laibson 2006, Grubb 2009). I show that incorporating a plausible *no-free-lunch* constraint means that biased beliefs can soften price competition in such markets, by forcing firms to compete on add-on fees

rather than fixed fees. Moreover, inattention exacerbates the softening of competition due to biased beliefs and makes consumers even worse off. 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 the sense that they always achieve at least their outside options.<sup>5</sup> In contrast, I show that inattention allows consumers to be exploited and can drastically exacerbate the cost of biased beliefs to consumers, even in fairly competitive markets.

### 3 Base Model: ex ante homogeneous and unbiased consumers

This section develops the underlying model structure used throughout the paper. The base assumptions that are relaxed later are that consumers have correct beliefs and are 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. The primary result in this section is an equivalence result: neither inattention nor price-posting regulation affect substantive market outcomes.

#### 3.1 Model

Game players are mass 1 of consumers and  $N \geq 1$  firms. Consumers privately learn a vector of  $N$  firm-specific (brand) taste shocks  $\mathbf{x}$  that is mean zero (and 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 their 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  $F$

---

<sup>5</sup>In 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.

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  $P^i(q_1, q_2)$  to firm  $i$ , as a function of past quantity choices. Firm  $i$ 's offered contract can be any deterministic price schedule:<sup>6</sup>

$$P^i(q_1, q_2) = p_0^i + p_1^i q_1 + p_2^i q_2 + p_3^i q_1 q_2,$$

characterized by the vector of prices  $\mathbf{p}^i = (p_0^i, p_1^i, p_2^i, p_3^i)$ .

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 choices  $q_t$ , private taste shocks  $v_t$ , and payment to the firm:

$$u(\mathbf{q}, \mathbf{v}) = v_0 + q_1 v_1 + q_2 v_2 - P^i(q_1, q_2).$$

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

A consumer's total expected payoff,  $U^i + x^i$ , includes brand taste  $x^i$ . Thus, fraction  $G(U^i; U^{-i})$  of consumers of type  $s$  buy from firm  $i$  if firm  $i$  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 are equal to payments less fixed costs (normalized to zero) and marginal cost  $c \in (0, 1)$  per unit served. Thus firm  $i$ 's expected profits are

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

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

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

---

<sup>6</sup>See Rochet and Stole (2002) for an insightful discussion of this assumption.

### 3.2 Consumer Strategies

The first step in analyzing the game is to solve the consumers problem. As I do so below, I suppress the firm  $i$  superscript from my notation.

The optimal decision rule for an attentive consume 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  $v^*(q^{t-1}, t)$  which 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

$$v_2^*(q_1) = p_2 + p_3 q_1, \quad (1)$$

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

$$v_1^* = p_1 + (1 - F(p_2 + p_3)) p_3 + \int_{p_2}^{p_2 + p_3} (v_2 - p_2) f(v_2) dv_2. \quad (2)$$

The intuition is that  $v_1^*$  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,  $\int_{p_2}^{p_2 + p_3} (v_2 - p_2) f(v_2) dv_2$ . Note that the period two threshold is a function of past usage if and only if the penalty fee is nonzero ( $p_3 \neq 0$ ). Formally, the attentive consumers optimal consumption rule can be written as:

$$q^A(\mathbf{v}; \mathbf{p}) = (1_{v_1 \geq v_1^*}, 1_{v_1 \geq v_1^*} 1_{v_2 \geq (p_2 + p_3)} + 1_{v_1 < v_1^*} 1_{v_2 \geq p_2}). \quad (3)$$

An inattentive consumer is one who cannot condition her strategy on the date  $t$  or on past usage  $q^{t-1}$  because she does not keep track of these variables. She exhibits imperfect recall. (Note that while everyone knows the calendar date, it takes more effort to track the date within ones billing cycle for any particular service.) Otherwise, I assume that inattentive consumers are entirely rational and, in particular, are aware of their own inattention and plan accordingly.<sup>7</sup> The optimal strategy of an inattentive consumer is also to consume if and only if value  $v_t$  is above a threshold  $v^*$ , but an inattentive consumer's threshold is simply a constant, since it cannot be conditioned on  $t$  or  $q^{t-1}$ .

Formally, the consumer's decision problem exhibits Piccione and Rubinstein's (1997b) *ab-sentmindedness*. Piccione and Rubinstein's (1997b) paradoxical absent minded driver example

---

<sup>7</sup>Inattentive consumers are unaware of past shocks  $v^{t-1}$ , usage  $q^{t-1}$ , or the current date  $t$ . They are aware of this limitation, the distribution of their taste shocks  $F$ , and can remember their chosen consumption thresholds  $v^*$ . Assuming that consumers do not attend to the date makes the model more tractable, but does not qualitatively change the primary welfare results.

shows that analysis of such decision problems can be problematic, and there are different views on how to handle them (Piccione and Rubinstein 1997b, Piccione and Rubinstein 1997a, Gilboa 1997, Battigalli 1997, Grove and Halpern 1997, Halpern 1997, Lipman 1997, Aumann, Hart and Perry 1997a, Aumann, Hart and Perry 1997b). In particular, optimal strategies need not be time consistent. In this case, however, there is no problem.<sup>8</sup> Consumers' optimal thresholds from an ex ante planning view point are time consistent and also optimal during execution. Hence the standard Bayesian Nash Equilibrium is an appropriate solution concept. Note that I assume that consumers plan ahead and choose a consumption strategy at the time they sign a contract. This rules out suboptimal equilibria that exist in the game modeled between multiple selves.

A feasible inattentive strategy is a function  $b(v_t)$  which describes a purchase probability for each valuation  $v_t$  to be implemented at all  $t > 0$  independently of date or past usage. 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  $v^*$ :  $\mathbf{q}^I(\mathbf{v}; \mathbf{p}) = (1_{v_1 \geq v^*}, 1_{v_2 \geq v^*})$ . The optimal consumption threshold  $v^*$  is equal to the expected marginal price conditional on purchasing in the current period and satisfies the first order condition:*

$$v^* = \frac{p_1 + p_2}{2} + (1 - F(v^*))p_3. \quad (4)$$

*Equation (4) 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 (4) 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.*

**Proof.** Assume that at the contracting stage a consumer plans to take strategy  $b^*$  but later considers a one time deviation to strategy  $b$ . At the planning stage, the consumer chooses  $b^*$  to maximize  $U(b^*, b^*)$ :

$$U(b^*, b^*) = v_0 - p_0 + 2 \int_0^1 \left( v - \frac{p_1 + p_2}{2} \right) b^*(v) dF(v) - p_3 \left( \int_0^1 b^*(v) dF(v) \right)^2.$$

The plan is time consistent if, when considering a one time deviation to strategy  $b$  at the imple-

---

<sup>8</sup>In Piccione and Rubinstein's (1997b) paradoxical absent minded driver example, time inconsistency arises because past decisions to exit or stay on the free-way determine whether or not the decision is faced again in the future. Thus simply arriving at a free-way exit is informative about which exit it is. In this paper, a consumer always has exactly two consumption choices to make, and hence being presented with a choice is entirely uninformative about past purchasing.

mentation stage, the resulting payoff  $U(b^*, b)$  is maximized at  $b = b^*$ .

$$U(b^*, b) = v_0 - p_0 + \int_0^1 \left( v - \frac{p_1 + p_2}{2} \right) b^*(v) dF(v) + \int_0^1 \left( v - \frac{p_1 + p_2}{2} \right) b(v) dF(v) - p_3 \left( \int_0^1 b^*(v) dF(v) \right) \left( \int_0^1 b(v) dF(v) \right).$$

Inspection of the first order conditions for point-wise maximization at the planning and implementation stages,

$$\frac{dU(b^*, b)}{db(v)} = \frac{1}{2} \frac{dU(b^*, b^*)}{db^*(v)} = f(v) \left( v - \frac{p_1 + p_2}{2} - p_3 \int_0^1 b^*(v) dF(v) \right),$$

shows that the optimal strategy at the planning stagey is a threshold strategy satisfying equation (4) and that it is time consistent. A non-negative penalty fee is sufficient for  $\frac{d}{dv^*} \left( \frac{1}{f(v^*)} \frac{dU(v^*)}{dv^*} \right) = -2(1 + f(v^*)p_3)$  to be strictly negative, which in turn is a sufficient second order condition for the consumer's maximization problem. ■

Note that given fixed prices and a positive penalty fee, equation (4) implies that  $v^*$  and  $(1 - F(v^*))$  both increase as the distribution of values  $F$  increases in a first order stochastic dominance 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.<sup>9</sup>

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

I also consider an alternative regulation which prohibits the use of penalty fees.

---

<sup>9</sup>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  $q^{t-1}$  yields the same predictions as the current model if penalty fees are exogenously or endogenously restricted to be not too large. 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 in and of itself would endogenously limit the size of penalty fees, but would not qualitatively affect the primary welfare predictions.

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

It will be a recurring result throughout the paper that firms optimally offer attentive consumers two-part tariffs with zero penalty fees. Thus 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.

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 price-posting regulation 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 constant-marginal-price regulations would both restrict equilibrium prices but have no effect on allocations, firm profits, or consumer surplus.*

The equivalence result in Proposition 2 captures the 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). However, the result relies heavily on the joint assumptions of 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 banks' choices not to voluntarily post transaction prices, by alerting consumers at the point of sale whether a given transaction will result in an overdraft charge, nor their expressed aversion to regulatory intervention (Federal Reserve Board 2009a).<sup>10</sup> Similarly, Proposition 2 does not explain why cellular phone companies do not actively alert consumers to accruing overage charges, as the FCC is now considering making a requirement.

---

<sup>10</sup>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.

## 4 Unbiased but ex ante Heterogeneous Consumers

In this section, I relax the assumption of ex ante homogeneity imposed in the base model, and show that 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 price-posting regulation does affect substantive market outcomes. In particular, price-posting regulation will be counter-productive in fairly competitive markets.

### 4.1 Model

Game players are mass 1 of consumers who have unbiased beliefs, but are heterogeneous ex ante, and  $N \geq 1$  firms. At the contracting stage ( $t = 0$ ), each consumer privately receives one of two private signals  $s \in \{L, H\}$ , where  $\Pr(s = H) = \beta$ . In addition, consumers privately learn a vector of  $N$  firm-specific taste shocks  $\mathbf{x}$  that is mean zero conditional on  $s$ . Each firm  $i$  simultaneously offers a menu with a choice of two contracts,  $s \in \{L, H\}$ . 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,  $t \in \{1, 2\}$ , consumers privately learn a taste shock  $v_t$ , which measures a consumer's value for a unit of add-on service. Conditional on receiving signal  $s$ , a consumer's consumption taste shocks  $v_t$  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):  $F_L(v) \geq F_H(v)$ . After learning their taste shocks  $v_t$ , 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  $P^i(q_1, q_2; \hat{s})$  to firm  $i$ , as a function of past quantity choices and the chosen contract  $\hat{s}$ :

$$P^i(q_1, q_2; \hat{s}) = p_{0\hat{s}}^i + p_{1\hat{s}}^i q_1 + p_{2\hat{s}}^i q_2 + p_{3\hat{s}}^i q_1 q_2,$$

characterized by the vector of prices  $\mathbf{p}_{\hat{s}}^i = (p_{0\hat{s}}^i, p_{1\hat{s}}^i, p_{2\hat{s}}^i, p_{3\hat{s}}^i)$ .

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 choices  $q_t$ , private taste shocks  $v_t$ , and chosen contract  $\hat{s}$ :

$$u(\mathbf{q}, \mathbf{v}, \hat{s}) = v_0 + q_1 v_1 + q_2 v_2 - P^i(q_1, q_2; \hat{s}).$$

Conditional on signing a contract with prices  $\mathbf{p}$ , a consumer's optimal consumption strategy can be described by a function mapping valuations to quantity choices:  $\mathbf{q}(\mathbf{v}; \mathbf{p})$ . The expected base utility of a consumer of type  $s$  who chooses contract  $\hat{s}$  from firm  $i$  at time zero and makes optimal

consumption choices thereafter is  $U_{s\hat{s}}^i = E[u(\mathbf{q}(\mathbf{v}; \mathbf{p}_s^i), \mathbf{v}, \hat{s}) | s]$ . Define  $U_s^i \equiv U_{ss}^i$  to be the expected base utility of a consumer who chooses the intended contract from firm  $i$ . Similarly, let  $S_s = v_0 + E\left[\sum_{t=1}^2 (v_t - c) q_t(\mathbf{v}; \mathbf{p}_s^i) | s\right]$  be the expected 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  $U_s^{-i}$ :

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

Firm profits per consumer are equal to payments less fixed costs (normalized to zero w.l.o.g.) and marginal cost  $c \in [0, 1)$  per unit served. Thus, suppressing competitors offers  $U_s^{-i}$  and firm  $i$  superscripts from the notation, the firm's expected profit maximization problem is:

$$\begin{aligned} & \max_{\mathbf{P}_L, \mathbf{P}_H} ((1 - \beta) G_L(U_L) (S_L - U_L) + \beta G_H(U_H) (S_H - U_H)) \\ & \text{s.t. } U_s \geq U_{s\hat{s}} \quad \forall s, \hat{s} \in \{L, H\}. \end{aligned}$$

This initial statement of the firm's problem encompasses both attentive and inattentive cases. They vary only by the consumers' optimal consumption rule  $\mathbf{q}(\mathbf{v}; \mathbf{p})$ , which is given as a function of prices by equation (3) in the attentive case but by Proposition 1 in the inattentive case.

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 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):  $G_L(x) = G_H(x) = \mathbb{1}_{x \geq 0}$ , which captures a monopolist serving customers with zero outside option. (2) Heterogeneous outside options (HOO):  $G_s(x)$  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.<sup>11</sup> Define  $\mu_s^* = S_s^{FB} - \hat{U}_s$  to be the optimal markup given first best allocations and ignoring ex ante incentive constraints ( $\hat{U}_s \equiv \arg \max_U G_s(U) (S_s^{FB} - U)$ ).

---

<sup>11</sup>The stronger assumption,  $\frac{G_s(U_s)}{g_s(U_s)}$  increasing, is equivalent to  $G(U)$  log concave, and pass-through-rate less than 1.

In the first case (ZOOM),  $\hat{U}_s = 0$  and  $\mu_s^* = S_s^{FB}$ . In the latter case (HOO),  $\mu_s^* = \frac{G_s(\hat{U}_s)}{g_s(\hat{U}_s)}$  where  $\hat{U}_s$  uniquely satisfies  $S_s^{FB} = \hat{U}_s + \frac{G_s(\hat{U}_s)}{g_s(\hat{U}_s)}$ . 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 benchmark

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 et al. (2009). Because values are conditionally independent,  $p_{3L} = p_{3H} = 0$ , the solution reduces to a repetition of the Courty and Li (2000) solution. My assumptions do not fit exactly within Pavan et al.'s (2009) framework (because I assume period-zero types are discrete rather than continuous), and I also need to extend the results to allow for heterogeneous outside-options. Incorporating heterogeneous outside-options is essential so that I can move beyond monopoly pricing and analyze imperfect competition.

A precise statement of the firm's problem with attentive consumers is given in Appendix A.3. Proposition 3 characterizes the solution to a single firm's problem, and Corollary 2 applies the result to a Hotelling duopoly.

**Proposition 3** *Given  $U_s + \frac{G_s(U_s)}{g_s(U_s)}$  increasing and  $c > 0$ : The firm offers a menu of two two-part tariffs,  $P_s(q_1, q_2) = p_{0s} + v_s^A(q_1 + q_2)$ . Penalty fees are zero,  $p_{3L} = p_{3H} = 0$ , marginal prices are date independent,  $p_{1s} = p_{2s} = v_s^A$ , and profits are:*

$$\begin{aligned} \Pi(U_L, U_H, v_L^A, v_H^A) &= (1 - \beta) G_L(U_L) \left( 2 \int_{v_L^A}^1 (v - c) f_L(v) dv - U_L \right) \\ &\quad + \beta G_H(U_H) \left( 2 \int_{v_H^A}^1 (v - c) f_H(v) dv - U_H \right). \end{aligned}$$

There are three cases. (1) If  $\mu_L^* = \mu_H^*$ , then a single marginal cost contract is offered and both types receive the first best allocation,  $v_L^A = v_H^A = c$ . (2) If  $\mu_H^* > \mu_L^*$  then the high type receives the first best allocation  $v_H^A = c$ , while the low type's allocation is distorted downwards below first best,  $v_L^A > c$ . The triple  $\{U_L, U_H, v_L^A\}$  maximize firm profits such that  $U_H = U_L + 2 \int_{v_L^A}^1 (F_L(v) - F_H(v)) dv$  (IC-H) and  $v_L^A$  must satisfy the first order condition:

$$v_L^A = c + \frac{\beta}{1 - \beta} \frac{F_L(v_L^A) - F_H(v_L^A)}{f_L(v_L^A)} \frac{-\partial \Pi / \partial U_H}{\beta G_L(U_L)}.$$

For the zero-outside-option monopoly case, this reduces to the Courty and Li (2000) solution:

$$v^{CL} = c + \frac{\beta}{1 - \beta} \frac{F_L(v^{CL}) - F_H(v^{CL})}{f_L(v^{CL})}.$$

(3) If  $\mu_H^* < \mu_L^*$ , then the low type receives the first best allocation  $v_L^A = c$ , while the high type's allocation is distorted upwards above first best,  $v_H^A < c$ . The triple  $\{U_L, U_H, v_H^A\}$  maximize firm profits such that  $U_H = U_L + 2 \int_{v_H^A}^1 (F_L(v) - F_H(v)) dv$  (IC-L) and  $v_H^A$  must satisfy the first order condition:

$$v_H^A = c - \frac{1 - \beta}{\beta} \frac{F_L(v_H^A) - F_H(v_H^A)}{f_H(v_H^A)} \frac{-\partial \Pi / \partial U_L}{(1 - \beta) G_H(U_H)}.$$

Optimal pricing with attentive consumers characterized by Proposition 3 is essentially a repetition of the Courty and Li (2000) solution modified for heterogeneous outside-options by an additional term in the first-order conditions. There are two features of the solution worth highlighting. First, Proposition 3 shows that optimal pricing for attentive consumers requires a constant marginal price and can be implemented as a two part tariff as a function of total usage:  $P_s(q_1, q_2) = p_{0s} + v_s^A(q_1 + q_2)$ . The constant marginal price result from Proposition 3 implies that nondisclosure of  $\{t, q^{t-1}\}$  is weakly optimal when consumers are inattentive:

**Corollary 1** *It is weakly optimal to conceal  $\{t, q^{t-1}\}$  from inattentive consumers.*

**Proof.** If a firm chose to make inattentive consumers attentive by disclosing  $\{t, q^{t-1}\}$  at the point of sale, then Proposition 3 shows that the optimal tariff has constant marginal price. Given a constant marginal price, inattentive and attentive consumers behave identically. Thus the same outcome could have been achieved by offering the same tariff and concealing  $\{t, q^{t-1}\}$ . ■

Analysis in the following section shows that nondisclosure is in fact strictly optimal whenever consumers are inattentive and unconstrained optimal-markups differ across low and high-types. The second feature of the attentive solution worth highlighting is that allocations are first best only when unconstrained optimal-markups are identical for low and high types. As Corollary 2 shows, this implies that allocations are only efficient in a Hotelling duopoly when both market segments have identical transportation costs.

**Corollary 2** *Let duopolists with strictly positive 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. (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 tariff 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 below first best. For  $\tau_H < \tau_L$ , low-types receive first best, while high types' allocation is distorted upwards.

The knife-edge efficiency-result in Proposition 3 and Corollary 2 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. In contrast, the striking result in the next section is that firms can charge different markups to different segments without distorting allocations if consumers are inattentive.

### 4.3 Inattentive case

Let  $v_{s\hat{s}}^*$  be the optimal consumption threshold of an inattentive consumer of type  $s$  who chooses contract  $\hat{s}$ , and let  $v_s^* = v_{ss}^*$ . The first order condition for  $v_{s\hat{s}}^*$  is a natural extension of equation (4):

$$v_{s\hat{s}}^* = \frac{p_{1\hat{s}} + p_{2\hat{s}}}{2} + p_{3\hat{s}} (1 - F_s(v_{s\hat{s}}^*)). \quad (5)$$

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

$$U_{s\hat{s}} = v_0 - p_{0\hat{s}} + 2 \int_{v_{s\hat{s}}^*}^1 v dF_s(v) - (p_{1\hat{s}} + p_{2\hat{s}}) (1 - F_s(v_{s\hat{s}}^*)) - p_{3\hat{s}} (1 - F_s(v_{s\hat{s}}^*))^2, \quad (6)$$

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

$$S_s = \int_{v_s^*}^1 (v - c) dF_s(v). \quad (7)$$

Define  $\bar{p}_s = (p_{1s} + p_{2s})/2$ . When consumers are inattentive, any pair  $\{p_{1s}, p_{2s}\}$  which have the same average  $\bar{p}_s$  are equivalent, both in terms of allocations and surplus division. I focus on symmetric pricing  $p_{1s} = p_{2s} = \bar{p}_s$ , 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, it is convenient to think of the firm choosing offered utility levels  $U_s$  rather than setting fixed fees  $p_{0s}$ . In this case

the base fee  $p_{0s}$  is given by equation (8):

$$p_{0s} = -U_s + v_0 + 2 \int_{v_s^*}^1 v dF_s(v) - 2\bar{p}_s (1 - F_s(v_s^*)) - p_{3s} (1 - F_s(v_s^*))^2. \quad (8)$$

Second, it is convenient to 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<sup>12</sup> for  $\bar{p}_s$  to satisfy the first order condition:

$$\bar{p}_s = v_s^* - p_{3s} (1 - F_s(v_s^*)). \quad (9)$$

The firm's problem can be written as:

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

where  $U_{s\hat{s}}$ ,  $S_s$ ,  $p_{0s}$ , and  $\bar{p}_s$  are given by equations (6) through (9).

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 (9) 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 4** *Increasing  $p_{3s}$  weakly relaxes both ex ante incentive constraints. It is weakly optimal to choose non-negative penalties  $p_{3s}$  as large as possible.*

**Proof.** Substituting equations (8-9) into equation (6) yields

$$U_{s\hat{s}} = U_{\hat{s}} + 2 \int_{v_{s\hat{s}}}^1 (v - v_{\hat{s}}) dF_s(v) - 2 \int_{v_{\hat{s}}}^1 (v - v_{\hat{s}}) dF_{\hat{s}}(v) - p_{3\hat{s}} (F_{\hat{s}}(v_{\hat{s}}) - F_s(v_{s\hat{s}}))^2. \quad (10)$$

By the envelope condition:

$$\frac{d}{dp_{3\hat{s}}} U_{s\hat{s}} = \frac{\partial}{\partial p_{3\hat{s}}} U_{s\hat{s}} = - (F_{\hat{s}}(v_{\hat{s}}) - F_s(v_{s\hat{s}}))^2 \leq 0. \quad (11)$$

---

<sup>12</sup>Up to the fact that all thresholds above one are equivalent, and all thresholds below zero are equivalent.

■

Proposition 4 suggests that the solution to the firm's problem could involve unreasonably high penalty fees. There are many forces which could endogenously limit penalty fees, some of which I discuss in Section 4.4. For simplicity, I exogenously impose one of two restrictions. Either I impose a cap on the penalty fees, or I require marginal prices to be non-negative. Both restrictions can be expressed as upper bounds on penalty fees:  $p_{3s} \leq h_s(v_s)$ . A cap on penalty fees corresponds to  $h_s(v_s) = p^{\max} > 0$ , while non-negative marginal prices correspond to  $h_s(v_s) = v_s / (1 - F_s(v_s))$ . Notice that all prior results and statements remain true with this addition to the problem.<sup>14</sup>

I solve the firm's problem separately for three cases. In each case I relax one or both ex ante incentive compatibility constraints and then confirm that the relaxed solution satisfies the ignored constraints and therefore solves the original problem. In the attentive problem, both ex ante incentive constraints can be relaxed and 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. This can be achieved because strictly positive penalty fees relax the ex ante incentive constraints when consumers are inattentive.

To state the proposition, first define  $X_H \equiv 2 \int_c^{v_{HL}} (v - c) dF_H(v) + (v_{HL} - c) (F_L(c) - F_H(v_{HL})) > 0$  where  $v_{HL}$  uniquely satisfies  $v_{HL} = c + h_L(c) (F_L(c) - F_H(v_{HL}))$  and  $X_L \equiv 2 \int_{v_{LH}}^c (c - v) dF_H(v) - (c - v_{LH}) (F_L(v_{LH}) - F_H(c)) > 0$  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 5** *Assume (1)  $U_s + \frac{G_s(U_s)}{g_s(U_s)}$  increasing and (2) either penalty fees  $p_{3s}$  are exogenously restricted to be less than  $p^{\max}$  ( $h_s(v_s) = p^{\max}$ ), or marginal prices are exogenously restricted to be non-negative ( $h_s(v_s) = v_s / (1 - F_s(v_s))$ ). If  $\mu_H^* \neq \mu_L^*$ , then the firm strictly prefers to conceal  $(t, q^{t-1})$ , offers a menu of two distinct contracts, and sets at least one penalty fee strictly positive. If  $\mu_H^* > \mu_L^*$ , then any weakly positive penalty fee  $p_{3H} \geq 0$  is optimal on the high contract, but the low contract must charge a strictly positive penalty fee  $p_{3L} > 0$ . The reverse is true for  $\mu_H^* < \mu_L^*$ . There are three cases: (1) If*

$$-X_L \leq \mu_H^* - \mu_L^* \leq X_H, \quad (12)$$

*then both types receive the first best allocation,  $v_L^* = v_H^* = c$ , and contract mark-ups are  $\mu_L^*$  and  $\mu_H^*$  respectively. (2) If  $\mu_H^* - \mu_L^* > X_H$ , then the high type receives the first best allocation,  $v_H^* = c$ , and any weakly positive penalty fee  $p_{3H} \geq 0$  is optimal on the high contract. However the downward*

---

<sup>14</sup>In particular, the constraint is symmetric such that any pair  $\{p_{1s}, p_{2s}\}$  which have the same average  $\bar{p}_s$  are still equivalent.

*incentive-constraint (IC-H) binds and the low-type's allocation is distorted downwards below first best:  $v_L^* > c$ . Moreover, the low type pays a strictly positive penalty fee  $p_{3L} = h_L(v_L^*) > 0$  and  $v_L^*$  must satisfy the first order condition:*

$$v_L = c + \frac{\beta}{1-\beta} \frac{F_L(v_L) - F_H(v_{HL})}{f_L(v_L)} \frac{-\partial\Pi/\partial U_H}{\beta G_L(U_L)} \left( (1 + p_{3L} f_L(v_L)) + \frac{1}{2} (F_L(v_L) - F_H(v_{HL})) h'_L(v_L) \right), \quad (13)$$

*where  $v_{HL} = v_L + p_{3L} (F_L(v_L) - F_H(v_{HL}))$ . (3) If  $\mu_H^* - \mu_L^* < -X_L$ , then the low type receives the first best allocation  $v_L^* = c$  and any weakly positive penalty fee  $p_{3L} \geq 0$  is optimal on the low contract. However the upward incentive-constraint (IC-L) binds and the high type's allocation is distorted upwards above first best:  $v_H^* < c$ . Moreover, the high type pays a strictly positive penalty fee  $p_{3H} = h_H(v_H^*) > 0$  and  $v_H^*$  must satisfy the first order condition:*

$$v_H = c - \frac{1-\beta}{\beta} \frac{F_L(v_{LH}) - F_H(v_H)}{f_H(v_H)} \frac{-\partial\Pi/\partial U_L}{(1-\beta) G_H(U_H)} \left( (1 + p_{3H} f_H(v_H)) - \frac{1}{2} (F_L(v_{LH}) - F_H(v_H)) h'_H(v_H) \right), \quad (14)$$

*where  $v_{LH} = v_H - p_{3H} (F_L(v_{LH}) - F_H(v_H))$ .*

Comparing Propositions 3 and 5 shows that the combination of penalty fees and consumer inattention can be socially valuable in reducing allocative distortions due to price discrimination when unconstrained optimal-markups across different consumer segments are different, but not too different.

Suppose that  $\mu_L^* < \mu_H^*$ . If a firm were unconstrained by ex ante incentive constraints (as in the case of third-degree price discrimination between low and high types) it would always be optimal to offer both groups first best allocations via marginal-cost pricing. The firm would then like to offer a discounted fixed fee to low-types. This is not feasible if consumers are attentive, as high-types would always choose the discounted contract intended for low-types. To satisfy the ex ante incentive constraint and give low-types a discount, the firm distorts the allocation of the low-type downwards. The striking result for inattentive consumers is that this is no longer the case for small discounts. Via the use of penalty fees, the firm can ensure that both low and high-types face expected marginal prices equal to marginal cost on their respective contracts, low-types pay a discounted markup, and that the ex ante incentive constraint is satisfied. This is possible because a positive penalty fee on contract L which makes expected marginal price equal to marginal cost for low-types also makes expected marginal price exceed marginal cost for deviating high-types.

Corollary 3 states the first of two main results in the paper. The combination of penalty fees and consumer inattention are socially valuable and price-posting regulation is counter-productive whenever markets are fairly competitive.

**Corollary 3** *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 pure strategy equilibrium, all customers are served, allocations are first best, and mark-ups are  $\mu_s = \tau_s$ . Moreover, the set of equilibrium prices includes  $p_{1s}^i = p_{2s}^i = 0$  and  $p_{3s}^i = c / (1 - F_s(c))$ . (2) Price-posting regulation (or constant-marginal-price regulation) would strictly decrease welfare and firm profits. Low types would be losers while high types would be winners.*

The intuition behind the result in Corollary 3 that PPR is socially detrimental is as follows. Consider starting at the inattentive equilibrium and introducing PPR. At existing prices, PPR would cause the downward incentive-constraint (IC-H) to be violated, and firms could no longer charge markups that were so different. To restore incentive compatibility, firms would reduce markups on contract  $H$ , increase markups on contract  $L$ , and distort allocations on contract  $L$  downwards to reduce the need to adjust markups even further. The changes in markups drive the consumer surplus results, while the allocative distortion causes the reduction in social welfare. Firm market shares are unaffected in equilibria, 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 expensive to raise markups on contract  $L$  in terms of market share.<sup>15</sup>

In contrast with fairly-competitive markets, sufficient market power implies that penalty fees and inattentive consumers do not produce efficient outcomes. Corollary 4 illustrates this for the zero outside option monopoly.

**Corollary 4** *Let the firm be a monopolist serving consumers with zero outside option and  $F_H < F_L$  for all  $v \in (0, 1)$  (a strong form of strict FOSD). The upward ex ante incentive constraint binds and the low-type's allocation is distorted below first best:  $v_L^* > c$ .*

**Proof.** By assumption,  $G_s(U_s) = 1_{U_s \geq 0}$  and  $\beta$  is sufficiently small that it makes sense to serve the low types. (If not  $v_L^* = 1 > c$  and the result is true as well). Hence, at the optimum,  $G_L(U_L) = G_H(U_H) = 1$  and  $\frac{-\partial \Pi / \partial U_H}{\beta G_L(U_L)} = 1$ . When neither IC-L nor IC-H bind,  $U_L = U_H = 0$ . However, the high type can always mimic the low type by choosing contract  $L$  and a threshold  $v_{HL}$  such that  $F_H(v_{HL}) = F_L(v_L)$ . In this case, the high type makes the same expected payments and

---

<sup>15</sup>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.

the same number of purchases, but at FOSD higher valuations. Thus  $U_{HL} > U_L = U_H = 0$ , which violates IC-H. ■

When there is sufficient market power the impact of regulation becomes ambiguous. Let the firm be a monopolist serving consumers with zero outside option. Without a binding revenue raising requirement, a regulator with sufficient information and authority would optimally set marginal price equal to marginal cost to achieve efficient allocations. In this case inattention and price-posting regulation have no effect on outcomes. If a revenue raising requirement was binding, then a regulator setting optimal Ramsey prices would keep marginal prices hidden from inattentive consumers for the same reason an unregulated firm would: inattention allows revenues to be more efficiently extracted from high types. If a regulator is unable to directly regulate prices, but could require marginal prices to be posted at the time of transaction, such regulation may or may not be beneficial. Proposition 6 gives sufficient conditions for such price-posting regulation to be beneficial and sufficient conditions for price-posting regulation to be harmful.

**Proposition 6** *Let the firm be a monopolist serving consumers with zero outside option (ZOOM). Suppose that there is an exogenous restriction that  $p_{3s} \leq h_s(v_s)$  for  $h_s(v_s) > 0$  and  $h'_s(v_s) \geq 0$ . Assume that  $f_H$  crosses  $f_L$  once from below at  $v = c^* > 0$ . (1) If  $c < c^*$  and  $f_H$  is weakly decreasing above  $c$ , then for  $\beta > 0$  sufficiently small, price-posting regulation improves welfare. (2) If  $c > c^*$  and  $h_s(v_s) = p^{\max} > 0$ , then either for  $p^{\max}$  sufficiently small or for  $f_H$  weakly increasing above  $c$ , price-posting regulation reduces welfare.*

#### 4.4 Constraints on penalty fees

Corollary 3 shows that, given sufficient competition, case (1) of Proposition 5 applies, ex ante incentive constraints are slack, and finite penalty fees are optimal. Thus with sufficient competition, restrictions on penalty fees do not bind, and the precise form of restriction does not matter. Hence Corollary 3 and the result it highlights – that in competitive markets the combination of penalty fees and consumer inattention can be socially valuable – are robust to a variety of restrictions on penalty fees.

When equation (12) isn't satisfied in equilibrium, then it is strictly optimal to set at least one penalty fee as high as possible. Without restriction this leads to the unreasonable prediction of negative infinity base marginal prices and positive infinity penalty fees. For simplicity and tractability, in the preceding analysis I imposed one of two exogenous constraints on penalty fees: either (a) that penalty fees must be below some exogenous upper bound  $p^{\max}$ , or (b) that marginal prices be non-negative. However, there are many natural economic forces absent from the model that would endogenously restrict penalty fees. This is particularly true because profits are bounded

(strictly) below first best surplus. Thus as penalty fees grow large, the remaining profit increase from increasing them all the way to infinity becomes arbitrarily small. Hence any arbitrarily small cost of raising penalty fees would be sufficient to endogenously limit penalty fees to finite levels.

Economic forces that would endogenously restrict penalty fees include: (1) Limited liability; (2) Mild consumer risk aversion; (3) A small risk of regulatory intervention that increases in the size of penalty fees; (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 their while; (6) Consumers who attend to the date and could condition  $v^*$  on the date.

(1) Limited liability restricts total price to always be below a consumer's wealth:  $p_{0s} \leq W$ ,  $p_{0s} + p_{1s} \leq W$ ,  $p_{0s} + p_{2s} \leq W$ , and  $p_{0s} + p_{1s} + p_{2s} + p_{3s} \leq W$ . Combining equations (8) and (9), the last constraint imposes an upper bound on the penalty fee  $p_{3s}$  for any fixed  $v_s^*$  and  $U_s$ :

$$p_{3s} \leq h(v_s, U_s) = \frac{W + U_s - v_0 - 2 \int_{v_s^*}^1 v dF_s(v) - 2v_s^* F_s(v_s^*)}{2F_s^2(v_s^*)}.$$

(2) Low base marginal fees combined with high penalty fees ensure that an inattentive consumer's bill is a lottery, the size of which will be limited by even mild risk aversion. (3) Regulatory threat is self explanatory. (4) Any consumers who are attentive can ensure they purchase one and only one unit of the add-on service. If penalty fees are too high this will be costly to the firm, since (combining equations (8) and (9)) the firm would end up paying them a subsidy of at least<sup>16</sup>

$$-(p_{0s} + \bar{p}_s) = U_s - v_0 - 2 \int_{v_s}^1 (v - v_s) dF_s(v) - v_s + p_{3s}(1 - F_s(v_s)) F_s(v_s), \quad (15)$$

which for any fixed  $U_s$  and  $v_s$  is increasing linearly in the penalty  $p_{3s}$ . (5) Rational inattention limits penalty fees because increasing penalty fees increases consumers' returns to attention and thus the number of consumers who endogenously choose to be attentive. (6) Consumers who attend to the date would restrict penalty fees because if penalty fees were sufficiently high, a consumer who attended to the date would never buy in the first period, but always buy in the second (or vice-versa), thereby always avoiding the penalty fee, but receiving at least the subsidy in equation (15).

As already noted, the pricing predictions would be qualitatively robust under any of these modifications given strong competition, since for strong competition small penalty fees are sufficient. Clearly introducing additional players (attentive consumers), or costs (risk aversion) would affect welfare predictions, but only slightly if the additions are small. With sufficient market power,

---

<sup>16</sup>The subsidy would be higher if the firm chose asymmetric prices  $p_{1s} \neq p_{2s}$ .

modifications (4) or (5) could qualitatively change pricing predictions by making asymmetric prices ( $p_{1s} \neq p_{2s}$ ) optimal. This would be in response to the information asymmetry between periods in the attentive consumers dynamic programming problem. There would be no qualitative change in pricing predictions from modifications (2), (3), or (6). The limited liability constraint on penalty fees is relaxed when utility offers are increased, and hence would have an additional affect on markups (beyond the indirect affect via limiting penalty fees), but otherwise would not qualitatively affect pricing predictions.

An additional endogenous restriction on penalty fees would come from the existence of 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. The existence of such potential customers imposes what I call the *no-free-lunch* (NFL) constraint. This restricts consumer payments to be non-negative at all allocations:  $p_{0s} \geq 0$ ,  $p_{0s} + p_{1s} \geq 0$ ,  $p_{0s} + p_{2s} \geq 0$ , and  $p_{0s} + p_{1s} + p_{2s} + p_{3s} \geq 0$ . Otherwise, the large pool of attentive consumers with zero value for the product would purchase exactly the right quantity to get paid by the firm. This limits penalty fees, since holding  $v_s$  and  $U_s$  fixed, increasing  $p_{3s}$  towards infinity sends  $p_{0s} + \bar{p}_s$  towards negative infinity (equation (15)). In fact, the NFL constraints  $p_{0s} + p_{1s} \geq 0$  and  $p_{0s} + p_{2s} \geq 0$  are equivalent to:

$$p_{3s} \leq h(v_s, U_s) = \frac{v_0 - U_s + 2 \int_{v_s}^1 (v - v_s) dF_s(v) + v_s}{(1 - F_s(v_s)) F_s(v_s)}.$$

I explore the NFL constraint further (under the assumption that consumer beliefs are biased) in Section 5.2.

## 5 Homogeneous consumers with biased beliefs

The previous section showed that when consumers are inattentive, penalty fees may be used to more efficiently price discriminate between customer segments with stochastically low and high demand for an add-on good or service. This provides an explanation for two choices by cellular-phone companies: first to offer tariffs with steep penalty charges for high usage and second to avoid actively notifying consumers of the accruing charges until the end of the month. The analysis also suggested caution with respect to the bill-shock regulation under consideration by the FCC, since if the cellular market is sufficiently competitive (and consumers are unbiased) then actively notifying customers about accruing charges could undermine the social benefits of consumer inattention.

Unfortunately, the analysis in the previous section sheds no light on why, prior to the Federal Reserve Board's adoption of an opt-in rule, Bank of America and other banks charged high (\$35) overdraft fees on debit and ATM transactions without notifying customers at the point of sale.

Banks like Bank of America do price discriminate by offering multiple types of checking accounts with different terms into which different customer segments self select. However, Bank of America and others typically did not use overdraft charges as a tool to encourage self selection. On the contrary, the terms of overdraft charges were typically the same across different types of accounts. (For example, Figure 1 shows Bank of America’s March 1st, 2010 menu of 4 types of checking accounts and Figure 2 describes overdraft fees which were the same for all 4 types of checking accounts.)

This section explores an explanation for firms’ valuation of penalty fees and consumer inattention that does apply to the case of overdraft fees: that consumers have biased beliefs and underestimate their consumption of the add-on good or service. Consumer inattention may exacerbate or ameliorate allocative distortions created by biased beliefs. When marginal costs are extreme relative to the distribution of consumer valuations, inattention creates allocative distortions that are worse than those with biased beliefs alone, thereby lowering total welfare. When marginal costs are high, the allocative distortion is overconsumption and there are surplus reducing trades. However, the effect of first-order importance may be on surplus distribution rather than total surplus. Inattention means that consumers can be exploited and receive payoffs far below their outside options. Price-posting regulation ensures that consumers receive at least their outside option.

## 5.1 Continuous taste shocks and welfare

If attentive consumers underestimate their demand for the service ex ante, then we know that firms have an incentive to set marginal charges above marginal cost, irrespective of competition (e.g. Grubb (2009)). Return to the assumption in the base model that consumers all have the same distribution of taste shocks  $F$ . Now, however, assume that consumers believe that the distribution is  $F^*$ , which like the true distribution  $F$  is continuous and strictly increasing on  $[0, 1]$ . Moreover, assume that  $F$  first-order-stochastically-dominates  $F^*$  so that consumers underestimate their demand for the add-on services.<sup>17</sup>

A consumer’s true base expected payoff from contracting with firm  $i$  at the contracting stage and making optimal consumption choices thereafter remains

$$U^i = E [u(\mathbf{q}(\mathbf{v}; \mathbf{p}^i), \mathbf{v}) \mid F] = \int_0^1 \int_0^1 u(\mathbf{q}(\mathbf{v}; \mathbf{p}^i), \mathbf{v}) dF(v_1) dF(v_2).$$

However, a consumer’s perceived expected payoff differs because expectations are taken with respect

---

<sup>17</sup>To capture overconfidence with only two subperiods, consumers would need to underestimate the correlation in  $v_i$  across periods.

to consumer beliefs:

$$U^{*i} = E [u(\mathbf{q}(\mathbf{v}; \mathbf{p}^i), \mathbf{v}) | F^*] = \int_0^1 \int_0^1 u(\mathbf{q}(\mathbf{v}; \mathbf{p}^i), \mathbf{v}) dF^*(v_1) dF^*(v_2).$$

The fraction  $G(U^{*i}; U^{*-i})$  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:

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

Thus firm  $i$ 's expected profits are

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

which can be rewritten in terms of total surplus and consumers' true and perceived expected-utilities:

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

### 5.1.1 Attentive benchmark

Proposition 7 characterizes optimal pricing in the attentive case.<sup>18</sup>

**Proposition 7** *If all consumers are attentive and homogeneously underestimate demand, then the optimal contract is a two part tariff ( $p_3 = 0$ ,  $p_1 = p_2 = p$ ) with marginal price above marginal cost,*

$$p = c + \frac{F^*(p) - F(p)}{f(p)} > c,$$

*and allocations are inefficiently low. All consumers are weakly better off than choosing their outside options, and all transactions generate positive surplus.*

Proposition 7 shows the potential for biased beliefs to reduce welfare in the absence of inattention by distorting consumption downwards. It also points out that when attentive consumers underestimate their value for a good or service they cannot be exploited (they must receive at least their outside option) and there are no surplus reducing trades.

---

<sup>18</sup>Marginal pricing is the unit-demand analog of that characterized by Grubb (2009) for continuous demand and  $T = 1$ , repeated in each subperiod  $t \in \{1, 2\}$ .

### 5.1.2 Inattentive case

Now consider the inattentive case. The consumption threshold chosen by an inattentive consumer with biased beliefs satisfies the first order condition,

$$v^* = \frac{p_1 + p_2}{2} + p_3 (1 - F^*(v^*)) \quad (16)$$

which substitutes consumer beliefs in place of the true distribution of tastes in equation (4). As before, I focus on symmetric pricing  $p_1 = p_2 = \bar{p}$  and it is useful to reframe the firm's problem in two ways. First, it is convenient to think of the firm choosing perceived expected-utility  $U^*$  rather than setting fixed fee  $p_0$ . In this case the fixed fee  $p_0$  is given by equation (17):

$$p_0 = -U^* + v_0 + 2 \int_{v^*}^1 v dF^*(v) - 2\bar{p}(1 - F^*(v^*)) - p_3 (1 - F^*(v^*))^2. \quad (17)$$

Second, it is convenient to think of the firm first choosing consumer threshold  $v^*$  and then choosing the best marginal prices  $\bar{p}$  and  $p_3$  which implement  $v^*$ . Given any fixed choice of perceived expected-utility  $U^*$  and consumer threshold  $v^*$ , by Proposition 1, it is necessary for  $\bar{p}$  to satisfy the first order condition:

$$\bar{p} = v^* - p_3 (1 - F^*(v^*)). \quad (18)$$

Using equations (17) and (18), firm profits can be written as a function of perceived expected-utility  $U^*$ , penalty  $p_3$ , and consumer threshold  $v^*$ :

$$\Pi = G(U^*) \left( -U^* + 2 \int_{v^*}^{\infty} \left( v - c - \frac{F^*(v) - F(v)}{f(v)} \right) f(v) dv + p_3 (F^*(v^*) - F(v^*))^2 \right). \quad (19)$$

Note that profits increase linearly in the penalty fee  $p_3$ . Thus the optimal penalty fee will be positive, in which case the local incentive constraint of equation (18) is sufficient for  $v^*$  to be globally optimal. Moreover, without any additional constraints, firms optimally choose  $p_3 = \infty$  and  $v^* \in (0, 1)$ . This contract transfers infinite wealth from consumers to the firm. Infinite penalty fees are implausible because many forces will restrict the size of penalty fees in practice, as discussed in Section 4.4. An important difference with biased beliefs is that the returns to increasing penalty-fees are constant rather than decreasing. Thus a fraction of consumers who are attentive would still endogenously restrict penalty fees, but only if the fraction were sufficiently large. For simplicity I

impose a maximum penalty fee  $p^{\max}$ . The firm's problem can then be written as:

$$\begin{aligned} \max_{U^*, v^*, p_3} & G(U^*) \left( -U^* + 2 \int_{v^*}^{\infty} \left( v - c - \frac{F^*(v) - F(v)}{f(v)} \right) f(v) dv + p_3 (F^*(v^*) - F(v^*))^2 \right) \\ \text{s.t. } & p_3 \leq p^{\max} \end{aligned}$$

Equation (19) shows that for any fixed finite-penalty-fee  $p_3$ , profits are increasing in the size of the disagreement between consumer and firm about the consumer's per period purchase probability  $|F^*(v^*) - F(v^*)|$ . Given a restriction on penalty fees, firms have an incentive to adjust consumers' threshold choice  $v^*$  to increase this disagreement. In general, the incentive to maximize disagreement could increase or decrease distortions relative to the attentive case. However, if marginal costs are extreme relative to consumer valuations, then maximizing disagreement entails increased inefficiency and price-posting regulation increases welfare (Proposition 8). 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  $[0, 1]$ ,  $F < \hat{F}$  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 8** *Assume penalty fees are exogenously restricted by an upper bound  $p^{\max} > 0$ . (1) Holding  $F$  and  $\hat{F}$  fixed, if  $c$  and  $\gamma$  are sufficiently close to zero, then inattention exacerbates underconsumption ( $v^* > v^A > c$ ). (2) If  $p^{\max}$  is sufficiently large, then for marginal costs in a neighborhood around  $c = 1$ , inattention creates overconsumption worse than the attentive underconsumption ( $v^* < c \leq v^A$ ). In both cases, price-posting regulation strictly improves welfare.*

Note that Proposition 8 points out that when consumers are inattentive, consumers who underestimate their demand for a product or service may be induced to overconsume. For instance when  $c$  is slightly above 1, all product sales are inefficient. Yet because inattentive consumers underestimate their likely values for the product, sales take place. Price posting regulation would increase consumer surplus and total welfare by ending sales of these products. This is potentially why Bank of America chose to stop offering overdraft protection on debit-card transactions following the Federal Reserve's new 'opt-in' requirement, 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).

## 5.2 Bernoulli taste shocks and surplus distribution

The effects of inattention and price-posting regulation on total welfare may in fact be second order relative to their effects on the distribution of surplus. In some situations, the welfare effects are likely to be small, for instance because costs and values are similar or small, or because valuations have a concentrated distribution. But more importantly, since inattentive consumers who underestimate their demand can be exploited, surplus distribution effects of price-posting regulation are not limited by first best surplus but can be orders of magnitude higher.

To focus on distributional issues, I make an alternative assumption about the distribution of taste shocks. For the rest of the paper, assume taste shocks have a Bernoulli distribution:  $v_t$  are drawn independently and are equal to 1 with probability  $\alpha$  and zero otherwise. Consumers underestimate their demand and believe that  $v_t$  equals 1 with probability  $\alpha' < \alpha$ . Also assume  $c \in (0, 1)$ . Finally, rather than exogenously imposing an upper bound on penalty fees or imposing that marginal prices be non-negative, I will endogenously restrict penalty fees by imposing the no-free-lunch constraint.

### 5.2.1 Attentive benchmark

To solve the firm's problem in both attentive and inattentive cases I proceed in two steps. First, I fix a perceived utility  $U^*$  to be offered and solve for the optimal price vector  $\mathbf{p}$  which implements  $U^*$  subject to the NFL constraint. This price vector determines the allocation, expected surplus  $S$ , and true expected-utility  $U$ . Hence the first step derives an optimal markup,  $\mu(U^*)$ , to be charged as a function of  $U^*$ . The second step is to choose the perceived expected-utility  $U^*$  which maximizes profits,  $\Pi = G(U^*) \mu(U^*)$ , subject to feasibility under NFL.

If consumers are attentive, firm's offer two part tariffs ( $p_3 = 0, p_1 = p_2 = p$ ) with no penalty fees and marginal price  $p \in (0, 1]$  which induces efficient consumption. Further, the optimal marginal price  $p$  and markup  $\mu(U^*)$  are characterized as a function of perceived expected-utility by Proposition 9.

**Proposition 9** *Given Bernoulli taste shocks, attentive consumers who underestimate demand ( $\alpha' < \alpha$ ),  $c \in (0, 1)$ , and the no-free-lunch constraint: Firms set zero penalty fees ( $p_3 = 0$ ) and offer the first best allocation. Moreover, (1) Conditional on offering  $U^* \in [0, v_0]$ , optimal prices and markups are*

$$p_3 = 0, p_1 = p_2 = 1, p_0 = v_0 - U^*$$

$$\mu(U^*) = S^{FB} - U^*. \tag{20}$$

(2) Conditional on offering  $U^* \in [v_0, v_0 + 2\alpha']$ , optimal prices and markup are:

$$p_3 = p_0 = 0, p_1 = p_2 = 1 - (U^* - v_0) / 2\alpha',$$

$$\mu(U^*) = (S^{FB} - U^* - (\alpha/\alpha' - 1)(U^* - v_0)). \quad (21)$$

(3) Offering  $U^* > v_0 + 2\alpha'$  is not feasible under NFL.

The no-free-lunch constraint requires that all payments from consumers to the firm be non-negative. Increasing perceived utility  $U^*$  while holding the allocation fixed at first best entails lowering prices. Hence the no-free-lunch constraint is increasingly difficult to satisfy as the offered  $U^*$  rises. This explains the three pricing regions in Proposition 9. Absent the NFL constraint, it would always be optimal to charge the maximum marginal price that induces first-best consumption ( $p = 1$ ) and implement  $U^*$  by appropriate choice of  $p_0$ . For  $U^* \in [0, v_0]$ , NFL is slack and this is the solution. For  $U^* \in (v_0, v_0 + 2\alpha']$ , however, this would require a negative fixed-fee and the NFL constraint  $p_0 \geq 0$  binds. To achieve  $U^* \in (v_0, v_0 + 2\alpha']$  with a non-negative fixed-fee requires charging a lower marginal-price  $p < 1$ . Finally,  $U^* = v_0 + 2\alpha'$  is the maximum perceived expected-utility that can be offered with non-negative prices.

### 5.2.2 Inattentive case

If consumers are inattentive, firms charge positive penalty fees, but still induce efficient consumption. Thus total surplus is first best irrespective of inattention, price-posting regulation, or consumers' biased beliefs. The distribution of surplus, however, varies significantly with inattention and price-posting regulation.

Given Bernoulli taste shocks, an inattentive consumer's strategy is described by the pair  $\{b_0, b_1\}$ . These are the probabilities of purchase conditional on realizing  $v_t = 0$  or  $v_t = 1$  respectively:  $b_0 = \Pr(q_t(v_t = 0) = 1)$  and  $b_1 = \Pr(q_t(v_t = 1) = 1)$ . A consumer's perceived expected-utility  $U^*$  is given by equation (22) as a function of prices and the strategy  $\{b_0, b_1\}$ :

$$U^*(b_0, b_1) = -p_0 + v_0 + 2(1 - \alpha')b_0(-\bar{p}) + 2\alpha'b_1(1 - \bar{p}) - ((1 - \alpha')b_0 + \alpha'b_1)^2 p_3. \quad (22)$$

Firm profits, as a function of prices, perceived expected-utility  $U^*$ , and the allocation  $\{b_0, b_1\}$  are given by equation (23):

$$\Pi = G(U^*) \left( p_0 + 2((1 - \alpha)b_0 + \alpha b_1)(\bar{p} - c) + ((1 - \alpha)b_0 + \alpha b_1)^2 p_3 \right). \quad (23)$$

The first result is that it will be optimal for firms to set prices which induce the efficient allocation  $\{b_0, b_1\} = \{0, 1\}$ .

**Lemma 1** *Given Bernoulli taste shocks, inattentive consumers who underestimate demand ( $\alpha' < \alpha$ ),  $c \in (0, 1)$ , and the no-free-lunch constraint, firms set prices which induce the efficient allocation: consumers buy if and only if  $v_t = 1$ .*

To induce the efficient allocation, the firm must set expected marginal price conditional on a purchase to be between zero and one:  $0 \leq \bar{p} + \alpha' p_3 \leq 1$ . Applying Lemma 1, the firm's problem can thus be reduced to the following:<sup>19</sup>

$$\begin{aligned} \max_{U^*, \bar{p}, p_3} \Pi &= G(U^*) (p_0 + 2\alpha(\bar{p} - c) + \alpha^2 p_3) \\ \text{such that} &: \\ \text{IC:} & \quad 0 \leq \bar{p} + \alpha' p_3 \leq 1 \\ \text{NFL:} & \quad p_0 \geq 0, p_0 + \bar{p} \geq 0, p_0 + 2\bar{p} + p_3 \geq 0 \\ \text{Fixed Fee} &: \quad p_0 = -U^* + v_0 + 2\alpha'(1 - \bar{p}) - \alpha'^2 p_3 \end{aligned}$$

Proposition 10 characterizes optimal prices given a fixed perceived-expected-utility  $U^*$ . For low utility offers, the NFL constraint  $p_0 + \bar{p} \geq 0$  and the IC constraint  $\bar{p} + \alpha' p_3 \leq 1$  both bind. For medium utility offers, the two NFL constraints  $p_0 \geq 0$  and  $p_0 + \bar{p} \geq 0$  both bind. Higher utility offers above  $v_0 + 2\alpha'$  are not feasible given the NFL constraint.

**Proposition 10** *Given Bernoulli taste shocks, inattentive consumers who underestimate demand ( $\alpha' < \alpha$ ),  $c \in (0, 1)$ , and the NFL constraint: (1) Conditional on offering  $U^* \in [0, v_0 + \alpha']$ , optimal prices and markup are*

$$\begin{aligned} p_1 = p_2 = -p_0 &= -\frac{v_0 + \alpha' - U^*}{1 - \alpha'}, p_3 = \frac{v_0 + 1 - U^*}{(1 - \alpha')\alpha'}, \\ \mu(U^*) &= S^{FB} - U^* + \frac{(\alpha - \alpha')^2}{\alpha'(1 - \alpha')} (1 + v_0 - U^*). \end{aligned} \tag{24}$$

(2) Conditional on offering  $U^* \in [v_0 + \alpha', v_0 + 2\alpha']$ , optimal prices and markup are:

$$p_1 = p_2 = p_0 = 0, p_3 = (2\alpha' + v_0 - U^*) / \alpha'^2,$$

---

<sup>19</sup>It is strictly optimal to set prices symmetrically,  $p_1 = p_2 = \bar{p}$ , since keeping  $\bar{p}$  constant but setting  $p_1 < p_2$  would tighten the NFL constraint  $p_0 + p_1 \geq 0$  without otherwise effecting consumer incentives or firm profits. Similarly, setting  $p_2 < p_1$  would tighten the NFL constraint  $p_0 + p_2 \geq 0$ .

$$\mu(U^*) = \left( S^{FB} - U^* - \left( (\alpha/\alpha')^2 - 1 \right) (U^* - v_0) + 2(\alpha - \alpha') \alpha/\alpha' \right). \quad (25)$$

(3) Offering  $U^* > v_0 + 2\alpha'$  is not feasible under NFL.

Propositions 9 and 10 characterize optimal prices and markup  $\mu(U^*)$  as a function of perceived expected-utility  $U^*$ . Corollary 5 applies Propositions 9 and 10 to a zero-outside-option monopoly for which the optimal utility offer is  $U^* = 0$ . The result compares attentive and inattentive cases and evaluates the effect of price-posting regulation:

**Corollary 5** *Assume a zero-outside-option monopoly, the no-free-lunch constraint, Bernoulli taste-shocks, consumers who underestimate demand ( $\alpha' < \alpha$ ), and  $c \in (0, 1)$ . If consumers are attentive, the monopolist charges  $p_0 = v_0$ ,  $p_1 = p_2 = 1$ , and  $p_3 = 0$ , induces efficient consumption, and captures the full surplus ( $\Pi = S^{FB}$ ,  $CS = 0$ ). Let  $Y \equiv (\alpha - \alpha')^2 / (\alpha'(1 - \alpha'))$ . If consumers are inattentive, the monopolist charges  $p_1 = p_2 = -p_0 = -(v_0 + \alpha') / (1 - \alpha')$  and  $p_3 = (v_0 + 1) / (\alpha'(1 - \alpha'))$ . While still inducing efficient consumption, the monopolist now captures more than the entire first best surplus ( $\Pi = S^{FB} + (1 + v_0)Y$ ) and consumers are exploited, receiving less than their outside option ( $CS = -(1 + v_0)Y < 0$ ). Price posting regulation does not affect total welfare, but redistributes  $(1 + v_0)Y$  from firm to consumers and eliminates consumer exploitation.*

**Proof.** A direct application of Propositions 9 and 10 given that the optimal utility offer is  $U^* = 0$  given ZOOM. ■

Note that my choice of the no-free-lunch constraint, rather than an alternative restriction on penalty fees, does not qualitatively effect the results in Corollary 5, only the magnitude of the shift in surplus  $(1 + v_0)Y$  would vary with alternative constraints. The assumption has a more substantive role in competitive markets however. For instance, with a simple upper bound of  $p^{\max}$  imposed on penalty fees, the redistributive effects of price-posting regulation would vanish with Hotelling competition, because additional profits extracted from inattentive consumers through penalty fees would be rebated through fixed fees due to competition. The no-free-lunch constraint, however, restricts fixed fees to be non-negative. Once firms reduce fixed fees to zero, they are forced to compete on either base marginal charges or penalty fees. This softens price competition and raises profits, because consumers underweight the chance of paying both base marginal charges and penalty fees and hence are less price sensitive to them than to fixed fees. The effect is larger for penalty fees, used with inattentive consumers, than with base marginal charges, used with attentive consumers. As a result, the no-free-lunch constraint implies that the redistributive effects of price-posting regulation persist under competition.

		Inattentive	Attentive (PPR)	Redistribution
Monopoly (ZOOM)	$\Pi$	$S^{FB} + (1 + v_0)Y$	$S^{FB}$	$(1 + v_0)Y$
	$U$	$-(1 + v_0)Y$	0	
Duopoly (Hotelling)	$\Pi$	$\tau(\alpha/\alpha')^2$	$\tau(\alpha/\alpha')$	$\tau(\alpha/\alpha')(\alpha/\alpha' - 1)$
For $\tau < \alpha'(1 - 2c)$	$U$	$S^{FB} - \tau(\alpha/\alpha')^2$	$S^{FB} - \tau(\alpha/\alpha')$	

Table 1: Summary of surplus distribution results from Corollaries 5 and 6. Profits, and consumer surplus under zero outside option monopoly and Hotelling duopoly with and without price posting regulation.

Corollary 6 applies Propositions 9 and 10 to a fairly competitive Hotelling duopoly, solves for equilibrium utility offers  $U^*$ , and compares attentive and inattentive cases to evaluate the effect of price-posting regulation.

**Corollary 6** *Assume duopoly competition on a uniform Hotelling line, the no-free-lunch constraint, Bernoulli taste shocks, consumers who underestimate demand ( $\alpha' < \alpha$ ) and  $c \in (0, 1/2)$ . Let  $\tau \in (0, \alpha'(1 - 2c)]$ . The market is fully covered. (1) If consumers are attentive, then the fixed fee and penalty fee are zero ( $p_0 = p_3 = 0$ ) and firms compete on base marginal charges  $p_1 = p_2 = c + \tau/2\alpha'$ . This softens price competition so that firms earn profits above  $\tau$ :  $\Pi = \tau(\alpha/\alpha')$ ,  $U = S^{FB} - \tau(\alpha/\alpha') \geq 0$ . Consumers are not exploited. (2) If consumers are inattentive, then in the unique pure-strategy equilibrium, fixed fee and base marginal-charges are zero ( $p_0 = p_1 = p_2 = 0$ ) and firms compete on penalty fees  $p_3 = 2c/\alpha + \tau/\alpha'^2$ , offering  $U^* = v_0 + 2\alpha'(1 - 2c(\alpha'/\alpha)) - \tau$ . This further softens price competition and leads to still higher industry profits,  $\Pi = \tau(\alpha/\alpha')^2$  and lower consumer surplus  $U = S^{FB} - \tau(\alpha/\alpha')^2$ . Consumers may be exploited and will be for sufficiently large bias (small  $\alpha'$ ). (3) Price posting regulation does not affect total welfare, but it does redistribute  $\tau(\alpha/\alpha')(\alpha/\alpha' - 1)$  from firms to consumers and eliminate any consumer exploitation which could have been present despite competition.*

Corollaries 5 and 6 capture the second main result in the paper – that, combined with biased beliefs, inattention can cause consumers to receive payoffs far below their outside option and that price-posting regulation will eliminate this exploitation. The surplus distribution results from Corollaries 5 and 6 are summarized in Table 1. The following numerical example takes advantage of the fact that Corollary 6 holds under the weaker sufficient condition  $\tau < \frac{4}{3}\alpha'(1 - c)$  and  $\tau < \alpha'(1 - 2c(\alpha'/\alpha))$ .

**Example 1** *Let  $\alpha = 9/10$ ,  $\alpha' = 1/10$ ,  $c = 4/9$ , and  $v_0 = 0$ . For the Hotelling duopoly, let  $\tau = 1/20$ . Given these parameters, first best surplus is  $S^{FB} = 1$ . Pricing, profits, and consumer*

		Inattentive	Attentive (PPR)	Redistribution
Monopoly (ZOOM)	$\{p_0, p, p_3\}$	$\{.1, -.1, 11.1\}$	$\{0, 1, 0\}$	
	$\Pi$	8.1	1	7.1
	$U$	-7.1	0	
Duopoly (Hotelling)	$\{p_0, p, p_3\}$	$\{0, 0, 6\}$	$\{0, 0.7, 0\}$	
	$\Pi$	4.05	0.45	3.6
	$U$	-3.05	0.55	

Table 2: Example 1: Pricing, profits, and consumer surplus under zero outside option monopoly and Hotelling duopoly with and without price posting regulation.

*surplus under monopoly and duopoly with and without price-posting regulation are given in Table 2.*

## 6 Conclusion

If consumers have unbiased beliefs, but have heterogeneous forecasts of their future demand for an add-on good or service, the combination of consumer inattention and penalty fees can help firms price discriminate between customer segments with stochastically low and high demand forecasts. Price-posting regulation, by providing inattentive consumers with the same information recalled by attentive consumers, can help consumers avoid penalty fees. While this is good for consumers holding prices fixed, it undermines the value of penalty fees and will cause firms to change their prices. When firms have substantial market power, it is ambiguous whether this will increase or decrease total welfare. In fairly competitive markets, however, 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.

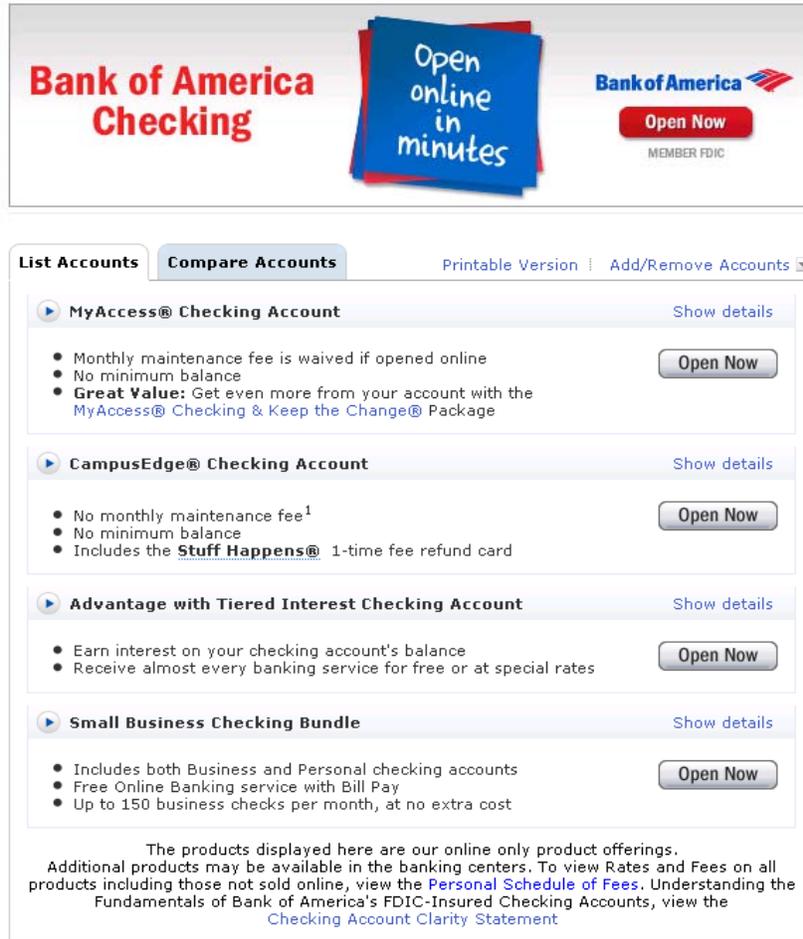
The model provides an explanation for two facts about cellular-phone service in the US. First, customers are charged steep penalty fees for exceeding usage allowances, and the variation in usage allowances across calling plans is an essential instrument for encouraging consumers to self select into different calling plans. Second, firms do not actively alert customers to accruing charges prior to the end of the month. If one believes that cellular phone customers have correct beliefs and the cellular market is sufficiently competitive, then the FCC's considered bill-shock regulation, which requires carriers to alert consumers to rapidly accruing charges, would be counterproductive. However, Grubb (2009) and Grubb and Osborne (2010) present compelling evidence that cellular phone customers have biased beliefs about their likely usage. Moreover it is not clear how competitive the

market for cellular-phone service is. As a result, the welfare impact of price-posting regulation is ambiguous and caution should be applied in adopting the FCC's considered bill-shock regulation.

When consumers underestimate their demand for an add-on service or good, the combination of consumer inattention and penalty fees can be highly profitable for firms. In fact, they 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 they are worse off than had they never done business with the firm. It is ambiguous whether price-posting regulation would increase or decrease welfare, but such changes in total welfare could be overshadowed by much larger changes in the distribution of surplus and the elimination of consumer exploitation. In both monopoly and competitive markets, price-posting regulation eliminates consumer exploitation and can increase consumer surplus by orders of magnitude more than the entire social surplus of the transaction.

This is one explanation for the high revenues (\$20Bn in 2009) from overdraft charges for ATM and one-time debit-card transactions, which is consistent with the fact that Bank of America cancelled its \$2.2Bn service when required by the Federal Reserve Board to ask consumers to opt-in (Martin 2010, Sidel and Fitzpatrick 2010). Moreover, it suggests that the Federal Reserve Board's regulation will substantially benefit consumers, and that banks will not be able to recoup the lost overdraft revenue simply by raising monthly fees on accounts. (Although they may of course find other equally profitable penalty fees to exploit.) It also suggests that the bill-shock regulation under consideration by the FCC could have substantial benefits for consumers, in particular as applied to fees such as roaming charges, which typically are the same across calling plans and are not used for purposes of price discrimination.

## 7 Figures



**Bank of America Checking** **Bank of America** 

**Open online in minutes**

**Open Now**  
MEMBER FDIC

---

**List Accounts** | **Compare Accounts** | [Printable Version](#) | [Add/Remove Accounts](#)

- MyAccess® Checking Account** [Show details](#)
  - Monthly maintenance fee is waived if opened online
  - No minimum balance
  - Great Value:** Get even more from your account with the MyAccess® Checking & Keep the Change® Package
- CampusEdge® Checking Account** [Show details](#)
  - No monthly maintenance fee<sup>1</sup>
  - No minimum balance
  - Includes the **Stuff Happens®** 1-time fee refund card
- Advantage with Tiered Interest Checking Account** [Show details](#)
  - Earn interest on your checking account's balance
  - Receive almost every banking service for free or at special rates
- Small Business Checking Bundle** [Show details](#)
  - Includes both Business and Personal checking accounts
  - Free Online Banking service with Bill Pay
  - Up to 150 business checks per month, at no extra cost

The products displayed here are our online only product offerings. Additional products may be available in the banking centers. To view Rates and Fees on all products including those not sold online, view the [Personal Schedule of Fees](#). Understanding the Fundamentals of Bank of America's FDIC-Insured Checking Accounts, view the [Checking Account Clarity Statement](#)

Figure 1: Bank of America’s menu of 4 checking accounts, offered online at [www.bankofamerica.com](http://www.bankofamerica.com) on March 1, 2010.

Other Account Fees			
Fee Category	Fee Name / Description	Fee Amount	Accounts Qualifying for Waiver of this Fee
<i>Overdraft Items (an overdraft item)</i>	Overdraft Item Fee	\$35.00 each item	N/A
	NSF: Returned Item Fee	\$35.00 each item	N/A
	Extended Overdrawn Balance Charge	\$35.00 – charged when we determine your account is overdrawn for 5 or more consecutive business days.	N/A

Figure 2: The overdraft fees associated with Bank of America’s checking accounts shown in Figure 1. They are the same across all accounts. Source [www.bankofamerica.com](http://www.bankofamerica.com), March 1, 2010.

## References

- Armstrong, Mark and John Vickers**, “Competitive Price Discrimination,” *RAND Journal of Economics*, 2001, 32 (4), 579–605.
- Aumann, Robert J., Sergiu Hart, and Motty Perry**, “The Absent-Minded Driver,” *Games and Economic Behavior*, 1997, 20 (1), 102–116.
- , — , and — , “The Forgetful Passenger,” *Games and Economic Behavior*, 1997, 20 (1), 117–120.
- Ausubel, Lawrence M. and Haiyan Shui**, “Time inconsistency in the credit card market,” Working Paper 2005.
- Bank of America**, “Personal Schedule of Fees,” Technical Report March 1 2010. <https://www3.bankofamerica.com/efulfillment/documents/91-11-3000ED.20100201.htm>.
- Baron, David P. and David Besanko**, “Regulation and Information in a Continuing Relationship,” *Information Economics and Policy*, 1984, 1 (3), 267–302.
- Battigalli, Pierpaolo**, “Dynamic Consistency and Imperfect Recall,” *Games and Economic Behavior*, 1997, 20 (1), 31–50.
- Borenstein, Severin**, “To What Electricity Price Do Consumers Respond? Residential Demand Elasticity Under Increasing-Block Pricing,” Preliminary Draft April 30 2009.
- Broadbent, Donald E.**, *Perception and Communication*, New York: Pergamon Press, 1958.
- Bubb, Ryan and Alex Kaufman**, “Consumer Biases and Firm Ownership,” Working Paper 2009.
- Bulow, Jeremy I., John D. Geanakoplos, and Paul D. Klemperer**, “Multimarket Oligopoly: Strategic Substitutes and Complements,” *The Journal of Political Economy*, 1985, 93 (3), 488–511.
- Cardon, James H. and Igal Hendel**, “Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey,” *The RAND Journal of Economics*, 2001, 32 (3), 408–427.
- Courty, Pascal and Hao Li**, “Sequential Screening,” *The Review of Economic Studies*, 2000, 67 (4), 697–717.
- Dash, Eric and Nelson D. Schwartz**, “Banks Seek to Keep Profits as New Oversight Rules Loom,” July 15 2010.
- DellaVigna, Stefano**, “Psychology and economics: Evidence from the field,” *Journal of Economic Literature*, 2009, 47 (2), 315–372.
- and **Ulrike Malmendier**, “Contract Design and Self-Control: Theory and Evidence,” *The Quarterly Journal of Economics*, 2004, 119 (2), 353–402.
- and — , “Paying Not to Go to the Gym,” *The American Economic Review*, 2006, 96 (3), 694–719.
- Diamond, Peter A.**, “A model of price adjustment,” *Journal of Economic Theory*, 1971, 3 (2), 156–168. doi: DOI: 10.1016/0022-0531(71)90013-5.

- Edlin, Aaron S. and Chris Shannon**, “Strict monotonicity in comparative statics,” *Journal of Economic Theory*, 1998, 81 (1), 201–219.
- Eliaz, Kfir and Ran Spiegler**, “Contracting with Diversely Naive Agents,” *The Review of Economic Studies*, 2006, 73 (3), 689–714.
- and — , “Consumer Optimism and Price Discrimination,” *Theoretical Economics*, 2008, 3 (4), 459–497.
- Ellison, Glenn**, “A Model of Add-on Pricing,” *The Quarterly Journal of Economics*, 2005, 120 (2), 585–637.
- FCC**, “Comment Sought on Measures Designed to Assist U.S. Wireless Consumers to Avoid ”Bill Shock”,” Public Notice May 11 2010. CG Docket No. 09-158, [http://hraunfoss.fcc.gov/edocs\\_public/attachmatch/DA-10-803A1.pdf](http://hraunfoss.fcc.gov/edocs_public/attachmatch/DA-10-803A1.pdf).
- Federal Reserve Board**, “Federal Register notice: Regulation E final rule,” Technical Report November 11 2009.
- , “Federal Reserve announces final rules prohibiting institutions from charging fees for overdrafts on ATM and one-time debit card transactions,” Press Release November 12 2009.
- Gabaix, Xavier and David Laibson**, “Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets,” *Quarterly Journal of Economics*, 2006, 121 (2), 505–540.
- Gaynor, Martin S., Yunfeng Shi, Rahul Telang, and William B. Vogt**, “Cell Phone Demand and Consumer Learning - An Empirical Analysis,” *SSRN eLibrary*, 2005.
- Gilboa, Itzhak**, “A Comment on the Absent-Minded Driver Paradox,” *Games and Economic Behavior*, 1997, 20 (1), 25–30.
- Grove, Adam J. and Joseph Y. Halpern**, “On the Expected Value of Games with Absentmindedness,” *Games and Economic Behavior*, 1997, 20 (1), 51–65.
- Grubb, Michael D.**, “Selling to Overconfident Consumers,” *American Economic Review*, 2009, 99 (5), 1770–1807.
- and **Matthew Osborne**, “Cellular service demand: Tariff choice, usage uncertainty, biased beliefs, and learning,” Working Paper 2010.
- Hadar, Josef and William R. Russell**, “Rules for Ordering Uncertain Prospects,” *American Economic Review*, 1969, 59 (1), 25.
- Halpern, Joseph Y.**, “On Ambiguities in the Interpretation of Game Trees,” *Games and Economic Behavior*, 1997, 20 (1), 66–96.
- Huang, Ching-I.**, “Estimating Demand for Cellular Phone Service Under Nonlinear Pricing,” *Quantitative Marketing and Economics*, 2008, 6 (4), 371–413.
- Lambrecht, Anja, Katja Seim, and Bernd Skiera**, “Does Uncertainty Matter? Consumer Behavior under Three-Part Tariffs,” *Marketing Science*, 2007, 26 (5), 698–710.

- Liebman, Jeffrey B. and Richard Zeckhauser**, “Schmeduling,” Working Paper October 2004.
- Lipman, Barton L.**, “More Absentmindedness,” *Games and Economic Behavior*, 1997, 20 (1), 97–101.
- Martin, Andrew**, “Bank of America to End Debit Overdraft Fees,” Technical Report, The New York Times March 10 2010.
- McAfee, R. Preston and Vera L. te Velde**, “Dynamic Pricing in the Airline Industry,” in T.J. Hendershott, ed., *Handbook on Economics and Information Systems*, Elsevier Handbooks in Information Systems 2007.
- Miravete, Eugenio J.**, “Screening Consumers Through Alternative Pricing Mechanisms,” *Journal of Regulatory Economics*, 1996, 9 (2), 111–132.
- , “The Welfare Performance of Sequential Pricing Mechanisms,” *International Economic Review*, 2005, 46 (4), 1321–1360.
- Pavan, Alessandro, Ilya Segal, and Juuso Toikka**, “Dynamic Mechanism Design: Incentive Compatibility, Profit Maximization and Information Disclosure,” Working Paper 2009.
- Piccione, Michele and Ariel Rubinstein**, “The Absent-Minded Driver’s Paradox: Synthesis and Responses,” *Games and Economic Behavior*, 1997, 20 (1), 121–130.
- and —, “On the Interpretation of Decision Problems with Imperfect Recall,” *Games and Economic Behavior*, 1997, 20 (1), 3–24.
- Reiss, Peter C. and Matthew W. White**, “Household Electricity Demand, Revisited,” *The Review of Economic Studies*, 2005, 72 (3), 853–883.
- Riordan, Michael H. and David E. M. Sappington**, “Awarding Monopoly Franchises,” *American Economic Review*, 1987, 77 (3), 375–387.
- Rochet, Jean-Charles and Lars A. Stole**, “Nonlinear Pricing with Random Participation,” *The Review of Economic Studies*, 2002, 69 (1), 277–311.
- Saez, Emmanuel**, “Do Taxpayers Bunch at Kink Points?,” Working Paper June 2002.
- Sandroni, Alvaro and Francesco Squintani**, “Overconfidence, Insurance and Paternalism,” *American Economic Review*, 2007, 97 (5), 1994–2004.
- Sidel, Robin and Dan Fitzpatrick**, “End Is Seen to Free Checking,” June 16 2010.
- Stango, Victor and Jonathan Zinman**, “What do Consumers Really Pay on Their Checking and Credit Card Accounts? Explicit, Implicit, and Avoidable Costs,” *American Economic Review Papers and Proceedings*, 2009, 99 (2).
- and —, “Limited and Varying Consumer Attention: Evidence from Shocks to the Salience of Overdraft Fees,” July 2010.

- Stole, Lars A.**, “Price Discrimination and Competition,” in Mark Armstrong and Robert K. Porter, eds., *Handbook of Industrial Organization*, Vol. Volume 3, Elsevier, 2007, pp. 2221–2299.
- Weyl, Eric G. and Michal Fabinger**, “Pass-Through as an Economic Tool,” *SSRN eLibrary*, 2009.
- Wilson, Robert B.**, *Nonlinear Pricing*, New York, NY: Oxford University Press, 1993.

## A Proofs

### A.1 Derivation of equation (2)

Given  $v_2^* = p_2 + q_1 p_3$ , the expected utility from choosing first period threshold  $v_1^*$  is:

$$U(v_1^*) = v_0 - p_0 + \int_{v_1^*}^1 \left( v_1 - p_1 + \int_{p_2 + p_3}^1 (v_2 - p_2 - p_3) f(v_2) dv_2 \right) f(v_1) dv_1 + F(v_1^*) \int_{p_2}^1 (v_2 - p_2) f(v_2) dv_2.$$

The first order condition,

$$\frac{dU}{dv_1^*} = f(v_1^*) \left( -v_1^* + p_1 + \int_{p_2}^{p_2 + p_3} (v_2 - p_2) f(v_2) dv_2 + (1 - F(p_2 + p_3)) p_3 \right) = 0,$$

yields equation (2). Moreover, this identifies the global maximum since for  $v^* > p_1 + \int_{p_2}^{p_2 + p_3} (v_2 - p_2) f(v_2) dv_2 + (1 - F(p_2 + p_3)) p_3$ ,  $\frac{dU}{dv^*} < 0$  and vice-versa.

### A.2 Proof of Proposition 2

Firm profits can be written as  $\Pi = G(U)(S - U)$ . For any fixed utility offer  $U$ , profits are maximized by choosing marginal prices  $p_1$ ,  $p_2$ , and  $p_3$  to achieve first best surplus, while adjusting the fixed fee  $p_0$  to keep  $U$  constant. The offered utility  $U$  is set via the fixed fee  $p_0$  to balance rent extraction versus participation, as in a basic monopoly pricing problem. Given attentive consumers and continuous taste shocks,  $p_1 = p_2 = c$  and  $p_3 = 0$  are the unique marginal prices which achieve  $S^{FB}$ . Given inattentive consumers and continuous taste shocks, any marginal prices which implement  $v^* = c$  are optimal. These include all marginal prices which satisfy  $p_3 \geq 0$  and equation (4) at  $c = v^*$  since equation (4) is sufficient as well as necessary for incentive compatibility given  $p_3 \geq 0$ .

### A.3 Proof of Proposition 3

The results in the paper are stated for the case  $T = 2$ . However, the proofs in this section are written for the more general case  $T \geq 1$ .

At time 0, consumers receive signals  $s \in \{L, H\}$  ( $\Pr(s = H) = \beta$ ) and choose a tariff  $\hat{s}$ . At time  $t \in \{1, 2, \dots, T\}$  consumers realize taste shock  $v_t \mid s \sim^{iid} F_s(v_t)$  make report  $\hat{v}_t$  and receive allocation  $q_t(\hat{s}, \hat{v}^t) \in [0, 1]$  (the probability of receiving the unit), where  $\hat{v}$  is the vector of reports to date  $[\hat{v}_1, \dots, \hat{v}_t]$ . At time  $T$ , consumers pay  $P(\hat{s}, \hat{v}^T)$ . Define  $U_t(s, \hat{s}, v^t, \hat{v}^t)$  to be expected utility at time  $t$  conditional on realizations  $(s, v^t)$  and reports  $(\hat{s}, \hat{v}^t)$  to date as well as a plan to report truthfully

from  $t + 1$  onwards. Consumers utility is quasi-linear and time separable, with unit demand each period, so that  $U_T(s, \hat{s}, v^T, \hat{v}^T) = \sum_{t=1}^T v_t q_t(\hat{s}, \hat{v}^t) - P(\hat{s}, \hat{v}^T)$ . Moreover let

$$U_t(s, v^t, \hat{s}, \hat{v}^t) = E \left[ v_t q_t(\hat{s}, \hat{v}^t) + \sum_{\tau=t+1}^T v_\tau q_\tau(\hat{s}, \hat{v}^t, v_{t+1} + \dots + v_\tau) - P(\hat{s}, \hat{v}^t, v_{t+1} + \dots + v_T) \mid (s, v^t, \hat{s}, \hat{v}^t) \right]$$

for  $t \geq 1$  and let  $U_0(s, \hat{s})$  be the expected utility of someone who has signal  $s$ , reports  $\hat{s}$ , and reports all  $v^T$  truthfully:

$$U_0(s, \hat{s}) = E \left[ \sum_{t=1}^T v_t q_t(\hat{s}, v^t) - P(\hat{s}, v^T) \mid s, \hat{s} \right].$$

Let  $U_{s\hat{s}} = U_0(s, \hat{s})$  and  $U_s = U_0(s, s)$  be the expected utility of someone who plans to be entirely truthful conditional on realization of signal  $s$ . Let  $G_s(U_s)$  be the fraction of consumers of type  $s$  with outside option below  $U_s$ . Let costs be  $C(q^T) = c \sum_{t=1}^T q_t$ , so that surplus is  $\sum_{t=1}^T (v_t - c) q_t$ . Define  $S_s$  to be the expected surplus from a type  $s$  consumer who reports truthfully:  $S_s = E \left[ \sum_{t=1}^T (v_t - c) q_t(s, v^t) \mid s \right]$ .

Invoking the revelation principle, the monopolist's problem may then be written as:

$$\max_{\substack{q^T(s, v^T) \in [0, 1] \\ P(s, v^T)}} (1 - \beta) G_L(U_L) (S_L - U_L) + \beta G_H(U_H) (S_H - U_H)$$

such that

1. Truthful history IC  $t \geq 1$   $U_t(s, s, v^t, v^t) \geq U_t(s, s, v^t, [v^{t-1}, \hat{v}_t]) \quad \forall t, s, v^t$  and  $\hat{v}_t$
2. Truthful history IC  $t = 0$   $U_0(s, s) \geq U_0(s, \hat{s}) \quad \forall s, \hat{s} \in S$
3. Any history IC  $U_t(s, s, v^t, v^t) \geq U_t(s, \hat{s}, v^t, \hat{v}^t) \quad \forall t, s, \hat{s}, v^t$  and  $\hat{v}^t$

**Lemma 2 Monotonicity:** *A necessary condition for incentive compatibility is that  $q_t(s, v^t)$  be non-decreasing in  $v_t$ .*

**Proof.** This follows from conditional independence of  $v_t$  and the single crossing property of the unit demand preferences. Consider two customers who have both truthfully reported identical  $(s, v^{t-1})$ , but have drawn different  $v_t > v'_t$  and are considering what reports  $\hat{v}_t$  to make. Conditional on making the same report  $\hat{v}_t$  at time  $t$ , the fact that both had different realizations of the true  $v_t$  at time  $t$  does not effect incentives going forward. Both face the same contract and the same distribution over future taste shocks. They will follow the same reporting strategy for  $t+1$  onwards. As a result, the difference  $U_t(s, s, v^{t-1}, v_t, v^{t-1}, \hat{v}_t) - U_t(s, s, v^{t-1}, v'_t, v^{t-1}, \hat{v}_t)$  is equal to  $q_t(s, v^{t-1}, \hat{v}_t) (v_t - v'_t)$ . Thus if  $q_t(s, v^{t-1}, v_t) < q_t(s, v^{t-1}, v'_t)$  and type  $v'_t$  prefers reporting  $v_t$  over  $v'_t$ , then  $v_t$  would prefer to report  $v'_t$  over  $v_t$  by at least  $(q_t(s, v^{t-1}, v'_t) - q_t(s, v^{t-1}, v_t)) (v_t - v'_t) > 0$  which violates incentive

compatibility. ■

**Lemma 3** *Local IC: A necessary condition for incentive compatibility is  $\frac{d}{dv_t}U_t(s, s, v^t) = \frac{\partial}{\partial v_t}U_t(s, s, v^t) = q_t(s, v^t)$ .*

**Proof.** This follows from conditional independence of  $v_t$  and application of an envelope theorem, which is valid by Proposition 1 of Pavan et al. (2009), since my setting fits within the Pavan et al. (2009) framework for  $t \geq 1$ . ■

I begin by solving a relaxed problem. I impose monotonicity ( $q_t(s, v^t)$  non-decreasing in  $v_t$ ), local incentive compatibility for  $t \geq 1$  ( $\frac{d}{dv_t}U_t(s, s, v^t) = q_t(s, v^t)$ ), and an ex ante incentive constraints IC-H ( $U_0(H, H) \geq U_0(H, L)$ ) and IC-L:  $U_0(L, L) \geq U_0(L, H)$ . However I relax all other incentive constraints. In particular, I am only checking incentive compatibility against one step deviations, rather than multiple step deviations. After solving the relaxed problem, I will need to check (1) incentive compatibility against multiple step deviations and (2) for global incentive compatibility of  $v^T$  reporting to confirm that the relaxed solution solves the original problem.

By the envelope condition  $\frac{d}{dv_t}U_t(s, s, v^t) = \frac{\partial}{\partial v_t}U_t(s, s, v^t) = q_t(s, v^t)$  and the FTC,

$$U_t(s, s, v^t) = U_t(s, s, v^{t-1}, \underline{v}_t) + \int_{\underline{v}_t}^{v_t} q_t(s, v^{t-1}, x) dx.$$

Moreover, since

$$U_T(s, \hat{s}, v^t, \underline{v}_{t+1} \dots \underline{v}_T) = \sum_{\tau=0}^t v_\tau q_\tau(\hat{s}, v^\tau) + \sum_{\tau=t+1}^T v_\tau q_\tau(\hat{s}, v^t, \underline{v}_{t+1}, \dots, \underline{v}_\tau) - P(\hat{s}, v^t, \underline{v}_{t+1} \dots \underline{v}_T),$$

it is true that

$$\begin{aligned} \frac{d}{dv_t}U_T(s, \hat{s}, v^t, \underline{v}_{t+1} \dots \underline{v}_T) &= q_t(\hat{s}, v^t) + \sum_{\tau=0}^t v_\tau \frac{d}{dv_t}q_\tau(\hat{s}, v^\tau) + \sum_{\tau=t+1}^T v_\tau \frac{d}{dv_t}q_\tau(\hat{s}, v^t, \underline{v}_{t+1}, \dots, \underline{v}_\tau) \\ &\quad - \frac{d}{dv_t}P(\hat{s}, v^t, \underline{v}_{t+1} \dots \underline{v}_T) \\ &= \frac{d}{dv_t}U_T(\hat{s}, \hat{s}, v^t, \underline{v}_{t+1} \dots \underline{v}_T). \end{aligned}$$

Thus, by FTC

$$U_T(s, \hat{s}, v^t, \underline{v}_{t+1} \dots \underline{v}_T) = U_T(s, \hat{s}, v^{t-1}, \underline{v}_t \dots \underline{v}_T) + \int_{\underline{v}_t}^{v_t} q_t(\hat{s}, v^{t-1}, x) dx,$$

and

$$U_T(s, \hat{s}, v^T) = U_T(s, \hat{s}, v^{T-1}, \underline{v}_T) + \int_{\underline{v}_T}^{v^T} q_T(\hat{s}, v^{T-1}, x) dx.$$

Now by iterated expectations,

$$U_{T-1}(s, \hat{s}, v^{T-1}) = \int_{\underline{v}_T}^{\bar{v}_T} U_T(s, \hat{s}, v^T) f_s(v_T) dv_T.$$

Substituting in the envelope condition and integrating by parts gives

$$U_{T-1}(s, \hat{s}, v^{T-1}) = U_T(s, \hat{s}, v^{T-1}, \underline{v}_T) + \int_{\underline{v}_T}^{\bar{v}_T} q_T(\hat{s}, v^T) (1 - F_s(v_T)) dv_T.$$

This can now be repeated recursively to yield

$$U_0(s, \hat{s}) = U_T(s, \hat{s}, \underline{v}_1, \underline{v}_2, \dots, \underline{v}_T) + \sum_{t=1}^T \int_{\underline{v}_t}^{\bar{v}_t} q_t(\hat{s}, v^t) (1 - F_s(v_t)) dv_t. \quad (26)$$

Equation (26) pins down  $U_{HL}$  as a function of the allocation  $q^T(L, v^T)$  and  $U_L$ , and thus IC-H is:

$$U_H \geq U_L + \sum_{t=1}^T \int_{\underline{v}_t}^{\bar{v}_t} q_t(L, v^t) (F_L(v_t) - F_H(v_t)) dv_t$$

Similarly, IC-L is

$$U_L \geq U_H - \sum_{t=1}^T \int_{\underline{v}_t}^{\bar{v}_t} q_t(H, v^t) (F_L(v_t) - F_H(v_t)) dv_t$$

or together:

$$\sum_{t=1}^T \int_{\underline{v}_t}^{\bar{v}_t} q_t(L, v^t) (F_L(v_t) - F_H(v_t)) dv_t \leq U_H - U_L \leq \sum_{t=1}^T \int_{\underline{v}_t}^{\bar{v}_t} q_t(H, v^t) (F_L(v_t) - F_H(v_t)) dv_t$$

Note that given FOSD, monotonicity of  $q_t(s, v^t)$  in  $s$ ,  $q_t(H, v^t) \geq q_t(L, v^t)$ , implies IC-L follows from binding IC-H and vice versa. Also, given FOSD, IC-H implies  $U_H \geq U_L$ . Finally, either binding IC-H or binding IC-L implies  $\frac{dU_H}{dU_L} = 1$ .

**Lemma 4** (1) IC-L slack implies  $q_t(H, v^t) = q^{FB}(v_t)$  and (2) IC-H slack implies  $q_t(L, v^t) = q^{FB}(v_t)$ .

**Proof.** (1) Suppose not. Then moving  $q_t(H, v^t)$  towards  $q^{FB}(v_t)$  a little bit while maintaining monotonicity in  $v^t$  and keeping  $U_H$  constant keeps IC-H unaffected, participation unaffected, but increases profit from type  $H$ , without violating IC-L since it has been relaxed. (2) Similar argument.

■

I now solve the problem separately for three cases.

**Case (1)**,  $\mu_L^* = \mu_H^*$ . Relax both IC-L and IC-H. Then allocations are first best and unconstrained optimal markups  $\mu_L^*$ ,  $\mu_H^*$  are charged on each contract. Since both allocations and markups are the same, the L and H contracts are the same, and hence IC-L and IC-H are satisfied.

**Case (2)**  $\mu_H^* > \mu_L^*$ . Relax IC-L. By Lemma 4,  $q_t(H, v^t) = q^{FB}(v_t)$ .

**Lemma 5** *Relaxed IC-L and  $\mu_H^* > \mu_L^*$  imply IC-H is binding such that  $U_H = U_{HL}$  in the relaxed problem.*

**Proof.** Suppose IC-H is slack:  $U_H > U_{HL}$ . Given IC-L is relaxed, Lemma 4 implies  $q_t(H, v^t) = q_t(L, v^t) = q^{FB}(v_t)$ . The slack IC-H therefore reduces to  $U_H - U_L > 2 \int_c^{\bar{v}} (F_L(v_t) - F_H(v_t)) dv_t = S_H^{FB} - S_L^{FB}$ .

(a) ZOOM: Given  $U_H - U_L > S_H^{FB} - S_L^{FB} > 0$ , IR-L ( $U_L \geq 0$ ) implies IR-H is slack ( $U_H > 0$ ). Hence the firm can raise  $p_{0H}$  and raise profits without violating IC-H or IR-H. This contradicts optimality.

(b) HOO: Profit maximization requires  $\frac{\partial \Pi}{\partial U_H} = \frac{\partial \Pi}{\partial U_L} = 0$ , since otherwise the firm could adjust  $U_H$  or  $U_L$  in either direction to raise profits without violating either constraint (since IC-H is slack and IC-L is relaxed). By assumption,  $U_s + \frac{G_s(U_s)}{g_s(U_s)}$  is increasing, so both first order conditions have unique solutions. Given the first best allocations, these solutions are the unconstrained optimal markups  $\mu_H^*$  and  $\mu_L^*$ . So  $\mu_H^* > \mu_L^*$  implies  $S_H^{FB} - U_H > S_L^{FB} - U_L$  or  $U_H - U_L < S_H^{FB} - S_L^{FB}$ , which contradicts IC-H slack. ■

Now we can simplify the doubly relaxed problem, by substituting the local incentive constraints summarized by equation (26) for  $U_0(s, s)$  into the objective function, eliminating marginal fees from the problem. (Fixed fees depend on  $U_T(s, \hat{s}, \underline{v}_1, \underline{v}_2, \dots, \underline{v}_T)$ ):

$$\max_{\substack{q^T(L, v^T) \geq 0 \\ U_L}} (1 - \beta) G_L(U_L) (S_L - U_L) + \beta G_H(U_H) (S_H^{FB} - U_H)$$

such that

1. IC-H  $U_H = U_L + \sum_{t=1}^T \int_{\underline{v}_t}^{\bar{v}_t} q_t(L, v^t) (F_L(v_t) - F_H(v_t)) dv_t$
2. monotonicity  $q_t(s, v^t)$  non-decreasing in  $v_t$ .

Some ingredients to the  $q_t(L, v^t)$  first order condition are:

$$\frac{dU_H}{dq_t(L, v^t)} = \int_{\underline{v}_t}^{\bar{v}_t} (F_L(v_t) - F_H(v_t)) dv_t,$$

$$\frac{dS_L}{dq_t(L, v^t)} = \int_{\underline{v}_t}^{1_t} (v_t - c) f_L(v_t) dv_t.$$

The  $q_t(L, v^t)$  first order condition is:

$$\begin{aligned}
\frac{d\Pi}{dq_t(L, v^t)} &= \frac{\partial\Pi}{\partial q_t(L, v^t)} + \frac{\partial\Pi}{\partial U_H} \frac{dU_H}{dq_t(L, v^t)} = \frac{\partial\Pi}{\partial q_t(L, v^t)} + \frac{\partial\Pi}{\partial U_H} \int_{\underline{v}_t}^{\bar{v}_t} (F_L(v_t) - F_H(v_t)) dv_t \\
&= (1 - \beta) G_L(U_L) \int_{\underline{v}_t}^{\bar{v}_t} (v_t - c) f_L(v_t) dv_t + \frac{\partial\Pi}{\partial U_H} \int_{\underline{v}_t}^{\bar{v}_t} (F_L(v_t) - F_H(v_t)) dv_t \\
&= (1 - \beta) G_L(U_L) \int_{\underline{v}_t}^{\bar{v}_t} \left( v_t - c - \frac{\beta}{1 - \beta} \frac{F_L(v_t) - F_H(v_t)}{f_L(v_t)} \frac{-\partial\Pi/\partial U_H}{\beta G_L(U_L)} \right) f_L(v_t) dv_t.
\end{aligned}$$

Define  $v_L^A$  such that:

$$v_L^A = c + \frac{\beta}{1 - \beta} \frac{F_L(v_L^A) - F_H(v_L^A)}{f_L(v_L^A)} \frac{-\partial\Pi/\partial U_H}{\beta G_L(U_L)}. \quad (27)$$

Maximizing point-wise implies that  $q_t(L, v^t) = 1$  if

$$\left( v_t - c - \frac{\beta}{1 - \beta} \frac{F_L(v_t) - F_H(v_t)}{f_L(v_t)} \frac{-\partial\Pi/\partial U_H}{\beta G_L(U_L)} \right) > 0.$$

(which is equivalent to  $v_t > v_L^A$  if  $v_L^A$  is unique) and zero otherwise. If  $v_L^A$  is not unique, the solution is obtained by ironing because monotonicity is necessary. Ironing will still yield a threshold, such that the allocation rule is  $q_t(s, v^t) = 1_{v_t \geq \phi(s, v^{t-1})}$  for some  $\phi$ . Picking a threshold to maximize the integral will mean picking one of the fixed points of 27. (If it weren't at a fixed point, we could always improve profits by moving towards one or another without violating monotonicity.) Thus

$$v_L^A = \arg \max_{v_L} \left\{ (1 - \beta) G_L(U_L) \int_{v_L}^{\bar{v}} (v - c) f_L(v) dv + \frac{\partial\Pi}{\partial U_H} \int_{v_L}^{\bar{v}} (F_L(v) - F_H(v)) dv \right\}.$$

Since we are assuming that IC-L is slack, there is no constraint on raising  $U_H$  (which only relaxes IC-H). Therefore we know that  $\frac{\partial\Pi}{\partial U_H} \leq 0$ , and hence (given FOSD)  $q_t(L, v^t)$  is weakly distorted downwards ( $q_t(L, v^t) \leq q^{FB}(v^t)$ ). Moreover,  $\mu_H^* > \mu_L^*$  implies that  $\frac{\partial\Pi}{\partial U_H} < 0$  and therefore that the distortion is strict ( $v_L^A > c$ ). (1) For ZOOM,  $\frac{\partial\Pi}{\partial U_H} = -\beta$ . (2) For HOO,  $\frac{d\Pi}{dU_L} = 0$  implies  $\frac{\partial\Pi}{\partial U_L} = -\frac{\partial\Pi}{\partial U_H}$ , so we can have  $\frac{\partial\Pi}{\partial U_H} = 0$  only if  $\frac{\partial\Pi}{\partial U_L} = 0$ . Given first best allocations and decreasing marginal revenue, this means, charging unconstrained optimal markups  $\mu_L^*$  and  $\mu_H^*$ . However, with identical allocations, IC-H requires that the H contract have a weakly lower markup, contradicting  $\mu_H^* > \mu_L^*$ .

Next, the downward distortion on the L contract implies monotonicity in signals:  $q_t(L, v^t) \leq q_t(H, v^t)$ . Therefore, by FOSD, binding IC-H implies IC-L, so it was okay to relax IC-L.

What about the optimal  $U_L$ ? There are two cases to consider.

(1) ZOOM (zero outside option monopoly where  $G_s = 1_{U_s \geq 0}$ ). Then  $U_L = 0$  and  $G_L(U_L) = 1$

(assuming  $\beta$  small enough that  $L$  types served) and also  $G_H(U_H) = 1$  since IC-H implies IR. (H types can always mimic L types, but would have stochastically higher values for the same consumption by FOSD. Alternatively,  $\sum_{t=1}^T \int_{v_t}^{\bar{v}_t} q_t(L, v^t) (F_L(v_t) - F_H(v_t)) dv_t \geq 0$  by FOSD and  $q \in [0, 1]$ ). In this case  $\frac{\partial \Pi}{\partial U_H} = -\beta$  and the first order condition for  $q_t(L, v^t)$  reduces to

$$v_L^{CL} = c + \frac{\beta}{1 - \beta} \frac{F_L(v_L^{CL}) - F_H(v_L^{CL})}{f_L(v_L^{CL})},$$

which is the Courty and Li (2000) solution.

(2) Heterogeneous outside options ( $G_s$  differentiable etc.). Then

$$\begin{aligned} \frac{\partial \Pi}{\partial U_L} &= (1 - \beta) g_L(U_L) (S_L - U_L) - (1 - \beta) G_L(U_L) \\ \frac{\partial \Pi}{\partial U_H} &= \beta g_H(U_H) (S_H^{FB} - U_H) - \beta G_H(U_H) \end{aligned}$$

and the first order condition for  $U_L$  is

$$\frac{\partial \Pi}{\partial U_L} = -\frac{\partial \Pi}{\partial U_H}$$

**Case (3)**  $\mu_H^* < \mu_L^*$ : This follows symmetrically to Case (2).

So, I've solved the relaxed problem in three cases. Does it solve the original problem? I have two constraints left to check:

(1) Multiple-step deviation IC: Notice that  $q_t(s, v^t) = q(s, v_t)$ . In other words, allocations depend on  $\hat{s}$  and  $v_t$  only, they do not depend on  $t$  or on  $v^{t-1}$ . As a result, one-step deviation IC is sufficient for multiple-step deviation IC.

(2) Global IC on  $v^T$ : Allocations  $q_s(s, v^t) = 1_{v_t \geq v_s^A}$  are simply implementable with two-part tariffs with marginal price  $v_s^A$ . These are globally IC by inspection.

#### A.4 Proof of Corollary 2

(1)  $\tau_H = \tau_L = \tau$ : Firm  $A$ 's residual demand from consumers of type  $s$  is  $G_s(U_s^A) = \frac{1}{2\tau_s} (U_s^A - U_s^B + \tau_s)$ .

In the proposed symmetric equilibrium, this implies  $\mu_s^* = \tau_s$ . Proposition 3 implies firm offers are best responses to each other. There are no other symmetric pure strategy equilibria, since with any set of symmetric offers  $\mu_s^* = \tau_s$ .

(2) If  $\tau_H \neq \tau_L$ , then all equilibria are inefficient: Suppose not, and in equilibrium we have efficient allocations. Then  $p_{3s}^i = 0$  and  $p_{1s}^i = p_{2s}^i = c$ . This means that we need to have  $p_{0L}^i = p_{0H}^i$ . As a result  $\mu_H^i = \mu_L^i = \mu^i$ . These statements hold for any offer in B's mixed strategy. A's expected

market share in segment  $s$  is therefore  $\frac{1}{2\tau_s} (E[\mu^B] - \mu_s^A + \tau_s)$ , and A's best response markup is  $\mu_s^{*A} = \frac{1}{2} (E[\mu^B] + \tau_s)$ . Thus  $\mu_L^{*A} \neq \mu_H^{*A}$  and by Proposition 3 A's best response includes an inefficient contract.

(3a) If  $\tau_H > \tau_L$ , then in all symmetric equilibria, high types receive first best allocations, while low types' allocation is distorted downwards below first best: We know that in a symmetric pure strategy equilibrium that for both firms, either  $\mu_L^* = \mu_H^*$ ,  $\mu_L^* < \mu_H^*$  or  $\mu_L^* > \mu_H^*$ . Part (2) rules out  $\mu_L^* = \mu_H^*$  if  $\tau_H > \tau_L$ . All that remains is to rule out  $\mu_L^* > \mu_H^*$ . This is ruled out by the assumption that the pass through rate is less than 1, which implies markups are strategic complements (Weyl and Fabinger 2009, Bulow, Geanakoplos and Klemperer 1985): Let  $\mu_s^{**}$  be the optimal markup unconstrained by ex ante IC at current allocation (i.e. that solves  $\frac{\partial \Pi}{\partial U_s} = \beta_s g_s(U_s) \left( S_s - U_s - \frac{G_s(U_s)}{g_s(U_s)} \right) = 0$  at the current allocation). I have assumed demand has a pass through rate less than 1 ( $\frac{G_s(U_s)}{g_s(U_s)}$  is increasing). This implies that  $\mu_s^{**}(S_s)$  is increasing in  $S_s$  and  $\mu_s^{**} \leq \mu_s^*$ . Also, if  $S_s - U_s < \frac{G_s(U_s)}{g_s(U_s)}$  then  $\mu_s^{**} < \frac{G_s(U_s)}{g_s(U_s)}$ . In any symmetric equilibrium,  $\frac{G_s(U_s)}{g_s(U_s)} = \tau_s$ , so if  $\frac{\partial \Pi}{\partial U_s} < 0$  then  $\mu_s < \mu_s^{**} < \tau_s$  and vice versa. Supposing  $\mu_L^* > \mu_H^*$ , then by Proposition 3, low contracts are first best. Hence  $\mu_L^{**} = \mu_L^*$  while  $\mu_H^{**} \leq \mu_H^*$ . Also,  $\frac{\partial \Pi}{\partial U_L} < 0$  and  $\frac{\partial \Pi}{\partial U_H} > 0$ , so  $\mu_L^{**} < \tau_L$  and  $\mu_H^{**} > \tau_H$ . Putting these together with  $\tau_L < \tau_H$  gives

$$\mu_L^* = \mu_L^{**} < \tau_L < \tau_H < \mu_H^{**} \leq \mu_H^*$$

which contradicts  $\mu_L^* > \mu_H^*$ .

(3b)  $\tau_H < \tau_L$  follows a symmetric argument.

## A.5 Proof of Proposition 5

Proposition 5 is stated for either of two restrictions: (1)  $p_{3s} \leq p^{\max}$  or (2)  $p_{3s} \leq v_s / (1 - F_s(v_s))$ . Both can be written as  $p_{3s} \leq h(v_s)$  for some  $h(v_s)$  that is strictly positive and non-decreasing. All but the last step of the proof work with the restrictions in this general form.

I. First consider half the parameter space:  $\mu_H^* \geq \mu_L^*$ .

By equation (10), IC-H is given by equation (28):

$$U_H \geq U_{HL} = U_L + 2 \int_{v_{HL}}^{\bar{v}} (v - v_L) dF_H(v) - 2 \int_{v_L}^{\bar{v}} (v - v_L) dF_L(v) - p_{3L} (F_L(v_L) - F_H(v_{HL}))^2 \quad (28)$$

Relax IC-L. There are two cases: either (1) IC-H is slack or (2) IC-H binds.

Case (1), IC-H is slack.

(a) Show that IC-L is satisfied. Since IC-L is relaxed by increasing  $p_{3H}$ , it is sufficient to check

at  $p_{3H} = 0$ . If both IC-L and IC-H are slack, then  $v_L = v_H = c$  and at  $p_{3H} = 0$ ,  $P_H(q_1, q_2) = T + c(q_1 + q_2)$ , so that  $U_H = S_H^{FB} - T$  and  $U_{LH} = S_L^{FB} - T$ . Thus IC-L,  $U_L \geq U_{LH}$ , is equivalent to  $S_H^{FB} - U_H \geq S_L^{FB} - U_L$ , or  $\mu_H^* \geq \mu_L^*$  at optimal offer  $\{U_H, U_L\}$  which is satisfied by assumption.

(b) Substituting  $v_L = c$  into equation (28), gives

$$U_H \geq U_{HL} = (U_L - S_L^{FB}) + 2 \int_{v_{HL}}^{\bar{v}} (v - c) dF_H(v) - p_{3L} (F_L(c) - F_H(v_{HL}))^2.$$

Noting that  $p_{3L}$  can be set to the maximum  $h_L(c)$ ,  $2 \int_{v_{HL}}^{\bar{v}} (v - c) dF_H(v) = S_H^{FB} - 2 \int_c^{v_{HL}} (v - c) dF_H(v)$ , and by definition at optimal utility offers,  $(S_H^{FB} - \hat{U}_H) - (S_L^{FB} - \hat{U}_L) = \mu_H^* - \mu_L^*$ , IC-H simplifies to:

$$(\mu_H^* - \mu_L^*) \leq 2 \int_c^{v_{HL}} (v - c) dF_H(v) + h_L(c) (F_L(c) - F_H(v_{HL}))^2.$$

Further,  $v_{HL} = c + h_L(c) (F_L(c) - F_H(v_{HL}))$  is uniquely defined by the FOC from equation (5) for  $v_{HL}$ , where  $\bar{p}_L$  is given by equation (9) and  $p_{3L} = h_L(c)$ . Note, if instead  $p_{3L} = 0$ , then IC-H reduces to  $(\mu_H^* - \mu_L^*) = 0$ . So we need  $p_{3L} > 0$  for any  $\mu_H^* > \mu_L^*$  even if IC-H doesn't bind at optimal prices - because it would bind at  $p_{3L} = 0$ .

Case (2), IC-H binds. If equation (12) is not satisfied, then IC-H cannot be relaxed. Moreover, it will bind with equality given either ZOOM (where  $\partial\Pi/\partial U_H = -\beta$  for all  $U_H > 0$ ) or HOO (where decreasing marginal revenue assumption,  $U_s + \frac{G_s(U_s)}{g_s(U_s)}$  increasing, implies concavity).

(a) Show  $v_H = c$ , derive FOC for  $v_L$ , and show  $v_L \geq v_H$ : (i)  $v_H = c$ : If IC-L is relaxed, then it is optimal to choose  $v_H = c$ . Moving from  $\hat{v}_H$  a little bit towards  $c$  holding  $U_H$  fixed increases profits from high types, leaves IC-H unaffected, and we are ignoring IC-L. (Note that at  $p_{3H} = 0$ , we will also have  $v_{LH} = c$ .) (ii) FOC for  $v_L$ : The profit maximization problem is

$$\begin{aligned} & \max_{U_L, v_L, p_{3L}} ((1 - \beta) G_L(U_L) (S_L - U_L) + \beta G_H(U_H) (S_H^{FB} - U_H)) \\ \text{s.t. } U_H &= U_L + 2 \int_{v_{HL}}^{\bar{v}} (v - v_L) dF_H(v) - 2 \int_{v_L}^{\bar{v}} (v - v_L) dF_L(v) - p_{3L} (F_L(v_L) - F_H(v_{HL}))^2 \\ S_L &= 2 \int_{v_L}^{\bar{v}} (v - c) dF_L(v) \\ p_{3L} &= h(v_L) \end{aligned}$$

By the envelope condition,  $\frac{dU_H}{dv_L} = \frac{\partial U_H}{\partial v_L}$ , so the FOC for  $v_L$  is:

$$\frac{d\Pi}{dv_L} = \frac{\partial\Pi}{\partial S_L} \frac{dS_L}{dv_L} + \frac{\partial\Pi}{\partial U_H} \left( \frac{\partial U_H}{\partial v_L} + \frac{\partial U_H}{\partial p_{3L}} h'(v_L) \right)$$

Taking derivatives and substituting equation (11) for  $\frac{\partial U_H}{\partial p_{3L}}$  gives:

$$\begin{aligned} \frac{d\Pi}{dv_L} &= -2(1-\beta)G_L(U_L)(v_L-c)f_L(v_L) \\ &\quad - \frac{\partial \Pi}{\partial U_H} \left[ 2(F_L(v_L) - F_H(v_{HL}))(1+p_{3L}f_L(v_L)) + (F_L(v_L) - F_H(v_{HL}))^2 h'(v_L) \right]. \end{aligned}$$

The FOC  $\frac{d\Pi}{dv_L} = 0$  simplifies to equation (13), or for non-negative marginal prices,  $h_s(v_s) = v_s/(1-F_s(v_s))$ , to:

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

Similar to case (1),  $v_{HL} = v_L + h_L(v_L)(F_L(v_L) - F_H(v_{HL}))$ , follows from equations (9) and (5).

Since IC-H is binding,  $\frac{\partial \Pi}{\partial U_H} \leq 0$ . (ZOOM:  $\frac{\partial \Pi}{\partial U_H} = -\beta$ , HOO:  $\frac{\partial \Pi}{\partial U_H} = \beta g_H(U_H) \left( S_H^{FB} - U_H - \frac{G_H(U_H)}{g_H(U_H)} \right)$ .)

Moreover,  $F_L(v_L) - F_H(v_{HL}) \geq 0$ , so  $v_L \geq c = v_H$ . (Why is  $F_L(v_L) - F_H(v_{HL}) \geq 0$ ? Suppose not. Then by  $v_{HL} = v_L + h(v_L)(F_L(v_L) - F_H(v_{HL}))$  and  $h(v_L) > 0$  we would have  $v_{HL} < v_L$  and hence  $F_L(v_L) - F_H(v_{HL}) > F_L(v_L) - F_H(v_L) \geq 0$  by FOSD. Contradiction.)

(b) Show that IC-L is satisfied: Suppose that  $\{U_L, v_L, p_{3L}, v_H, p_{3H}\}$  is the relaxed solution, with IC-H binding so that equation (28) holds with equality. Now consider the alternative contract menu  $\{U_L, v_L, \hat{p}_{3L} = 0, \hat{U}_H, v_H, \hat{p}_{3H} = 0\}$  with  $\hat{U}_H = \left( 2 \int_{v_{HL}}^{\bar{v}} (v - v_L) dF_H(v) - 2 \int_{v_L}^{\bar{v}} (v - v_L) dF_L(v) + U_L \right)$  (preserving IC-H with equality). In this case IC-H equality, FOSD, and  $v_L \geq v_H$  imply IC-L. This is the standard logic - high types are willing to pay more ex ante for a decrease in marginal price than are low types. If high-types are just indifferent to the upgrades, then low-types won't find it worthwhile. Next move back to the original contract menu, in two steps. First adjust the  $p_{3s}$  keeping  $\hat{U}_H$  fixed. We know that this relaxes the IC constraints, so it is still IC. Second decrease  $U_H$  back to IC-H binding. The decrease in  $U_H$  relaxes IC-L still further. So it is still satisfied.

**II.** Now consider the other half of the parameter space,  $\mu_H^* \leq \mu_L^*$ . The results follow by a nearly symmetrical argument. The only important difference is that for  $\mu_H^* - \mu_L^* < -X_L$ , the first order condition for  $v_H$ ,

$$v_H = c - \frac{1-\beta}{\beta} \frac{F_L(v_{LH}) - F_H(v_H)}{f_H(v_H)} \frac{-\partial \Pi / \partial U_L}{(1-\beta)G_H(U_H)} \left( (1 + p_{3H}f_H(v_H)) - \frac{1}{2} (F_L(v_{LH}) - F_H(v_H)) h'_H(v_H) \right),$$

may call for  $v_H > c$  if  $h'_H(v_H)$  is sufficiently positive, which would violate the relaxed IC-H condition. However, the proposition is stated for  $h_H(v_H) = p^{\max}$  or for  $h_H(v_H) = \frac{v_H}{1-F_H(v_H)}$  rather than for general  $h_H(v_H)$ . In the former case there is no issue, since  $h'_H(v_H) = 0$ . In the latter case, there is an additional step to show that  $v_H < c$ . Given  $p_{3H} = h_H(v_H) = \frac{v_H}{1-F_H(v_H)}$ , the

first order condition for  $v_H$  can be re-written as

$$v_H = c - \frac{1 - \beta}{\beta} \frac{F_L(v_{LH}) - F_H(v_H)}{f_H(v_H)} \frac{-\partial\Pi/\partial U_L}{(1 - \beta)G_H(U_H)} (1 + p_{3H}f_H(v_H)) \left(1 - \frac{1}{2} \frac{(F_L(v_{LH}) - F_H(v_H))}{(1 - F_H(v_H))}\right).$$

In this form, it is apparent by inspection that  $v_H < c$ , despite  $h' > 0$ .

## A.6 Proof of Corollary 3

(1) Show proposed equilibrium exists by construction: Impose  $p_{3s} \leq h_s(v_s) = v_s/(1 - F_s(v_s))$ . Assume that each firm offers  $p_{3s} = h_s(c)$ ,  $v_L = v_H = c$ , and  $U_s = S_s^{FB} - \tau_s$ . In this case,  $U_s = \hat{U}_s$  and  $\mu_s = \mu_s^* = \tau_s$ . As a result,  $(\mu_H^* - \mu_L^*) = \tau(H - L)$ . For  $\tau$  sufficiently small, this satisfies the condition for first best allocations in Proposition 5, which verifies that the proposed offers are best responses. If the constraint  $p_{3s} \leq h_s(v_s)$  were relaxed (no such constraint was imposed in the corollary) this would still be an equilibrium.

(2) Show that no other symmetric pure strategy equilibrium exist: There are three possibilities: (a)  $(\mu_H^* - \mu_L^*) < -X_L$ , (b)  $(\mu_H^* - \mu_L^*) > X_H$ , and (c)  $(\mu_H^* - \mu_L^*) \in [-X_L, X_H]$ . Given (c), the proposed equilibrium is unique. A symmetric equilibrium in case (a) is ruled out by  $\tau_H > \tau_L$  and pass-through rate less than 1 following a similar argument that was used in the proof of Corollary 2.<sup>20</sup> I rule out a symmetric equilibrium in case (b) by showing that there would exist a profitable deviation:

Suppose a symmetric equilibrium satisfied  $(\mu_H^* - \mu_L^*) > X_H$ . Then by Proposition 5, IC-H binds and IC-L is slack. By symmetry, at the equilibrium utility offers:  $\frac{G_H}{g_H} - \frac{G_L}{g_L} = (\tau_H - \tau_L)$ . I construct a profitable menu deviation in three steps, ignoring IC-H until the end. (i) Change  $U_s$  to the unconstrained optimum at current  $S_s$ , which increases profits. This means lowering  $U_H$  and raising  $U_L$ , and relaxing IC-L. Given the pass-through less than 1 assumption, this means lowering  $\frac{G_H}{g_H} - \frac{G_L}{g_L}$ . (ii) Change  $S_L$  to  $S_L^{FB}$  for L type, which increases profits and does not affect IC-L. (We already have  $S_H = S_H^{FB}$  by  $(\mu_H^* - \mu_L^*) > X_H$ .) (iii) Change  $U_L$  to  $\hat{U}_L$ , which increases profits now that  $S_L = S_L^{FB}$ . (We already have  $U_H = \hat{U}_H$  from step (i).) The change in  $U_L$  follows an increase in  $S_L$ , so is an increase in  $U_L$  by decreasing marginal revenue, and hence relaxes IC-L. By pass-through rate less than 1, this lowers  $\frac{G_H}{g_H} - \frac{G_L}{g_L}$ . The new contract has strictly higher profits and still satisfies IC-L. The new contract menu offers unconstrained optimal markups and first best

---

<sup>20</sup>There I showed that PTR less than 1 implies (i) that  $\mu_s^{**} \leq \mu_s^*$  and (ii) that in any symmetric equilibrium if  $\frac{\partial\Pi}{\partial U_s} < 0$  then  $\mu_s < \mu_s^{**} < \tau_s$  and vice versa. If  $(\mu_H^* - \mu_L^*) < -X_L$ , then by Proposition 5, low contracts are first best. Hence  $\mu_L^{**} = \mu_L^*$  while  $\mu_H^{**} \leq \mu_H^*$ . Also,  $\frac{\partial\Pi}{\partial U_L} < 0$  and  $\frac{\partial\Pi}{\partial U_H} > 0$ , so  $\mu_L^{**} < \tau_L$  and  $\mu_H^{**} > \tau_H$ . Putting these together with  $\tau_L < \tau_H$  gives  $\mu_L^* = \mu_L^{**} < \tau_L < \tau_H < \mu_H^{**} \leq \mu_H^*$  which contradicts  $\mu_L^* > \mu_H^*$ .

allocations, so  $\hat{\mu}_s = \frac{G_s(\hat{U}_s)}{g_s(\hat{U}_s)} = \mu_s^*$ . As a result

$$\mu_H^* - \mu_L^* = \frac{G_H(\hat{U}_H)}{g_H(\hat{U}_H)} - \frac{G_L(\hat{U}_L)}{g_L(\hat{U}_L)} \leq \frac{G_H(U_H)}{g_H(U_H)} - \frac{G_L(U_L)}{g_L(U_L)} = \tau(H - L)$$

(where  $U_s$  means original utility offer, and  $\hat{U}_s$  is the unconstrained optimal utility offer used in the new menu) and by Proposition 5, IC-H is satisfied for small  $\tau$ . Thus this deviation was strictly profitable, and the proposed contract menus cannot have been an equilibrium.

(3) Total welfare result: With price-posting regulation, equilibrium pricing matches the attentive case, and Corollary 2 implies that allocations are inefficient in all equilibria for any  $\tau > 0$ . Thus PPR strictly reduces welfare.

(4) Distributional result: Without PPR,  $\frac{\partial \Pi}{\partial U_L} = -\frac{\partial \Pi}{\partial U_H} = 0$ . With PPR, Corollary 2 implies that in any symmetric pure strategy equilibrium IC-H binds and  $\frac{\partial \Pi}{\partial U_L} = -\frac{\partial \Pi}{\partial U_H} > 0$ . This implies high types win,  $U_H^{PPR} > S_H^{FB} - \tau_H = \hat{U}_H$ , but low types  $U_L^{PPR} < S_L^{PPR} - \tau_L < S_L^{FB} - \tau_L = \hat{U}_L$  are losers. Firms still split both segments equally, but now make less on high types  $S_H^{FB} - U_H^{PPR} < \tau_H$ , but more on low types  $S_L^{PPR} - U_L^{PPR} > \tau_L$ . On average firms lose money. The first order condition under PPR,  $\frac{\partial \Pi}{\partial U_L} = -\frac{\partial \Pi}{\partial U_H} > 0$ , and symmetry ( $G_s/g_s = \tau_s$ ) imply that

$$\frac{1}{2} (S_L^{PPR} - U_L^{PPR} - \tau_L) (1 - \beta) = -\frac{\tau_L}{\tau_H} \frac{1}{2} (S_H^{FB} - U_H^{PPR} - \tau_H) \beta < -\frac{1}{2} (S_H^{FB} - U_H^{PPR} - \tau_H) \beta.$$

The inequality shows that the profit gain on low types (LHS) is less than the profit loss on high types (RHS).

## A.7 Proof of Proposition 6

$$\frac{d^2 \Pi}{dv_L dp_{3L}} = 2\beta \frac{F_L(v_L) - F_H(v_{HL})}{1 + p_{3L} f_H(v_{HL})} (f_L(v_L) - f_H(v_{HL}))$$

As shown earlier,  $F_L(v_L) - F_H(v_{HL}) > 0$  and  $v_{HL} > v_L$  for any finite  $p_{3L} \geq 0$ . (1) To show that price-posting regulation weakly increases welfare, it is sufficient to show that  $\frac{d^2 \Pi}{dv_L dp_{3L}} \geq 0$  for all  $p_{3L} \geq 0$  and for all  $v_L \in [c, v^{CL}]$ . (Standard monotone comparative statics results apply since  $h'(v_L) \geq 0$ .) Given  $c < c^*$ , for  $\beta$  sufficiently small  $v^{CL} < c^*$  as well ( $v^{CL} = c + \frac{\beta}{(1-\beta)} \frac{F_L(v^{CL}) - F_H(v^{CL})}{f_L(v^{CL})}$ ). Thus for all  $v_L \in [c, v^{CL}]$ , we have  $f_L(v_L) \geq f_H(v_H) \geq f_H(v_{HL})$  by  $v^{CL} < c^*$  and  $f_H$  weakly decreasing above  $c$ . Also, for  $\beta$  sufficiently small, the low types are served with or without price-posting regulation. (2) To show that price-posting regulation weakly decreases welfare, it is sufficient to show that  $\frac{d^2 \Pi}{dv_L dp_{3L}} \leq 0$  for all  $p_{3L} \in [0, p^{\max}]$  and for all  $v_L \geq v^{CL}$ . We know that for all  $v_L \geq v^{CL} \geq c > c^*$ ,  $f_L(v_L) < f_H(v_L)$ . For  $f_H$  weakly increasing, this implies

$f_L(v_L) < f_H(v_{HL})$ . As  $p^{\max}$  goes to zero, the constraint  $p_{3L} \leq p^{\max}$  implies that  $v_{HL}$  approaches  $v_L$  and hence the inequality  $f_L(v_L) < f_H(v_{HL})$  holds. Also imposing price-posting regulation could cause low types not to be served at all.

## A.8 Proof of Proposition 7

The results in the paper are stated for the case  $T = 2$ . However, the proofs in this section are written for the more general case  $T \geq 1$ .

At time  $t \in \{1, 2, \dots, T\}$  consumers realize taste shock  $v_t \sim^{iid} F(v_t)$  make report  $\hat{v}_t$  and receive allocation  $q_t(\hat{v}^t) \in [0, 1]$  (the probability of receiving the unit), where  $\hat{v}$  is the vector of reports to date  $[\hat{v}_1, \dots, \hat{v}_t]$ . Consumers believe  $v_t \sim^{iid} F^*(v_t)$ . At time  $T$ , consumers pay  $P(\hat{v}^T)$ . Define  $U_t(v^t, \hat{v}^t)$  to be expected utility at time  $t$  with respect to  $F$ , conditional on realizations  $v^t$  and reports  $\hat{v}^t$  to date as well as a plan to report truthfully from  $t+1$  onwards.  $U_t^*(v^t, \hat{v}^t)$  is perceived expected utility at time  $t$  with respect to  $F^*$ . Consumers utility is quasi-linear and time separable, with unit demand each period, so that  $U_T(v^T, \hat{v}^T) = \sum_{t=1}^T v_t q_t(\hat{v}^t) - P(\hat{v}^T)$ . For  $t \geq 1$ , let  $U_t(v^t, \hat{v}^t)$  and  $U_t^*(v^t, \hat{v}^t)$  be true and perceived expected utilities at time  $t$  of someone with type realizations and reports to date of  $(v^t, \hat{v}^t)$  who plans to report truthfully from period  $t+1$  onwards:

$$\begin{aligned} U_t(v^t, \hat{v}^t) &= E \left[ v_t q_t(\hat{v}^t) + \sum_{\tau=t+1}^T v_\tau q_\tau(\hat{v}^t, v_{t+1} + \dots + v_\tau) - P(\hat{v}^t, v_{t+1} + \dots + v_T) \mid (v^t, \hat{v}^t); F \right], \\ U_t^*(v^t, \hat{v}^t) &= E \left[ v_t q_t(\hat{v}^t) + \sum_{\tau=t+1}^T v_\tau q_\tau(\hat{v}^t, v_{t+1} + \dots + v_\tau) - P(\hat{v}^t, v_{t+1} + \dots + v_T) \mid (v^t, \hat{v}^t); F^* \right]. \end{aligned}$$

Also let  $U$  and  $U^*$  be true and perceived expected utilities at time zero of someone who plans to report all  $v^T$  truthfully:

$$\begin{aligned} U &= E \left[ \sum_{t=1}^T v_t q_t(v^t) - P(v^T) \mid F \right], \\ U^* &= E \left[ \sum_{t=1}^T v_t q_t(v^t) - P(v^T) \mid F^* \right]. \end{aligned}$$

Let  $G(U^*)$  be the fraction of consumers of type  $s$  with outside option below  $U^*$ . Let costs be  $C(q^T) = c \sum_{t=1}^T q_t$ , so that surplus is  $\sum_{t=1}^T (v_t - c) q_t$ . Define  $S_s$  to be the true expected surplus from a type  $s$  consumer who reports truthfully:  $S = E \left[ \sum_{t=1}^T (v_t - c) q_t(v^t) \mid F \right]$ .

Invoking the revelation principle, the monopolist's problem may then be written as:

$$\max_{\substack{q^T(v^T) \in [0,1] \\ P(v^T)}} G(U^*)(S - U)$$

such that

1. Truthful history IC  $t \geq 1$   $U_t^*(v^t, v^t) \geq U_t^*(v^t, [v^{t-1}, \hat{v}_t]) \quad \forall t, v^t$  and  $\hat{v}_t$
2. Any history IC  $U_t^*(v^t, v^t) \geq U_t^*(v^t, \hat{v}^t) \quad \forall t, v^t$  and  $\hat{v}^t$

**Lemma 6** *Monotonicity: A necessary condition for incentive compatibility is that  $q_t(v^t)$  be non-decreasing in  $v_t$ .*

**Proof.** Analogous to Lemma 2. ■

**Lemma 7** *Local IC: A necessary condition for incentive compatibility is  $\frac{d}{dv_t} U_t^*(v^t) = \frac{\partial}{\partial v_t} U_t^*(v^t) = q_t(v^t)$ .*

**Proof.** Analogous to Lemma 3. ■

I begin by solving a relaxed problem. I impose monotonicity ( $q_t(v^t)$  non-decreasing in  $v_t$ ) and local incentive compatibility for  $t \geq 1$  ( $\frac{d}{dv_t} U_t^*(v^t) = q_t(v^t)$ ). However I relax all other incentive constraints. In particular, I am only checking incentive compatibility against one step deviations, rather than multiple step deviations. After solving the relaxed problem, I will need to check (1) incentive compatibility against multiple step deviations and (2) for global incentive compatibility of  $v^T$  reporting to confirm that the relaxed solution solves the original problem.

By the envelope condition  $\frac{d}{dv_t} U_t^*(v^t) = \frac{\partial}{\partial v_t} U_t^*(v^t) = q_t(v^t)$  and the FTC,

$$U_t^*(v^t) = U_t^*(v^{t-1}, \underline{v}_t) + \int_{\underline{v}_t}^{v_t} q_t(v^{t-1}, x) dx.$$

Now by iterated expectations,

$$U_{T-1}^*(v^{T-1}) = \int_{\underline{v}_T}^{\bar{v}_T} U_T^*(v^T) f_s^*(v_T) dv_T.$$

Substituting in the envelope condition and integrating by parts gives

$$U_{T-1}^*(v^{T-1}) = U_T^*(v^{T-1}, \underline{v}_T) + \int_{\underline{v}_T}^{\bar{v}_T} q_T(v^T) (1 - F^*(v_T)) dv_T.$$

This can now be repeated recursively to yield

$$U^* = U_T(\underline{v}_1, \underline{v}_2, \dots, \underline{v}_T) + \sum_{t=1}^T \int_{\underline{v}_t}^{\bar{v}_t} q_t(v^t) (1 - F^*(v_t)) dv_t. \quad (29)$$

Similarly,

$$U = U_T(\underline{v}_1, \underline{v}_2, \dots, \underline{v}_T) + \sum_{t=1}^T \int_{\underline{v}_t}^{\bar{v}_t} q_t(v^t) (1 - F(v_t)) dv_t. \quad (30)$$

Or, alternatively,

$$U_T(\underline{v}_1, \underline{v}_2, \dots, \underline{v}_T) = U^* - \sum_{t=1}^T \int_{\underline{v}_t}^{\bar{v}_t} q_t(v^t) (1 - F^*(v_t)) dv_t,$$

and

$$U = U^* + \sum_{t=1}^T \int_{\underline{v}_t}^{\bar{v}_t} q_t(v^t) (F^*(v_t) - F(v_t)) dv_t.$$

Now we can simplify the doubly relaxed problem, by substituting the local incentive constraints summarized by equations (29-30) for  $U_T(\underline{v}_1, \underline{v}_2, \dots, \underline{v}_T)$  and  $U$  into the objective function, eliminating marginal fees from the problem. (Fixed fees depend on  $U^*$ ):

$$\max_{\substack{q^T \in [0,1] \\ U^*}} G(U^*) \left( S - U^* - \sum_{t=1}^T \int_{\underline{v}_t}^{\bar{v}_t} q_t(v^t) (F^*(v_t) - F(v_t)) dv_t \right)$$

such that

1. Monotonicity  $q_t(v^t)$  non-decreasing in  $v_t$ .

A term within the  $q_t(v^t)$  first order condition is:

$$\frac{dS}{dq_t(v^t)} = \int_{\underline{v}_t}^{\bar{v}_t} (v_t - c) f(v_t) dv_t.$$

The  $q_t(v^t)$  first order condition is:

$$\begin{aligned}
\frac{d\Pi}{dq_t(v^t)} &= G(U^*) \left( \frac{dS}{dq_t(v^t)} - \int_{\underline{v}_t}^{\bar{v}_t} (F^*(v_t) - F(v_t)) dv_t \right) \\
&= G(U^*) \left( \int_{\underline{v}_t}^{\bar{v}_t} (v_t - c) f(v_t) dv_t - \int_{\underline{v}_t}^{\bar{v}_t} (F^*(v_t) - F(v_t)) dv_t \right) \\
&= G(U^*) \int_{\underline{v}_t}^{\bar{v}_t} \left( v_t - c - \frac{F^*(v_t) - F(v_t)}{f(v_t)} \right) f(v_t) dv_t.
\end{aligned}$$

Define  $v^A$  such that:

$$v^A = c + \frac{F^*(v^A) - F(v^A)}{f(v^A)} \quad (31)$$

Maximizing point-wise implies that  $q_t(v^t) = 1$  if

$$\left( v_t - c - \frac{F^*(v_t) - F(v_t)}{f(v_t)} \right) > 0$$

(which is equivalent to  $v_t > v^A$  if  $v^A$  is unique) and zero otherwise. If  $v^A$  is not unique, the solution is obtained by ironing because monotonicity is necessary. Ironing will still yield a threshold such that the allocation rule is  $q_t(v^t) = 1_{v_t \geq \phi(v^{t-1})}$  for some  $\phi$ . Picking a threshold to maximize the integral will mean picking one of the fixed points of 31. (If weren't at a fixed point, could always improve by moving towards one or another without violating monotonicity.) Thus

$$v^A = \arg \max_{v^A} G(U^*) \left( -U^* + \sum_{t=1}^T \int_{v^A}^{\bar{v}} \left( v_t - c - \frac{F^*(v_t) - F(v_t)}{f(v_t)} \right) f(v_t) dv_t \right)$$

## A.9 Proof of Proposition 8

**The firm's problem:** Perceived and true expected utilities are:

$$U^* = v_0 - p_0 + 2 \int_{v^*}^1 v dF^*(v) - (p_1 + p_2)(1 - F^*(v^*)) - p_3(1 - F^*(v^*))^2, \quad (32)$$

and

$$U = v_0 - p_0 + 2 \int_{v^*}^1 v dF(v) - (p_1 + p_2)(1 - F(v^*)) - p_3(1 - F(v^*))^2, \quad (33)$$

respectively. Expected surplus is:

$$S = \int_{v^*}^1 (v - c) dF(v). \quad (34)$$

Substituting equations (17) and (18) into equation (33), yields true expected utility as a function of  $U^*$ ,  $v^*$ , and  $p_3$ :

$$U = U^* + 2 \int_{v^*}^1 \left( \frac{F^*(v) - F(v)}{f(v)} \right) f(v) dv - p_3 (F^*(v^*) - F(v^*))^2. \quad (35)$$

The firm's profit function in equation (19) is then obtained by substituting equations (34) and (35) into the expression  $\Pi = G(U^*)(S - U)$ .

**The proof:** (1) First note that given FOSD, the sign of cross partial derivative  $\partial^2 \Pi / \partial p_3 \partial v^*$  equals the sign of  $(f^*(v^*) - f(v^*))$ :

$$\frac{\partial^2 \Pi}{\partial p_3 \partial v^*} = 2 (F^*(v^*) - F(v^*)) (f^*(v^*) - f(v^*)).$$

Given  $F < F^*$  for all  $v \in (0, 1)$ , there is an interval  $[0, x)$  for which  $f^* > f$ . (In many natural cases  $f^*$  will cross  $f$  once from above at  $x$ ). Profits are strictly super modular in  $p_3$  and  $v^*$  ( $\partial^2 \Pi / \partial p_3 \partial v^* > 0$ ) over the interval  $(0, x)$ . Moreover,  $x$  is independent of  $\gamma$ . Let  $v^a$  be the solution  $v^a = c + \frac{F^*(v^a) - F(v^a)}{f(v^a)}$  with attentive consumers. The limit of  $v^a$  as  $\gamma$  approaches zero is  $c$ . Therefore, if  $c < x$  then for sufficiently small  $\gamma$ ,  $v^a < x$ . Given the constraint  $p_3 \leq p^{\max}$ , strict super modularity on  $(0, x)$  and  $v^a < x$  imply  $v^* > v^a$ . This follows (from Edlin and Shannon (1998)) because the change in the firm's maximization problem from attentive to inattentive customers is identical to the change when customers are inattentive but  $p_3$  exogenously increases from zero to  $p^{\max}$ . Note: given the constraint  $p_3 \leq h(v^*)$  for  $h(v^*)$  non-decreasing, the result continues to hold. If  $h(v^*)$  is strictly increasing, the constraint simply creates an additional incentive to raise  $v^*$  when consumers are inattentive: to relax the constraint on  $p_3$ .

(2) For  $c = 1$ ,  $v^a = 1$  and allocations are first best with attentive customers. However, fix any  $v^* \in (0, 1)$  and for  $p^{\max}$  sufficiently large,

$$2 \int_{v^*}^1 \left( v - 1 - \frac{F^*(v) - F(v)}{f(v)} \right) f(v) dv + p_3 (F^*(v^*) - F(v^*))^2 > 0,$$

which implies with inattentive customers there is overconsumption ( $v^* < 1$ ) and total welfare is strictly lower. By continuity, this is true for  $c$  in a neighborhood around  $c = 1$ .

## A.10 Proof of Proposition 9

**Step 1:** Show that the efficient allocation and zero penalty fees are optimal. To be completed.

**Step 2:** Given an efficient allocation, zero penalty fees  $p_3 = 0$ , and symmetric pricing  $p_1 =$

$p_2 = \bar{p}$ , perceived and actual payoffs are:

$$U^* = -p_0 + v_0 + 2\alpha'(1 - \bar{p})$$

$$U = -p_0 + v_0 + 2\alpha(1 - \bar{p})$$

$$\Pi = G(U^*) (S^{FB} - U^* - 2(\alpha - \alpha')(1 - p))$$

Note that

$$U - U^* = 2(\alpha - \alpha')(1 - \bar{p}) \geq 0$$

which implies that there is no exploitation:  $U^* \geq 0$ . Utility offer  $U^*$  is implemented by the following fixed fee:

$$p_0 = v_0 - U^* + 2\alpha'(1 - \bar{p}).$$

There are two cases to consider:

**(2a)**  $U^* \in [0, v_0]$ : Ignoring the NFL constraint, profits are increasing in  $\bar{p}$ . Thus the incentive constraint  $\bar{p} \leq 1$  will bind which implies  $\bar{p} = 1$ ,  $p_0 = v_0 - U^*$  and  $\mu(U^*) = S^{FB} - U^*$ . Given  $U^* \leq v_0$ , this satisfies NFL.

**(2b)**  $U^* \in (v_0, v_0 + 2\alpha']$ : Profits are increasing in  $\bar{p}$ , and the NFL constraint  $p_0 \geq 0$  will bind before the incentive constraint  $\bar{p} \leq 1$  since  $U^* > v_0$ . Thus NFL binds at  $p_0 = 0$  and  $p = 1 - (U^* - v_0)/2\alpha'$ . This implies a markup of  $\mu(U^*) = S^{FB} - U^* - (\alpha/\alpha' - 1)(U^* - v_0)$ .

## A.11 Proof of Lemma 1

Solving equation (22) for  $p_0$  yields:

$$p_0 = -U^* + v_0 + 2(1 - \alpha')b_0(-\bar{p}) + 2\alpha'b_1(1 - \bar{p}) - ((1 - \alpha')b_0 + \alpha'b_1)^2 p_3.$$

Substituting this for  $p_0$  into equation (23) gives:

$$\Pi = G(U^*) \left( \begin{aligned} & -U^* + v_0 + 2b_0(-(\alpha - \alpha')\bar{p} - (1 - \alpha)c) + 2b_1((\alpha - \alpha')\bar{p} + \alpha' - \alpha c) \\ & + \left( ((1 - \alpha)b_0 + \alpha b_1)^2 - ((1 - \alpha')b_0 + \alpha' b_1)^2 \right) p_3 \end{aligned} \right).$$

There are four alternatives to the efficient allocation to consider:

1.  $b_0 = b_1 = 1$ : Profits and the fixed fee are:

$$\Pi_1 = G(U^*) (-U^* + v_0 + 2(\alpha' - c)),$$

$$p_0 = -U^* + v_0 + 2\alpha' - 2\bar{p} - p_3.$$

If  $U^* \leq v_0 + 2\alpha'$ , then this allocation can be implemented without violating the NFL constraint with prices  $p_1 = p_2 = p_3 = 0$  and  $p_0 = -U^* + v_0 + 2\alpha'$ . If  $U^* > v_0 + 2\alpha'$ , then this allocation is not implementable without violating the NFL constraint. This follows from the fact that  $p_0 + 2\bar{p} + p_3 \geq 0$  is equivalent to  $U^* \leq v_0 + 2\alpha'$ . However, the efficient allocation could be implemented with identical prices, also satisfying the NFL constraint for  $U^* \leq v_0 + 2\alpha'$ , but yielding strictly higher profit,

$$\Pi = G(U^*) (-U^* + v_0 + 2(\alpha' - \alpha c)),$$

by saving production cost  $2(1 - \alpha)c$ . Thus  $b_0 = b_1 = 1$  is never optimal.

2.  $b_0 = b_1 = 0$ : Profits and the fixed fee are:

$$\Pi_2 = G(U^*) (-U^* + v_0),$$

$$p_0 = -U^* + v_0.$$

If  $U^* \leq v_0$  then this allocation is implementable without violating the NFL with prices  $p_0 = -U^* + v_0$ ,  $p_1 = p_2 = 1$ , and  $p_3 = 0$ . If  $U^* > v_0$ , then this allocation is not implementable without violating the NFL constraint. However, the efficient allocation can be implemented with identical prices, strictly raising profits by  $2\alpha(1 - c)$  from the additional sales. Thus  $b_0 = b_1 = 0$  is never optimal.

3.  $b_0 \in (0, 1)$ ,  $b_1 = 1$ : For this allocation to be implemented,  $b_0$  must satisfy first and second order conditions of the consumers' problem:

$$\frac{dU^*}{db_0} = -2(1 - \alpha')(\bar{p} + ((1 - \alpha')b_0 + \alpha')p_3) = 0,$$

and

$$\frac{d^2U^*}{db_0^2} = -2(1 - \alpha')^2 p_3 \leq 0.$$

This requires that  $p_3 \geq 0$  and  $\bar{p} = -((1 - \alpha')b_0 + \alpha')p_3$ . At these prices, the three NFL constraints, (a)  $p_0 \geq 0$ , (b)  $p_0 + \bar{p} \geq 0$ , and (c)  $p_0 + 2\bar{p} + p_3 \geq 0$  are:

$$\max \left\{ \frac{U^* - v_0 - 2\alpha'}{((1 - \alpha')b_0 + \alpha')^2}, \frac{U^* - v_0 - 2\alpha'}{(1 - \alpha')^2(1 - b_0)^2} \right\} \leq p_3 \leq \frac{2\alpha' + v_0 - U^*}{(1 - \alpha')(1 - b_0)((1 - \alpha')b_0 + \alpha')}$$

If  $p_3 \geq 0$ , the upper bound on penalty fees can only be satisfied if  $U^* \leq v_0 + 2\alpha'$ , in which case the lower bound is always satisfied. Moreover, profits are increasing in penalty fee  $p_3$ ,

$$\Pi_3 = G(U^*) \left( v_0 - U^* + 2(\alpha' - \alpha)c - 2b_0(1 - \alpha)c + (\alpha - \alpha')^2(1 - b_0)^2 p_3 \right),$$

so the optimal penalty fee satisfies the upper bound with equality:

$$p_3 = \frac{(2\alpha' + v_0 - U^*)}{(1 - \alpha')(1 - b_0)((1 - \alpha')b_0 + \alpha')}.$$

Given these prices, profits are strictly decreasing in  $b_0$ ,

$$\frac{d\Pi_3}{db_0} = G(U^*) \left( -2(1 - \alpha)c - \frac{(\alpha - \alpha')^2(1 - b_0)}{((1 - \alpha')b_0 + \alpha')} p_3 \right) < 0,$$

for all  $p_3 \geq 0$  and hence any NFL implementable allocation with  $b_0 \in (0, 1)$  is always dominated by the efficient allocation.

4.  $b_0 = 0$ ,  $b_1 \in (0, 1)$ : For this allocation to be implemented,  $b_1$  must satisfy first and second order conditions of the consumers' problem:

$$\frac{dU^*}{db_1} = +2\alpha'(1 - \bar{p}) - 2(\alpha')^2 b_1 p_3 = 0,$$

and

$$\frac{d^2U^*}{db_1^2} = -2(\alpha')^2 p_3 \leq 0.$$

This requires  $p_3 \geq 0$  and  $\bar{p} = 1 - \alpha' b_1 p_3$ . At these prices, the three NFL constraints, (a)  $p_0 \geq 0$ , (b)  $p_0 + \bar{p} \geq 0$ , and (c)  $p_0 + 2\bar{p} + p_3 \geq 0$  are:

$$\max \left\{ \frac{U^* - v_0}{\alpha'^2 b_1^2}, \frac{U^* - v_0 - 2}{(1 - \alpha' b_1)^2} \right\} \leq p_3 \leq \frac{1 + v_0 - U^*}{\alpha' b_1 (1 - \alpha' b_1)}$$

All three constraints can be satisfied only if  $U^* \leq v_0 + \alpha' b_1$ . (This is equivalent to  $\frac{U^* - v_0}{\alpha'^2 b_1^2} \leq \frac{1 + v_0 - U^*}{\alpha' b_1 (1 - \alpha' b_1)}$ , while  $\frac{U^* - v_0 - 2}{(1 - \alpha' b_1)^2} \leq \frac{1 + v_0 - U^*}{\alpha' b_1 (1 - \alpha' b_1)}$  is equivalent to the weaker condition  $U^* \leq 1 + v_0 + \alpha' b_1$ .) Otherwise, this allocation is not implementable without violating NFL. Profits are strictly increasing in  $p_3$ ,

$$\Pi_4 = G(U^*) \left( -U^* + v_0 + 2b_1\alpha(1 - c) + b_1^2(\alpha - \alpha')^2 p_3 \right),$$

so the optimal penalty fee  $p_3$  will equal the upper bound:

$$p_3 = \frac{1 + v_0 - U^*}{\alpha' b_1 (1 - \alpha' b_1)}.$$

Given these prices, profits are strictly increasing in  $b_1$ ,

$$\frac{d\Pi_4}{db_1} = G(U^*) \left( 2\alpha(1-c) + \frac{b_1(\alpha - \alpha')^2}{1 - \alpha' b_1} p_3 \right) > 0,$$

so any NFL implementable allocation with  $b_1 \in (0, 1)$  is dominated by the efficient allocation.

## A.12 Proof Proposition 10

NFL says prices can be no lower than  $p_0 = p_1 = p_2 = p_3 = 0$ , and hence offered perceived utility  $U^*$  can be no higher than  $v_0 + 2\alpha'$ . Optimal pricing need only be characterized for  $U^* \in [0, v_0 + 2\alpha']$ . By Lemma 1, the firm will induce the efficient allocation,  $b_0 = 0$ ,  $b_1 = 1$ . In this case, profits and fixed fees are:

$$\Pi = G(U^*) \left( -U^* + v_0 + 2((\alpha - \alpha')\bar{p} + \alpha' - \alpha c) + (\alpha^2 - \alpha'^2) p_3 \right),$$

$$p_0 = -U^* + v_0 + 2\alpha'(1 - \bar{p}) - \alpha'^2 p_3.$$

Incentive compatibility requires that the expected marginal price be between zero and one:  $0 \leq \bar{p} + \alpha' p_3 \leq 1$ , or alternatively that the penalty fee be between:  $-\bar{p}/\alpha' \leq p_3 \leq (1 - \bar{p})/\alpha'$ . The three NFL constraints, (a)  $p_0 \geq 0$ , (b)  $p_0 + \bar{p} \geq 0$ , and (c)  $p_0 + 2\bar{p} + p_3 \geq 0$  are:

$$\frac{U^* - v_0 - 2\alpha'(1 - \bar{p}) - 2\bar{p}}{1 - \alpha'^2} \leq p_3 \leq \frac{2\alpha'(1 - \bar{p}) + v_0 - U^*}{\alpha'^2} + \min \left\{ 0, \frac{\bar{p}}{\alpha'^2} \right\}.$$

There are two cases to consider.

Case I,  $U^* < \alpha' + v_0$ : Impose the NFL upper bound  $p_3 \leq (2\alpha'(1 - \bar{p}) + \bar{p} + v_0 - U^*)/\alpha'^2$  and the IC upper bound  $p_3 \leq (1 - \bar{p})/\alpha'$ , but relax the other three constraints. At  $\bar{p} = -\frac{\alpha' - (U^* - v_0)}{1 - \alpha'}$ , both constraints are the same and the optimal penalty fee would be the upper bound  $p_3 = \frac{1 - (U^* - v_0)}{(1 - \alpha')\alpha'}$ . For larger  $\bar{p}$ , the IC upper bound is tighter and the optimal penalty is  $p_3 = (1 - \bar{p})/\alpha'$ . In this case, profits are,

$$\Pi = G(U^*) \left( -U^* + v_0 + 2((\alpha - \alpha')\bar{p} + \alpha' - \alpha c) + (\alpha^2 - (\alpha')^2) (1 - \bar{p})/\alpha' \right),$$

and  $d\Pi/d\bar{p} = -G(U^*) (\alpha - \alpha')^2/\alpha' < 0$ , so it is optimal to reduce  $\bar{p}$  towards  $\bar{p} = -\frac{\alpha' - (U^* - v_0)}{1 - \alpha'}$ . For

$\bar{p}$  below  $-\frac{\alpha' - (U^* - v_0)}{1 - \alpha'}$ , the NFL upper bound is binding, and as shown under case 1, it is optimal to increase  $\bar{p}$ . Thus the optimal prices are:

$$\bar{p} = -p_0 = -\frac{v_0 + \alpha' - U^*}{1 - \alpha'}, p_3 = \frac{v_0 + 1 - U^*}{(1 - \alpha')\alpha'}.$$

The assumption  $U^* < v_0 + \alpha'$  ensures  $\bar{p}$  is negative, and hence the alternative NFL upper bound is satisfied. Substituting for prices, the NFL lower bound reduces to  $U^* \leq v_0 + \alpha' + 1$ , which is satisfied given  $U^* < v_0 + \alpha'$ . The IC lower bound is always satisfied when the upper bound is satisfied with equality. Substituting for prices, profits are

$$\Pi = G(U^*) \left( S^{FB} - U^* + \frac{(\alpha - \alpha')^2}{\alpha'(1 - \alpha')} (1 + v_0 - U^*) \right).$$

Case II,  $U^* \in [v_0 + \alpha', v_0 + 2\alpha']$ : Relax the incentive constraint and the NFL lower bound on the penalty fee. Since profits are increasing in both  $\bar{p}$  and  $p_3$ , for any fixed  $\bar{p}$ , the penalty fee  $p_3$  will be set at the NFL upper bound. If  $\bar{p} \geq 0$ , this implies  $p_3 = (2\alpha'(1 - \bar{p}) + v_0 - U^*)/\alpha'^2$ ,

$$\Pi = G(U^*) \left( 2\alpha(\bar{p} - c) + 2\alpha^2(1 - \bar{p})/\alpha' - (\alpha/\alpha')^2(U^* - v_0) \right),$$

and  $d\Pi/d\bar{p} = -2\alpha(\alpha - \alpha')/\alpha' < 0$ . Thus profits increases as  $\bar{p}$  is reduced towards zero. If  $\bar{p} \leq 0$ , this implies  $p_3 = (2\alpha'(1 - \bar{p}) + \bar{p} + v_0 - U^*)/\alpha'^2$ ,

$$\Pi = G(U^*) \left( 2\alpha(\bar{p} - c) - \bar{p} + (2\alpha'(1 - \bar{p}) + \bar{p} - (U^* - v_0))(\alpha/\alpha')^2 \right),$$

and  $d\Pi/d\bar{p} = (\alpha^2(1 - 2\alpha') - \alpha'^2(1 - 2\alpha))/\alpha'^2 > 0$ . Thus profits increase as  $\bar{p}$  is increased towards zero. As a result, optimal prices are  $\bar{p} = p_0 = 0$  and  $p_3 = (2\alpha' + v_0 - U^*)/\alpha'^2$ . Substituting for prices, the IC constraint is equivalent to the assumption  $U^* \in [v_0 + \alpha', v_0 + 2\alpha']$  and hence is satisfied. Similarly, the NFL lower bound is equivalent to  $U^* \leq v_0 + 2\alpha'$  and so is satisfied. Substituting for prices, profits are

$$\begin{aligned} \Pi &= G(U^*) \left( (2\alpha' + v_0 - U^*) (\alpha/\alpha')^2 - 2\alpha c \right) \\ &= G(U^*) \left( S^{FB} - U^* + \left( (\alpha/\alpha')^2 - 1 \right) (v_0 - U^*) + 2(\alpha - \alpha')\alpha/\alpha' \right) \end{aligned}$$

### A.13 Proof of Corollary 6

For sufficiently small transportation cost  $\tau$ , there will be full market coverage in equilibrium, with each firm receiving positive market share. In this case, if firms A and B offer perceived expected

utilities of  $U^A$  and  $U^B$  respectively, market share of firm A is:  $G(U^A, U^B) = \frac{1}{2\tau} (U^A - U^B + \tau) \geq 0$ . Profits are

$$\Pi^A = G(U^A, U^B) \mu(U^A)$$

where  $\mu(U^A)$  is the markup derived in Proposition 9 in the attentive case, or Proposition 10 in the inattentive case. In particular, in the attentive case,  $\mu(U^A)$  is given by equation (20) for  $U^A \in [0, v_0]$  and by equation (21) for  $U^A \in (v_0, v_0 + 2\alpha']$ . In the inattentive case,  $\mu(U^A)$  is given by equation (24) for  $U^A \in [0, v_0 + \alpha']$  and by equation (25) for  $U^A \in (v_0 + \alpha', v_0 + 2\alpha']$ . In both attentive and inattentive cases, the profit function is concave (with a kink at  $U^A = v_0$  in the attentive case, and with a kink at  $U^A = v_0 + \alpha'$  in the inattentive case), and hence firm A's best response is a continuous function of  $U^B$ . Away from the kink  $d^2\Pi^A/dU^A{}^2 = g(U^A, U^B) d\mu/dU^A < 0$ , and at the kink  $d\Pi^A/dU^A$  decreases. This follows since

$$\frac{d\Pi^A}{dU^A} = g(U^A, U^B) \mu(U^A) + G(U^A, U^B) \frac{d\mu}{dU^A},$$

and while  $G(U^A, U^B)$  is continuous and nonnegative, in the attentive case  $d\mu/dU^A$  decreases at  $U^A = v_0$  (Since  $d\mu/dU^A = -1$  for  $U^A < v_0$ ,  $d\mu/dU^A = -(\alpha/\alpha')$  for  $U^A > v_0$  and  $(\alpha/\alpha') > 1$ ), and in the inattentive case  $d\mu/dU^A$  decreases at  $U^A = v_0 + \alpha'$  (Since  $d\mu/dU^A = -(1+Y)$  for  $U^A < v_0 + \alpha'$ ,  $d\mu/dU^A = -(\alpha/\alpha')^2$  for  $U^A > v_0 + \alpha'$  and  $(\alpha/\alpha')^2 > 1+Y$ ).

The optimal  $U^A$  either solves the first order condition  $\mu(U^A) = -(U^A - U^B + \tau) d\mu/dU^A$ , or is located at the kink ( $v_0$  in the attentive case or  $v_0 + \alpha'$  in the inattentive case). In the attentive case, there are three sub-cases: (1)  $U^A < v_0$ : The first order condition is  $U^A = \frac{1}{2} (S^{FB} + U^B - \tau)$ . (2)  $U^A > v_0$ , the first order condition is

$$U^A = \frac{1}{2} (U^B - \tau + v_0) + \frac{1}{2} (\alpha'/\alpha) (S^{FB} - v_0).$$

(3)  $U^A = v_0$ .

In the inattentive case, there are also three sub-cases: (1)  $U^A < v_0 + \alpha'$ : The first order condition is

$$U^A = \frac{1}{2} (U^B + v_0 - \tau) + \frac{2\alpha(1-c) + Y}{2(1+Y)}$$

(2)  $U^A > v_0 + \alpha'$ : The first order condition is

$$U^A = \frac{1}{2} (U^B + v_0 - \tau) + \alpha' - \alpha c (\alpha'/\alpha)^2$$

(3)  $U^A = v_0 + \alpha'$ . By inspection, the best response by A has slope  $dU^A/dU^B$  of either zero or 1/2.

Since  $dU^A/dU^B \in [0, 1)$ , there is a unique pure strategy equilibrium, which is symmetric. This is true for both attentive and inattentive cases.

Attentive case: For an equilibrium with  $U^* > v_0$  and full market coverage,  $U^*$  solves  $U^* = \frac{1}{2}(U^* - \tau + v_0) + \frac{1}{2}(\alpha'/\alpha)(S^{FB} - v_0)$ , which yields:

$$U^* = 2\alpha'(1 - c) + v_0 - \tau.$$

The condition  $U^* > v_0$  is equivalent to  $\tau < 2\alpha'(1 - c)$ , and  $U^* > \tau/2$  (full market coverage) is equivalent to  $\tau < \frac{4}{3}\alpha'(1 - c) + \frac{2}{3}v_0$ . The joint assumption  $\tau < \frac{4}{3}\alpha'(1 - c)$  and  $v_0 \geq 0$  is sufficient for both  $U^* > v_0$  and full market coverage. By Proposition 9, the markup is  $\mu = \tau(\alpha/\alpha')$  and prices are  $p_3 = p_0 = 0$  and  $p_1 = p_2 = c + \tau/2\alpha'$ .

Inattentive case: For an equilibrium with  $U^* > v_0 + \alpha'$  and full market coverage,  $U^*$  solves  $U^* = \frac{1}{2}(U^* + v_0 - \tau) + \alpha' - \alpha c(\alpha'/\alpha)^2$ , which yields:

$$U^* = (v_0 - \tau) + 2\alpha'(1 - c(\alpha'/\alpha)).$$

The assumption  $\tau < \alpha'(1 - 2c(\alpha'/\alpha))$  is necessary and sufficient for the solution to satisfy  $U^* > v_0 + \alpha'$ . Moreover, it is sufficient for full market coverage ( $U^* > \tau/2$ ) since given  $v_0 \geq 0$

$$\tau \leq \alpha'(1 - 2c(\alpha'/\alpha)) < \alpha'(1 - c(\alpha'/\alpha)) < \frac{4}{3}\alpha'(1 - c(\alpha'/\alpha)) \leq \frac{4}{3}\alpha'(1 - c(\alpha'/\alpha)) + \frac{2}{3}v_0,$$

and  $\tau < \frac{4}{3}\alpha'(1 - c(\alpha'/\alpha)) + \frac{2}{3}v_0$  is equivalent to  $U^* > \tau/2$ . By Proposition 10, the markup is  $\mu = \tau(\alpha/\alpha')^2$  and prices are  $\bar{p} = p_0 = 0$  and  $p_3 = 2c/\alpha + \tau/\alpha'^2$ .

The assumption  $\tau \leq \alpha'(1 - 2c)$  is sufficient for both  $\tau < \frac{4}{3}\alpha'(1 - c)$  and  $\tau < \alpha'(1 - 2c(\alpha'/\alpha))$ .