Cambridge Working Papers in Economics CWPE 0430

UNIVERSITY OF
CAMBRIDGE
Department of Applied Economics

Retail Electricity Competition

Paul Joskow and Jean Tirole

The Cambridge-MIT Institute

Massachusetts Institute of Technology
Center for Energy and
Environmental Policy Research

CMI Working Paper 44

Cambridge Working Papers in Economics

UNIVERSITY OF
CAMBRIDGE
Department of
Applied Economics

Not to be quoted without permission

Massachusetts Institute of Technology
Center for Energy and
Environmental Policy Research

CMI Working Paper

Retail Electricity Competition*

Paul Joskow† and Jean Tirole‡

April 21, 2004

Abstract

We analyze a number of unstudied aspects of retail electricity competition. We first explore the implications of load profiling of consumers whose traditional meters do not allow for measurement of their real time consumption, when consumers are homogeneous up to a scaling factor. In general, the combination of retail competition and load profiling does not yield the second best prices given the non price responsiveness of consumers. Specifically, the competitive equilibrium does not support the Ramsey two-part tariff. By contrast, when consumers have real time meters and are billed based on real time prices and consumption, retail competition yields the Ramsey prices even when consumers can only partially respond to variations in real time prices. More complex consumer heterogeneity does not lead to adverse se1ection and competitive screening behavior unless consumers have real time meters and are not rational. We then examine the incentives competitive retailers have to install one of two types of advanced metering equipment. Competing retailers overinvest in real time meters compared to the Ramsey optimum, but the investment incentives are constrained optimal given load-profiling and retail competition. Finally, we consider the effects of physical limitations on the ability of system operators to cut off individual customers. Competing retailers have no incentive to determine the aggregate value of non-interruption of consumers in the zones they serve, preferring instead to free ride on other retailers serving consumers in the same zones.

*We are grateful to Claude Crampes, Bruno Jullien, Stephen Holland, Patrick Rey and the participants at the IDEI-CEPR conference on "Competition and Coordination in the Electricity Industry," January 16–17, 2004, Toulouse and at the ninth annual POWER conference, UC Berkeley, March 19, 2004 for helpful discussions and comments.

†Department of Economics, and Center for Energy and Environmental Policy Research, MIT.

‡IDEI and GREMAQ (UMR 5604 CNRS), Toulouse, CERAS (URA 2036 CNRS), Paris, and MIT.

1 Introduction

The paper analyzes a number of hitherto unstudied aspects of retail competition in electricity markets. Its starting point is that final consumers may not react to the real time prices that emerge in wholesale electricity markets for (at least) three reasons: First, they do not have incentives to properly adjust their consumption to real-time prices if only their total consumption over a given period is recorded, i.e., they are on a traditional meter. Second, even if their consumption is recorded on a real-time basis, transaction costs associated with monitoring the evolution of hourly prices and constantly optimizing the use of equipment are enormous for small consumers. Third, consumers, even if they want to, may not be able to adjust their consumption freely. They may be constrained by the physical attributes of distribution networks as they are presently configured; in particular, rationing usually occurs at the level of zones rather than individual consumers.

In order to analyze competition among electricity retailers or Load Serving Entitites (LSEs) for the final (retail) consumers, it is convenient to group the latter into four categories:[1]

Price-sensitive consumers are endowed with real-time (RT) meters and either autonomously or through communication with the LSE, adjust their demand efficiently to the evolution of the wholesale spot market price.

Price-insensitive consumers with real-time meters are endowed with RT meters, but are only partially aware or unaware of RT prices and therefore do not adjust their consumption perfectly as real time prices vary from

[1]The grouping in three categories is an oversimplification. There are a number of partially price sensitive categories, such as those subject to time-of-use pricing (retail prices are preset for certain blocks of time) or critical peak pricing (that combines time-of-use pricing with high retail prices for a number of critical hours per year to be declared by the utility). See Borenstein et al (2002) for a review of recent innovations.

minute to minute and hour to hour. At the extreme, they are fully (RT) price-insensitive. Such consumers are not irrational; rather they trade off the transaction costs invoked above and the savings in their electricity bill. While these consumers do not react to real time prices their actual real time consumption can be measured and assigned to their LSE for settlement purposes.

Consumers on traditional meters are metered only once a month or every few months (in some countries meters are read even less frequently), and pay a per-kWh electricity charge that is independent of the actual timing of their overall consumption. The case of consumers on traditional meters can be decomposed into two subcases, depending on the way the consumers' LSE is charged for its energy purchases. In the case of a *monopoly* local distribution company, this company pays the real-time price of the consumer's consumption: Even though the LSE is then unable to measure the realized profile of any given consumer with a traditional meter in its distribution area, it observes and pays for the realized total consumption profile of all such customers in the area.

Under *retail competition* by contrast, an LSE other than the local distribution grid owner and serving such a consumer pays a unit electricity charge based on the "load profile" of the consumer. That is, it pays *the average wholesale price for the load profile* that is representative of the consumer's class regardless of the actual time pattern of the individual customer's consumption and the relationship between this actual physical consumption and the contemporaneous RT wholesale prices.

Table 1 summarizes this taxonomy.

	Meter measures:	**Consumer RTP sensitive?**	**LSE's energy purchase cost corresponds to:**
1	entire consumption profile (RT meter)	yes	customer's RT profile
2	entire consumption profile (RT meter)	no / partial	customer's RT profile
3	aggregate consumption only (traditional meter)	no	customer's RT profile
4	aggregate consumption only (traditional meter)	no	load profiled consumption

Table 1

The case of price-sensitive consumers who react efficiently to real time prices (case 1 in Table 1) is the textbook representation of consumer demand. Borenstein and Holland (2003,a,b) study retail competition when price-sensitive consumers (case 1) and price-insensitive consumers endowed with real-time meters (case 2) co-exist, and LSEs can use only linear prices. This paper extends their analysis of case 2 (Section 3) and treats cases 3 and 4 in which consumers are endowed with traditional meters (Section 2).

The paper focuses on two possible failures of retail price signals to adequately reflect the scarcity conveyed by real time wholesale market price signals. The first failure arises at the consumer level when only her aggregate consumption is measured. Because the consumer then does not pay more when consuming mainly at peak when wholesale prices are high than when spreading consumption more equally across peak and off-peak hours, the consumer consumes too much at peak and too little off peak. The second failure occurs at the retailer's level, when the latter's individual consumers'

real time intake again is not measured by the system, which then charges the retailer on the basis of some estimated consumption load profile rather than the LSE's consumers' actual load profile.

Section 2 analyzes retail competition among load serving entities (LSEs) in a world in which consumers are homogeneous (possibly up to a scale parameter) and on traditional meters. Section 2.1 characterizes the second-best optimum and shows that it can be implemented in the absence of retail competition. By contrast, Section 2.2 and 2.3 show that under load profiling, retail competition (with or without the incumbent distributor) leads to a retail price equal to the average wholesale power cost and differing from the socially optimal retail price.

Section 3 takes on the case (case 2) in which consumers are on real-time meters, but do not react or only partially react to the real-time prices. We here build on the analysis of Borenstein and Holland (2003a,b) and expand on it in a number of ways: (i) We argue that partial responsiveness requires considering alternative representations of consumer demand; (ii) we allow competing LSEs to offer non-linear prices to retail consumers. Unlike Borenstein and Holland, we find that with homogeneous and rational consumers retail competition leads to the second-best optimum; this is no longer true with boundedly rational consumers.

Section 4 extends the analysis to situations in which consumers differ in other aspects than just scale, i.e., they have different load profiles, and investigates the possibility of adverse selection and competitive screening.

Section 5 shows that given the price inefficiencies associated with load profiling, LSEs face the right incentives when offering their customers enhanced metering equipment, and so subsidies for such equipments are not

warranted.[2]

Last, Section 6 analyzes the implications of limitations in the controllability of the distribution circuits. These limitations imply that price sensitive consumers may be rationed along with everyone else, and that LSEs cannot generally demand any specific level of rationing that they desire to reflect their consumers' valuations. At best one can then elicit only the aggregate willingness to pay for reliability in any given joint interruptibility zone. The section discusses both market mechanisms that are needed to reach a "third best" and the difficulties that make the phasing out of non-market mechanisms unlikely when there is retail competition.

2 Consumers on traditional meters

2.1 Model and social optimum

States of nature (or, equivalently, periods) are indexed by $i \in [0,1]$. f_i denotes the frequency of state i. Because we focus on competition on the demand side, we take the wholesale prices as exogenous, and we identify states of nature by the wholesale price p_i, with p_i increasing in i. [We will discuss this identification later on.]

For the sake of simplicity, let us ignore rationing for the moment. We consider a representative retail consumer on a traditional meter with demand $D_i(p)$ when facing price p in state i, with $D_i' < 0$. Let $S_i\left(D_i(p)\right)$ denote the associated gross surplus, with $S_i' = p$. Note that consumers are assumed to be homogeneous. They may differ in the size of their demand, though: That is, they can be indexed by $\sigma > 0$, such that a consumer of type σ has demand $q_i = \sigma D_i(p)$ and gross surplus $\sigma S_i\left(q_i/\sigma\right)$. We normalize σ to be equal to 1,

[2] We do not consider metering technologies where there are economies of scale or density.

but nothing is changed if consumers differ in their scale σ.[3] [More general forms of heterogeneity are discussed in Section 4.]

Assumption 1. The function $E\left[(p - p_i)\, D_i'(p)\right]$ is decreasing in p.[4]

The retail consumers are physically served by a local grid owner (usually also called the incumbent distributor, or transmission and distribution service provider). Because we are not interested here in the price of access to the grid, we normalize to zero any delivery, metering and customer service costs that continue to reflect responsibilities of the distribution grid owner. Thus, when we later introduce LSEs, their only cost will be either the purchase of energy from the wholesale market in real time or, in the case of load-profiled consumers, the load profiled variable charge for power supplied from the wholesale market to be paid for power delivered by the local grid owner. We consider two-part tariffs, consisting of a monthly subscriber charge and a per-kWh variable charge.[5] We will later note that focusing on two-part tariffs involves no loss of generality.

A consumer on a traditional meter cannot obtain the first-best utility, U^{FB}, that she would obtain if her demand were controlled to perfectly adjust to the RTP:

$$U^{FB} \equiv E\left[S_i\left(D_i\left(p_i\right)\right) - p_i D_i\left(p_i\right)\right]. \tag{1}$$

A Ramsey social planner for consumers with traditional meters chooses prices, namely single per unit retail price p^* and fixed fee A^*, so as to maximize the consumer's expected net surplus subject to the budget balance

[3]Neither the social planner nor the LSEs need to observe the consumer's scale σ in advance: They can infer it ex post from the consumer's total consumption.

[4]This assumption is made mainly for analytical convenience. It is satisfied in particular if the demand functions' curvature is small enough ($|D_i''/D_i'|$ small).

[5]Offers by retailers to residential customers in England and Texas that we have reviewed have a fixed monthly charge plus one or more tiers of kWh charges.

constraint:

$$U^* \equiv \max_{\{p^*, A^*\}} E\left[S_i\left(D_i\left(p^*\right)\right) - p^* D_i\left(p^*\right)\right] - A^*$$

s.t.

$$E\left[\left(p^* - p_i\right) D_i\left(p^*\right)\right] + A^* \geq 0. \tag{2}$$

At the optimum, the budget constraint is binding, and the Ramsey planner maximizes the joint surplus:

$$U^* = \max_{p^*} E\left[S_i\left(D_i\left(p^*\right)\right) - p_i D_i\left(p^*\right)\right], \tag{3}$$

yielding the following formula:

$$E\left[\left(p^* - p_i\right) D_i'\left(p^*\right)\right] = 0. \tag{4}$$

Assumption A1 implies that (4) has a unique solution.

To get some feel for what the Ramsey price entails, suppose for example that the elasticity of demand comes from the installation of air conditioning units. Suppose further that there are only two periods: off-peak (1) and peak (2), with respective wholesale prices p_1 and p_2. Then, the Ramsey price is $p^* = p_2$ in the US and $p^* = p_1$ in France, since summer is part of the peak in the US and is off peak in France.[6] Thus the Ramsey price would be greater than the average annual wholesale price of electricity in the US and below the average annual wholesale price in France.

Let us now consider the case of an LSE whose energy purchase cost corresponds to its customers' actual load profile (case 3 in Table 1). As we have argued, this is the case when customers in an area are served by a monopoly distribution company. The LSE then chooses the two-part tariff (p, A) so as

[6] If "installation" referred to electric heating, then $p^* = p_2$ in France since winter is the peak period.

to maximize its profit

$$E\left[\left(p-p_i\right)D_i(p)\right]+A$$

subject to the consumers being granted a certain utility level $\overline{U}$ (0 if the monopoly is unregulated, higher if regulated):

$$E\left[S_i\left(D_i(p)\right)-pD_i(p)\right]-A\geq\overline{U}.$$

This program is of course the dual of the Ramsey program above. We thus obtain:

Proposition 1 *Traditional meters give rise to consumer moral hazard: Consumers consume relatively too much on peak and too little off peak.*

(i) The Ramsey usage price is given by

$$E\left[\left(p^*-p_i\right)D_i'\left(p^*\right)\right]=0.$$

(ii) The Ramsey (second-best) allocation prevails in the absence of retail competition.

Remark (optimality of two-part tariffs): We have assumed that the Ramsey planner offers two-part tariffs. Could a better allocation be obtained through more complex pricing structures?

With traditional meters, the social planner (or an LSE for that matter) cannot do better than with a two-part tariff. At best he can hope to control total consumption through the marginal charge, while the load curve is chosen by the consumer without any concern for the actual cost of purchasing energy. More formally, the social planner is limited to total-consumption based tariffs $T(Q)$. Suppose that the planner selects the consumer's total consumption Q, and charges an amount T for this. The consumer then chooses her load curve so as to solve:

$$\max E\left[S_i\left(D_i\right)\right]\quad\text{subject to}\quad E\left[D_i\right]=Q.$$

Letting p denote the shadow price of the constraint, $S_i'(D_i) = p$, and so the allocation is the same as under a two-part tariff.

2.2 Retail competition for load-profiled consumers: independent retailers

We analyze the competitive outcome with load profiled customers in two environments. In the first, the local grid owner is subject to a line-of-business restriction. He provides access or delivery service to retailers, but is not allowed to compete for the final consumer. In the second, this line-of-business restriction is lifted and so the incumbent distributor is permitted to compete with independent retailers. We assume either that the distributor separates its retail "supply" business into a ring-fenced affiliate that is treated like any other retailer (as in the UK and in Texas), or that the retail arm maximizes the profit of the vertically integrated firm.[7]

In this subsection, we assume that (pure) retailers, but not the local grid owner, compete for load-profiled consumers and can offer two-part tariffs if they choose to do so.[8] Retailers' settlement obligations for wholesale power costs are then based on their customers' load-profiled consumption.[9] To com-

[7]A further complication is that when retail competition is first introduced the distributor as retailer initially cannot "compete" in the normal sense, but rather is required to offer default service at a regulated price. These default service prices have been set in many different ways. We view these regulated default service obligations as transition arrangements and focus our analysis on a post transition retail competition regime where there is no regulated default service requirement.

[8]We are interested solely in the price effects of retail competition. We thereby ignore some benefits of competition(such as improved incentives to offer better metering, tariffs, total energy management services or hedging packages) as well as some potential costs of retail competition (such as consumer churn and poaching, duplicative or misleading advertising expenditures, and competitive screening for credit quality and high volume consumers).

[9]The aggregate demand of all consumers served through a particular distribution network is measured on a real time basis. Since the aggregate real time consumption obligations must add up to the aggregate real time supplies of power delivered over the distribution network, a set of "load profiles" must be applied to the monthly, bi-monthly or quarterly consumption measured for customers without real-time meters. For example, consider a customer with a standard meter read on a monthly basis with 1000 kWh of

pute the price per kWh paid for wholesale energy by retailers for each customer they have signed up, a, suppose that, in equilibrium, retailers' variable (per-kWh) charge to consumers is p. Average consumption (load profiled) per consumer is $E[D_i(p)]$ and the wholesale price paid by the retailers for energy is

$$a(p) = \frac{E[p_i D_i(p)]}{E[D_i(p)]}. \tag{5}$$

We use the notation a for "access charge" by analogy with the economics literature on variable charges paid by entrants for access to regulated bottlenecks (local loop, etc.).[10] This access charge must be understood as the average wholesale power cost paid by retailers.

Let us define the "average wholesale cost price", $\hat{p}$, as the marginal retail price that balances an LSE's budget in the absence of a fixed charge:

$$E[(\hat{p} - p_i) D_i(\hat{p})] = 0.$$

Intuitively, $\hat{p}$ exceeds the Ramsey price p^* if the state of nature impacts

consumption recorder for the previous month. The 1000 kWh of monthly consumption then must be allocated to the 720 hours of the previous month for settlement purposes. This is accomplished by assigning the customer to a group or class of customers thought to have similar consumption. A consumption or load profile is developed for each group based on real-time metered consumption patterns of a sample of customers in each class. An individual customer who consumed no electricity during very hot summer days (because she was on vacation for half the month) would still have her measured monthly consumption allocated to some hot summer day hours based on her group's load profile. The load profile-based allocations must also satisfy an adding up property so that all power measured to have flowed through the distribution network is fully allocated to retail consumers. There are at least two ways to do this. One way is to load profile all customers without real-time meters whether they are served by competitive retailers or the distribution company providing default retail service. Another way is to load profile only the customers with traditional meters of competitive retailers and subtract the resulting hourly aggregates from the real-time metered consumption for the entire distribution system, leaving the distribution company/retailer with settlement obligations for the residual.

[10]Note that our setup is equivalent to assuming that the distribution grid owner purchases the power in the wholesale market and then resells it to each LSE based on the real time metered or load profiled consumption of the customers they have signed up. The access charge a is then the price LSEs pay to compensate the distribution grid for the costs of the wholesale power it has purchased on their behalf.

demand more than marginal demand. We will therefore be led to consider three cases:

$$\textit{Case 1}: \qquad \frac{E\left[p_i D_i(p)\right]}{E\left[D_i(p)\right]} > \frac{E\left[p_i D_i'(p)\right]}{E\left[D_i'(p)\right]}. \qquad \text{for all } p.$$

In this case, $p^* < \hat{p}$.

$$\textit{Case 2}: \qquad \frac{E\left[p_i D_i(p)\right]}{E\left[D_i(p)\right]} < \frac{E\left[p_i D_i'(p)\right]}{E\left[D_i'(p)\right]}. \qquad \text{for all } p.$$

In this case, $p^* > \hat{p}$.

$$\textit{Case 3}: \qquad \frac{E\left[p_i D_i(p)\right]}{E\left[D_i(p)\right]} = \frac{E\left[p_i D_i'(p)\right]}{E\left[D_i'(p)\right]}. \qquad \text{for all } p.$$

In this case, $p^* = \hat{p}$.

Examples: For the additive linear with state-contingent intercept case $D_i(p) = d_i - h(p)$, we are in case 1. For the multiplicative case, $D_i(p) = d_i h(p)$, then $p^* = \hat{p}$ (case 3).

Lemma 1. (i) *Cases 1 through 3 can be characterized by how the average wholesale cost price varies with the marginal retail prices:*

$a' > 0$ *in case 1*

$a' < 0$ *in case 2*

$a' = 0$ *in case 3.*

(ii) *In all cases,* $a(p) > p$ *for* $p < \widehat{p}$

$a(p) < p$ *for* $p > \widehat{p}$.

Proof: Part (i) is obtained by differentiating (5). To demonstrate part (ii), it suffices to show that $a'(p) < 1$ whenever $a(p) = p$, or after a few computa-

tions:

$$H(p) = E\left[(p - p_i)\, D_i' + D_i\right] > 0.$$

We know that $a(p) > p$ for p small (since $a(p) \geq E[p_i]$) and $a(p) \leq p_1 < p$ for p going to infinity. Hence, if the equation $a(p) = p$ has multiple solutions (an odd number greater than one) the function $H(p)$ must be increasing over at least some range. But $H'(p) = E\left[2D_i' + (p - p_i)\, D_i''\right] < E\left[D_i' + (p - p_i)\, D_i''\right] < 0$, a contradiction. ■

A retailer designs his offer so as to solve:

$$\max_{\{p,A\}} E\left[(p - a)\, D_i(p)\right] + A$$

s.t.

$$E\left[S_i\left(D_i(p)\right) - pD_i(p)\right] - A \geq \bar{U},$$

where $\overline{U}$ is the net surplus obtained by the consumer from subscribing with a rival retailer.

The retailer therefore selects p so as to maximize the joint surplus:

$$\max_p E\left[S_i\left(D_i(p)\right) - aD_i(p)\right],$$

or

$$(p - a)\, E\left[D_i'(p)\right] = 0,$$

yielding

$$p = a.$$

In equilibrium, a is given by (5). Hence

$$p = \hat{p}.$$

Furthermore, $A = 0$: Retailers charge no monthly fee and just pass their

variable cost of wholesale power through to the consumer.[11] Except in case 3, retail competition is, under load profiling, inconsistent with a Ramsey outcome.

For future reference, let U^{RC} ("RC" for "retail competition") denote the consumers' equilibrium utility:

$$U^{RC} \equiv E\left[S_i\left(D_i\left(\widehat{p}\right)\right) - \widehat{p}D_i\left(\widehat{p}\right)\right]. \tag{6}$$

Proposition 2 *Pure retail competition under load profiling delivers linear pricing at the average wholesale power cost $\widehat{p}$ despite the fact that LSEs have the possibility of offering two-part tariffs. The marginal price of electricity for the retail customer is therefore higher than the Ramsey price in case 1, and smaller in case 2; it is equal to the Ramsey price only in case 3.*

Remark 1: The Ramsey optimum can be achieved through a per customer subsidy or tax levied on retailers. Thus, let a retailer pay $\mathcal{A} + aQ$ when his customer consumes Q. The fixed charge $\mathcal{A}$ is over (or under) and beyond any delivery, metering and customer service costs that continue to reflect responsibilities of the distribution grid owner (these costs have been normalized at zero). Faced with an access tariff $(\mathcal{A}, a)$, retailers optimally pass this tariff through to their customers ($A = \mathcal{A}$ and $p = a$). The break-even constraint of the distribution grid owner is then:

$$\mathcal{A} + E\left[\left(a - p_i\right) D_i(a)\right] = 0.$$

The Ramsey outcome can be obtained by setting $a = p^*$, and then $\mathcal{A}$ so as to achieve budget balance, but (except in the non-generic case 3) this

[11] Borenstein-Holland (2003a,b) may appear to have derived an equivalent "third best" result. However, their analysis does not consider load profiling and assumes that LSEs must offer linear tariffs. We derive results for situations that reflect their assumptions in Section 3.

requires a departure from relying on load profiled consumption to calculate the wholesale price charged to retailers, in that the variable access charge differs (except in case 3) from the consumption-weighted average wholesale market price corresponding to the consumption induced by marginal price $p = a$.

Remark 2 (rationing): We have assumed away rationing. In the presence of rationing, the consumers' gross surplus and demand depend on the price they face, but also on the probability α_i of rationing in state of nature i. We refer to Joskow-Tirole (2004) for a detailed discussion; that paper in particular shows that in the absence of load profiling and provided that LSEs can choose the extent α_i of state-contingent rationing, then retail competition delivers the second-best outcome. For simplicity, let us focus here on the special case of perfectly foreseen outages associated with rolling blackouts. The consumers' gross surplus and demand in state i are then $\alpha_i S_i(D_i)$ and $\alpha_i D_i$. In the context of load profiling, it makes sense to assume that LSEs take α_i as exogenous, since rationing is zonal and retail competition with load profiling corresponds to competition within a zone. LSEs, as earlier, maximize the joint surplus:

$$\max_p \{E[\alpha_i [S_i(D_i(p)) - aD_i(p)]]\}$$

yielding:

$$E[\alpha_i (p-a) D_i'(p)].$$

Again, it is optimal for LSEs to pass the average wholesale price a onto consumers:

$$p = a.$$

And so

$$p = \frac{E[p_i \alpha_i D_i(p)]}{E[\alpha_i D_i(p)]}.$$

2.3 Incumbent distributor competing with independent retailers for load-profiled customers

Consider next the situation in which the distributor is also permitted to compete for load-profiled customers. We first assume that the LSE behaves so as to maximize profits for the parent company as a whole. We then observe that nothing is altered by a ring-fencing rule that requires the affiliate to maximize its own profits rather than those of the parent company.

a) Let us first show that the incumbent distributor's offers of the Ramsey tariff invites entry as long as $\hat{p} \neq p^*$. [As before $\widehat{p}$ is the marginal retail price that balances the LSE's budget in the absence of a fixed charge.] Suppose indeed that the distributor offer tariff (p^*, A^*). The load-profiled access charge or average wholesale power cost when the distributor serves all consumers is

$$a^* \equiv \frac{E\left[p_i D_i\left(p^*\right)\right]}{E\left[D_i\left(p^*\right)\right]}.$$

Consider an independent retailer contemplating a *small-scale entry* at some tariff $\left(\bar{p}, \bar{A}\right)$. We assume small-scale entry so that the entrant can take the access charge as given. Large-scale entry modifies the access charge that is assessed ex post, by modifying the average load profile. Alternatively, we could assume that a^* is fixed in advance based on Ramsey load profiles. This independent retailer entrant can make a positive profit provided that he offers a higher joint surplus than the Ramsey level.

Let

$$U\left(\bar{p}, a^*\right) \equiv E\left[S_i\left(D_i\left(\bar{p}\right)\right) - a^* D_i\left(\bar{p}\right)\right]$$

denote this joint surplus. Note that

$$U\left(p^*, a^*\right) = U^*.$$

Furthermore,

$$\frac{\partial U}{\partial \bar{p}}\left(\bar{p}, a^*\right) = E\left[\left(\bar{p} - a^*\right) D_i'\left(\bar{p}\right)\right],$$

and so the independent retailer optimally charges

$$\bar{p} = a^*.$$

The independent retailer entrant charges a higher variable price than the incumbent

$$a^* > p^*$$

if and only if

$$E\left[\left(p_i - p^*\right) D_i\left(p^*\right)\right] < 0 \quad \Longleftrightarrow \quad A^* > 0.$$

It may seem surprising that an entrant can (except in the non-generic case $a^* = p^*$ i.e., $\hat{p} = p^*$) enter against an incumbent offering the Ramsey tariff. The point is that the entrant benefits from an effective subsidy from the incumbent, who then operates at a loss given the entry.[12] The subsidy arises as a consequence of the fact that the distributor's obligation to wholesale suppliers is equal to the aggregate metered consumption for the entire distribution system net of the load profiled consumption assigned to independent retailers. As a corollary, the incumbent distributor cannot offer the Ramsey access charge.

b) Thus, assume that the incumbent distributor cum retailer is regulated so as to reach the Ramsey optimum in the presence of retail competition. That is, it is instructed to maximize social welfare subject to the budget balance condition; it charges prices (p, A). The variable charge paid by retailers for each kWh consumed by their retail customers, a, is based on the average load

[12] This loss is equal to

$$E\left[\left(p_i - a^*\right) D_i\left(a^*\right)\right] \propto \left[E\left[p_i D_i\left(a^*\right)\right] E\left[D_i\left(p^*\right)\right] - E\left[p_i D_i\left(p^*\right)\right] E\left[D_i\left(a^*\right)\right]\right].$$

profile of the incumbent's consumers; because the incumbent distributor can always duplicate what retailers do, we can assume without loss of generality that it serves the market (but, to serve the market, it must provide at least the net surplus offered by competitive retailers).

Let

$$V(p) \equiv U(p,p) = E\left[S_i\left(D_i(p)\right) - pD_i(p)\right]$$

with $V'(p) = -E\left[D_i(p)\right]$. The analysis in Section 2.2 implies that with load profiling competitive retailers optimally offer a linear tariff with price equal to the average wholesale power cost. And so competitive retailers offer consumer net surplus equal to $V\left(a(p)\right)$. Note further, that because entrants prefer to offer a linear price $a(p)$ to offering marginal price p and charging a fixed fee equal to the "deficit" $[a(p) - p]\, E\left[D_i(p)\right]$,

$$V\left(a(p)\right) \geq V(p) - [a(p) - p]\, E\left[D_i(p)\right]$$

with strict inequality unless $p = a(p)$, i.e., $p = \widehat{p}$.

The constrained Ramsey distributor cum retailer then maximizes the consumers' utility

$$\max_{\{p,A\}} \left[-A + V(p)\right]$$

subject to two constraints:

$$A + E\left[\left(p - p_i\right) D_i(p)\right] \geq 0$$

and

$$-A + V(p) \geq V\left(a(p)\right).$$

The first constraint is the incumbent distributor cum retailer's zero-profit condition, and the second is the contestability constraint created by the threat of entry by independent retailers.

From the budget constraint,

$$\begin{aligned} V(p) - A &\leq V(p) + E\left[(p - p_i) D_i(p)\right] = V(p) + \left[p - a(p)\right] E\left[D_i(p)\right] \\ &\leq V(a(p)), \end{aligned}$$

with strict inequality unless $a(p) = p$, or equivalently $p = \widehat{p}$. Hence, the incumbent distributor cum retailer can do no better than pure retail competition. Intuitively, the parent company by construction breaks even, and therefore the affiliate cannot do better than rival retailers, who compete with the same instruments.[13]

Remark on "ring-fencing": In the U.S. and UK there are affiliate rules that are designed to separate regulated lines of business (e.g. transmission and distribution) from unregulated lines of business (e.g. competitive generation and retailing). The rules typically require (a) cost separation to avoid cross-subsidization of unregulated lines of business by regulated lines of business, (b) information transfer restrictions that limit transfers of "private information" between regulated and unregulated affiliates, (c) transfer price rules requiring any services transferred from the regulated entity to the unregulated entity to reflect either their fair market value or a regulated price and (d) equal treatment regulations that require the regulated affiliates to offer services under the same terms and condition to unaffiliated companies competing with their unregulated affiliates as they offer to their unregulated affiliates. These rules are designed to define constraints on the ability of a vertically integrated firm to maximize the joint profits of the entire enterprise. In our set-up such additional constraints on the affiliates have no impact, as

[13]More generally, the incumbent distributor cannot deliver a net surplus to consumers in excess of $V(\widehat{p})$ by serving some consumers but not all. To see this, note that the retail affiliate must make a non-negative profit (since a is computed so that the parent company always breaks even). By the same reasoning as above, the retail affiliate cannot offer more than $V(p) + E\left[(p - a) D_i(p)\right] < V(a)$ unless $p = a$. But if $p = a$, everyone (affiliate, independent retailers) offers retail price a, and so $a = \widehat{p}$.

the combination of break-even access charges and retail competition completely deprives the vertically integrated incumbent of any discretion.[14]

Proposition 3 *Under load profiling and retail competition, the Ramsey optimum is generically not attainable. The incumbent retailer in the constrained Ramsey optimum charges the average wholesale power cost price* $\widehat{p}$.

Remark (lagged computation of the average wholesale power cost): We have assumed that settlements occur "ex post", so a is computed on the basis of the actual aggregate consumption pattern over the period. Alternatively, one could compute a^t at date t on load profiling using date-$(t-1)$ data. Suppose that the incumbent distributor is instructed to maximize intertemporal social welfare subject to an intertemporal budget balance condition with discount factor δ, and to the contestability condition:

$$-A^t + V\left(p^t\right) \geq V\left(a\left(p^{t-1}\right)\right) \quad \text{for all } t.$$

It can be shown that the resulting constrained Ramsey price is stationary: $p^t = p^{**}$, with:

$$E\left[\left(p^{**} - p_i\right) D_i'\left(p^{**}\right)\right] = -\delta\left(\frac{\mu - 1}{\mu}\right)\left(E\left[D_i\left(p^{**}\right)\right]\right) a'\left(p^{**}\right).$$

where μ is the shadow price of the intertemporal budget balance constraint.

For $\delta = 1$, the solution is $p^{**} = \widehat{p}$ (with $\mu = \infty$). For $\delta = 0$, then $p^{**} = p^*$. And, more generally, it can be shown that the optimal policy narrows the gap between the unconstrained Ramsey price p^* and the average wholesale power cost: $p^* < p^{**} < \widehat{p}$ in case 1, $p^* > p^{**} > \widehat{p}$ in case 2, $p^* = p^{**} = \widehat{p}$ in case 3.

[14] As suggested above, ring-fencing in practice serves a different purpose: It aims at preventing the shifting of the costs of the unregulated affiliate company to the regulated distribution company and thus to the ratepayers.

3 Partially price responsive consumers with real-time meters

Let us now follow Borenstein and Holland (2003a,b) and assume that consumers are equipped with a real-time meter, so that load profiling is not necessary.[15] Consumers react imperfectly to the real time prices $\widehat{p}_i$ that they face (these real time prices $\widehat{p}_i$ are chosen by the LSE and can therefore differ from the wholesale prices p_i). Borenstein and Holland (BH) depict this situation by assuming that (a) a consumer reacts only to the *average* usage price $\widehat{p} = E\left[\widehat{p}_i\right]$ that he pays and not to the state-contingent price $\widehat{p}_i$, and (b) his demand, $D_i(p)$, by contrast, is state-contingent. The BH representation presumes some bounded rationality on the consumer's part. For, a rational consumer ought to realize (at least) that the state of nature i she reacts to and the price she will pay, $\widehat{p}_i$, are correlated; for example, an American consumer should realize that the use of air conditioning in a hot weather condition is correlated with high electricity prices.

We now investigate sequentially the cases of rational and boundedly rational consumers. Rational consumers react imperfectly to the price profile that is offered to them by the LSE, but they make efficient use of the (endogenously imperfect) knowledge of this price profile and they trade off optimally the transaction cost involved in improving their monitoring of the price profile and in optimizing the usage of equipment, and the corresponding savings in their electricity bill.

[15] A potential argument against the use of RTP for consumers is that it would obfuscate price comparisons with existing tariffs; however, websites already facilitate such price comparisons in the case of consumers with traditional meters. Another potential argument against RTP relates to the consumers' solvency or risk aversion; LSEs however could bundle small-scale "contracts for differences" with their supply contracts for consumers with real-time meters.

3.1 Rational consumers

Let us first motivate our analysis of rational consumers by a couple of examples, and then build a general theory.

Example 1: Suppose that the state of nature is (ij) where i and j each belong to $[0,1]$. The joint density is denoted f_{ij}. The wholesale price is p_{ij}. The consumer observes i (the local weather), but not j (the weather elsewhere, or the availability of the transmission lines or generators).[16] The observable and unobservable components of uncertainty may be correlated. The consumer's gross surplus $S_i(q)$ depends only on the observable part of the state of nature. Let $\widehat{p}_i = E_j[\widehat{p}_{ij}]$ denote the average marginal price when the observable component is i. Thus, a rational consumer chooses his consumption q_i when observing event i so as to solve:

$$\max\left\{S_i(q_i) - \widehat{p}_i q_i\right\},$$

defining a demand function $q_i = D_i(\widehat{p}_i)$.

The Ramsey optimum is then given by:

$$\max_{\{\widehat{p}_\cdot, A\}} \left\{E\left[S_i\left(D_i\left(\widehat{p}_i\right)\right) - \widehat{p}_i D_i\left(\widehat{p}_i\right)\right] - A\right\}$$

s.t.

$$E\left[\left(\widehat{p}_i - p_{ij}\right) D_i\left(\widehat{p}_i\right)\right] + A \geq 0,$$

or

$$\max_{\{\widehat{p}_\cdot\}} \left\{E\left[S_i\left(D_i\left(\widehat{p}_i\right)\right) - p_i D_i\left(\widehat{p}_i\right)\right]\right\}$$

where $p_i \equiv E_j[p_{ij}]$ is the average wholesale price when the observable component is i. The optimal policy is therefore a passthrough of the wholesale

[16] We here take this information structure as given. Presumably, it results from some optimization as in the more general model considered below.

price: $\widehat{p}_i = p_i$, which can for example be obtained by:

$$\widehat{p}_{ij} = p_{ij},$$

and to charge no fixed fee: $A = 0$. Furthermore, LSE competition delivers this optimal passthrough.

Example 2: Let us next give an example in which the consumer does not observe the state of nature, yet his consumption is state-dependent. Consider equipment (e.g., space heater, air conditioning, pool heater) that, for a given quality of service s (e.g., indoor temperature set once and for all by the consumer) consumes a state-contingent amount of electricity. The real-time price profile of electricity affects the quality s (for example, an increase in winter prices lowers the indoor temperature chosen by the consumer or induces the consumer to switch to oil heat).

More formally, letting j be the full description of the state of nature, the consumer, who does not observe j, sets s so as to maximize his net surplus, equal to the gross surplus $S(s)$ minus the electricity bill:

$$E_j \left[S(s) - \widehat{p}_j D_j(s) \right] - A$$

where $D_j(s)$ is the state-contingent consumption needed to reach level s (for example a given swimming pool temperature requires a higher consumption of electricity when the weather is cold). Let $s\left(\widehat{p}_\cdot\right)$ denote the selected setting. The Ramsey optimum then solves:

$$\max_{\{\widehat{p}_\cdot, A\}} \left\{ E \left[S\left(s\left(\widehat{p}_\cdot\right)\right) - \widehat{p}_j D_j\left(s\left(\widehat{p}_\cdot\right)\right) \right] - A \right\}$$

s.t.

$$E\left[\left(\widehat{p}_j - p_j \right) D_j\left(s\left(\widehat{p}_\cdot\right)\right) \right] + A \geq 0,$$

or

$$\max_{\{\widehat{p}_\cdot\}} \left\{ E \left[S\left(s\left(\widehat{p}_\cdot\right)\right) - p_j D_j\left(s\left(\widehat{p}_\cdot\right)\right) \right] \right\}$$

Again, the Ramsey optimum (or the LSE's equilibrium offer for that matter) is obtained by a passthrough policy:

$$\widehat{p}_j = p_j.$$

Let us now consider a more general environment and further allow the consumer to choose his degree of awareness of the real time price. Namely, let ω denote the state of nature (for instance, $\omega = (ij)$ in Example 1). Let $\mathcal{P}$ denote the consumer's partition (for example, $\mathcal{P}((ij)) = i$ in Example 1). That is, the consumer observes that ω belongs to an event $\mathcal{P}(\omega)$. Let $C(\mathcal{P})$ denote the total transaction cost associated with partition $\mathcal{P}$; one has in mind that choosing a finer partition $\mathcal{P}$ (for example, keeping informed of the real time price) is costly, although we will not need to make this assumption.

The consumer in state ω takes a decision s that is measurable with respect to partition $\mathcal{P}$. This decision can be his electricity consumption as in Example 1, but can be different from the consumption, as illustrated by Example 2. Let $D(s(\mathcal{P}(\omega)), \omega)$ denote the associated consumption. Letting $S(s, \omega)$ denote the consumer's gross surplus, and $\widehat{p}_\omega$ the usage price charged to the consumer, for event P in the partition, $s(P)$ is given by

$$V(P) = \max_s \{E[S(s,\omega) - \widehat{p}_\omega D(s,\omega) \mid \omega \in P]\}$$

and $\mathcal{P}$ is given by:

$$\max_{\mathcal{P}} \{E_{P\in\mathcal{P}}[V(P)] - C(\mathcal{P}) - A\}.$$

The budget constraint writes:

$$E[(\widehat{p}_\omega - p_\omega) D(s(\mathcal{P}(\omega)), \omega)] + A \geq 0.$$

Hence, the consumer's utility is

$$\max_{\mathcal{P}} \left\{ E_{P \in \mathcal{P}} \left[\max_{s} E \left[S(s,\omega) - p_\omega D(s,\omega) \mid \omega \in P \right] \right] - C(\mathcal{P}) \right\}. \qquad (7)$$

This utility is maximized when the consumer is confronted with the wholesale prices: $\widehat{p}_\omega = p_\omega$.

Proposition 4 *With real-time meters and imperfectly reactive, but rational consumers:*

(i) the Ramsey optimum (consumption decision, consumer's information) is obtained when the consumer pays the real time wholesale price associated with her actual consumption pattern;

(ii) retail competition delivers the Ramsey optimum.

3.2 Boundedly rational consumers

Let us next assume that, as in Example 1 above, consumers observe component i of the state of nature (ij), although not the real time price p_{ij}, but fail to realize that they are representative of the consumer sample and that the wholesale price is correlated with their demand. Their demand D_i depends only on the average price $\widehat{p} \equiv E[\widehat{p}_{ij}]$ that they are offered: $D_i(\widehat{p})$. In a sense, these consumers suffer from what Kahneman and Tversky (1973) call the base-rate fallacy: they make insufficient use of their prior beliefs and incorrectly believe that because they do not observe the RT price, they are facing the average price.

The Ramsey planner solves

$$\max_{\{\widehat{p},A\}} \left\{ E\left[S_i\left(D_i(\widehat{p}) \right) - \widehat{p} D_i(\widehat{p}) \right] - A \right\}$$

s.t.

$$E\left[\left(\widehat{p} - p_{ij} \right) D_i(\widehat{p}) \right] + A \geq 0.$$

Eliminating A and maximizing with respect to $\widehat{p}$ yields

$$E\left[\left(\widehat{p}-p_i\right)D_i'\left(\widehat{p}\right)\right]=0$$

where $p_i \equiv E_j\left[p_{ij}\right]$. Retail competition as usual delivers the same outcome.

Proposition 5 *With real-time meters and boundedly rational consumers:*
(i) the Ramsey optimal average usage price $\widehat{p}$ is the same as under traditional meters:

$$E\left[\left(\widehat{p}-p_i\right)D_i'\left(\widehat{p}\right)\right]=0;$$

budget balance is achieved by setting A appropriately.
(ii) retail competition still delivers the (second-best) Ramsey optimum.

Remark: Part (i) of Proposition 5 can be found in Borenstein and Holland. They do not find the Ramsey outcome under retail competition (their outcome actually is identical to the outcome of retail competition under traditional meters and load profiling — see Section 2.2), because they constrain LSEs to offer linear tariffs.

4 Non-scale heterogeneity and competitive screening

For expositional simplicity, we have assumed that consumers are homogeneous (perhaps up to a size factor σ). This section investigates the implications of consumer heterogeneity for retail competition.

Suppose that there are different classes of consumers $h \in [0,1]$ with state-contingent demands $D_i^h(p)$ and state-contingent surplus $S_i^h\left(D_i^h(p)\right)$. Let n^h denote the frequencies of consumers of type h, and $E_h\left[\cdot\right]$ denote the expectations with respect to consumer types (the expectations with respect to the state of nature are now labeled $E_i\left[\cdot\right]$). Let us begin with a few general remarks.

The first is that under *load profiling*, the analysis of retail competition is a simple generalization of that in Section 2. The retailers charge a linear price $\widehat{p}$ given by

$$\widehat{p} = \frac{E_i\left[E_h\left[p_i D_i^h\left(\widehat{p}\right)\right]\right]}{E_i\left[E_h\left[D_i^h\left(\widehat{p}\right)\right]\right]}$$

and fail to achieve the (second-best) Ramsey optimum. In particular, the retailers face no adverse selection problem to the extent that they pay a per-kWh price $a = \widehat{p}$ that is independent of the type of consumers they end up attracting.

Neither do LSEs face an adverse selection problem when dealing with rational consumers on real-time meters:[17] Proposition 4 above showed that it is optimal for LSEs to pass the wholesale price through to the consumer. And so at the optimal contract, the LSE breaks even on usage in each state of nature. Its profit is therefore unaffected by the consumer's actual load profile. Consumer heterogeneity then has no impact on competitive outcomes.

Competitive screening[18] issues arise only in the case of bounded rational consumers with real time meters (Section 3.2). This is the case because passthrough of wholesale prices is then in general suboptimal, consumers differ in their load profiles and LSEs need to be careful of who they attract. This situation is reminiscent of Rothschild and Stiglitz's celebrated treatment of insurance markets (1976).

A complete analysis of competitive screening with boundedly rational consumers with real time meters in retail electricity markets lies out of the scope of the paper. Rather, we will content ourselves with an illustrative

[17] They may face forms of adverse selection unrelated to the consumer's load profile. For example, LSEs may try to obtain superior information about the probability of consumer default.

[18] Much of the theory of competitive non-linear pricing has been developed in the context of private values, that is when suppliers care solely about price and quantity, and not, per se, to whom they sell: See Rochet-Stole (2002, 2003). Here the context is one with common values.

example. Time is uniformly distributed on $[0,1]$. The state-contingent price (depicted by the dotted line in Figure 1) is increasing linearly in i:

$$p_i = i.$$

Consumers at any period $i \in [0,1]$ consume 0 or 1 unit of electricity. Suppose that there are two categories of consumers. Their state-contingent willingness to pay is depicted in Figure 1.

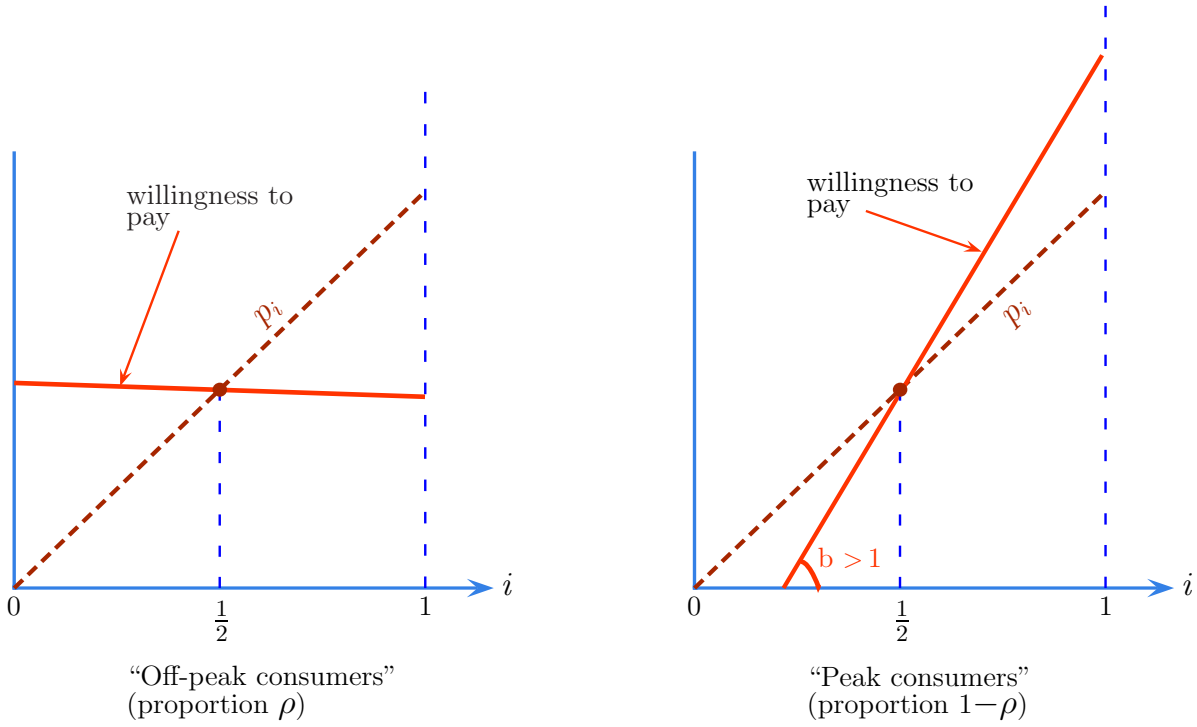

Figure 1

Off-peak consumers, when consuming q units (i.e., when consuming a fraction of time q), consume during $[0,q]$. Their gross surplus is then[19]

$$S_1(q) \equiv \frac{1}{2}q.$$

Peak consumers prefer to consume at peak and obtain gross surplus $\frac{1}{2} +$

[19] A constant willingness to pay is chosen for computational simplicity. To have them strictly prefer to consume their allotment q off peak, one can have in mind a gross surplus of unit q equal to $1/2 + \varepsilon(q)$ with $\varepsilon' < 0$, and then take the limit as $\varepsilon(q)$ converges to 0 uniformly.

$b\left(i-\frac{1}{2}\right)$ from consuming one unit in state i, where $b>1$; thus, their gross surplus from consuming a fraction q of the time is

$$S_2(q)=\int_{1-q}^{1}\left[\frac{1}{2}+b\left(i-\frac{1}{2}\right)\right]di=\frac{(b+1)q-bq^2}{2}.$$

The cost of wholesale electricity purchases for both types are, respectively:

$$C_1(q)=\int_0^q idi=\frac{q^2}{2}$$

and

$$C_2(q)=\int_{1-q}^{1} idi=\frac{2q-q^2}{2}>C_1(q)\quad\text{whenever } 0<q<1.$$

Letting ρ and $1-\rho$ denote the fraction of off-peak and peak consumers and ignoring in a first step incentive compatibility, the Ramsey optimum solves:

$$\max_{\{q_1,q_2\}}\left\{\rho\left[S_1\left(q_1\right)-C_1\left(q_1\right)\right]+(1-\rho)\left[S_2\left(q_2\right)-C_2\left(q_2\right)\right]\right\},$$

yielding, as one would expect,

$$q_1^*=q_2^*=\frac{1}{2}.$$

To be incentive compatible, the two types must pay the same total amount: $T_1^*=T_2^*$ since they consume the same amount. Accordingly, achieving the Ramsey allocation in the absence of retail competition requires cross-subsidies. As

$$\rho\left[T_1-C_1\left(q_1^*\right)\right]+(1-\rho)\left[T_2-C_2\left(q_2^*\right)\right]=0$$

$$C_1\left(q_1^*\right)<T_1^*=T_2^*<C_2\left(q_2^*\right).$$

Let us now consider retail competition. LSEs can safely offer to peak consumers their symmetric allocation contract $(q_2^*,\ T_2=C_2\left(q_2^*\right))$, since they would make money if the off-peak consumer were to take this contract. Let us look for conditions under which the market offers $(q_2^*,\ T_2=C_2\left(q_2^*\right))$ to the

peak consumers and $(q_1 = \widehat{q}_1,\, T_1 = \widehat{T}_1 = C_1(\widehat{q}_1)\ )$ to the off-peak consumers. The incentive-compatibility-constraint is

$$S_2(q_2^*) - C_2(q_2^*) \geq S_2(q_1) - C_1(q_1),$$

or

$$\frac{b-1}{8} \geq \frac{(b+1)(q_1 - q_1^2)}{2}.$$

This latter condition, satisfied with equality, defines a unique $\widehat{q}_1$ in $\left(\frac{1}{2}, 1\right)$.[20]

This separating equilibrium is an equilibrium provided that no LSE can offer a pooling contract with higher payoffs for both types. This is the case if $\rho \leq \rho^*$ for some $\rho^* \in (0,1)$.[21] However, this equilibrium does not achieve the Ramsey optimum.

Proposition 6 *With heterogeneous consumers:*

(i) Adverse selection does not arise when consumers are either on traditional meters and load profiled, or on real-time meters and rational. The analysis of Sections 2 and 3.1 thus generalizes to heterogeneous consumers.

(ii) By contrast, with boundedly rational consumers on real-time meters, adverse selection and the concomitant competitive screening prevent retail competition from achieving the Ramsey outcome, unlike in the case of homogeneous consumers.

5 Incentives to install real-time meters and communication equipment

Let us investigate the consequences of the previous analysis of the case where there are traditional meters, load profiling and retail competition for retail-

[20] In our example $S_1(q_1) - C_1(q_1) = S_1(q_1) - S_2(q_1) + U_2^*$ is symmetric around 1/2, and so type 1 is indifferent between separating at $\widehat{q}_1$ or $1 - \widehat{q}_1$.

[21] To prove this, maximize $S_1(q_1) - T_1$, subject to the incentive compatibility constraint: $U_2^* + \Delta \geq S_2(q_1) - T_1$ (where $U_2^* = S_2(q_2^*) - C_2(q_2^*)$ and $\Delta \geq 0$) and to the break-even condition: $\rho[T_1 - C_1(q_1)] - (1-\rho)\Delta \geq 0$.

ers' incentives to install real-time meters with or without communication, starting with the Ramsey incentives. Suppose that consumers have the same load profile but differ in the size σ of their demand: Consumer of type σ has demand $q_i = \sigma D_i(p)$ and surplus $\sigma S_i\left(q_i/\sigma\right)$. There is a continuous distribution of consumers σ on $[0, \infty)$.

Consumers initially have traditional meters and thus cannot react to the RTP. Two types of equipment can be added to a traditional meter:[22]

- *a real-time meter*, costing $m > 0$, that measures and makes verifiable the consumer's RT consumption, but makes this consumption imperfectly reactive to the RTP as in Section 3;

- *communication* (on top of real-time metering), costing $M > m$, that furthermore makes it possible for consumers to perfectly react to the RT prices through remote control of appliances and equipment.

Ramsey benchmark.

Consider a rational consumer with type $\sigma = 1$. Let U^{FB} be the utility that this consumer could obtain if her consumption could adjust efficiently to variations in the real time price (see (1) above). Let U^* be the second-best utility that could be achieved by a Ramsey social planner for the consumer with a traditional meter (see (3) above). And let $U^{**} \in \left(U^*, U^{FB}\right)$ denote her utility when endowed with a real-time meter without communication (See (7)) above. The utilities of a consumer with type σ are equal to σ times these utilities. The Ramsey planner (or a monopoly retailer) would endow consumers in (σ^*, σ^{**}) with a real-time meter, and those in (σ^{**}, ∞) with real

[22]Note that we assume that there are no returns to scale in installing equipments. In practice, LSEs incur costs, such as wireless bay stations enabling remote real time recording, that are common across consumers in a neighborhood. Such costs give rise to non-convexities and inefficiencies unless they are shared among LSEs.

time meters plus communication, where:[23]

$$\sigma^* = \frac{m}{U^{**} - U^*} \quad \text{and} \quad \sigma^{**} = \frac{M - m}{U^{FB} - U^{**}}.$$

Load profiling.

We keep the assumption that the consumption of retail consumers with traditional meters is load profiled using the load profile of the consumers in that class. Under perfect retail competition with load profiled consumers, the consumer obtains σU^{RC} when keeping a traditional meter, $\sigma U^{**} - m$ when equipped with a real-time meter, and $\sigma U^{FB} - M$ when equipped with communication.

Simple derivations yield:

Proposition 7 *(i) Under pure retail competition with load profiling:*

- *Consumers with type $\sigma \geq \sigma^{**}$ are equipped with communication, where σ^{**} is the Ramsey level.*
- *Consumers with type $\sigma \in \left[\sigma^{RC}, \sigma^{**}\right)$ are equipped with real-time meters, when $\sigma^{RC} = m/\left[U^{**} - U^{RC}\right] < \sigma^*$, the Ramsey level.*

(ii) Consequently, there is more investment in meters that measure real-time consumption than in the Ramsey optimum. Given the inefficiencies introduced by the combination of load profiling and retail competition, however investments are socially optimal.

The constrained efficiency of market-determined investment in metering equipment (part (ii) of the proposition) deserves some comment. There are really two Ramsey benchmarks, one unconstrained by retail competition and the other constrained by retail competition. The investments are socially

[23] Assuming $(U^{**} - U^*) M \geq \left(U^{FB} - U^*\right) m$. Otherwise, it is not optimal to install real-time meters without communication.

optimal given the inefficiencies created by retail competition with load profiling.

6 The joint interruptibility problem

In our companion paper (Joskow-Tiro1e 2004) we derive the efficient prices and investment program for an electricity market with demand uncertainty, price insensitive consumers, and LSEs that can choose any level of rationing they prefer contingent on the real time price. We then identify the assumptions required for a competitive wholesale and retail market equilibrium to achieve this efficient price and investment program. One of the key assumptions is that different users can choose and the system operator can implement different levels of priority in rationing that reflect users' individual preferences. The validity of this assumption requires the system operator to be able physically to cut off individual retail consumers.

There is no theoretical reason why individual customers cannot be rationed. It requires installing communications and control equipment between the customer's connection to the network and the control center. However, this equipment is costly. As a practical matter, except for very large customers that have direct control equipment, most directed interruptions must occur at points on the network ("zones") that can be controlled by the distribution network operator.[24] The affected zone has (a) customers served by multiple LSEs that compete with one another (so every house on a street can be "served" by a different LSE) and (b) customers with heterogeneous

[24]In reality, system operators generally try to squeeze out all of the price sensitive demand first before they start rolling blackouts. This may not be optimal of course. There is also some priority rationing in that circuits with hospitals and fire stations, etc. will often be placed on a "do not blackout list." In this case, all customers on the same circuit get the benefit of being near a fire station or hospital. This example illustrates the fact that different consumers may have different values of lost load, and that furthermore the dispatcher cannot fine-tune the intensity of rationing.

preferences.

An optimal dispatch when zones but not individual consumers are controlled by the system operator must elicit each zone's *aggregate* willingness to pay for being served. From the point of view of the set of LSEs and industrial users in a given zone, reliability is a *public* good.

In principle, one can make use of the theory of public goods in order to design incentive-compatible mechanisms of elicitation of individual preferences for reliability.[25] For instance, one could use the Clarke (1971)-Groves(1973) scheme. Suppose that, due to a shortage in supply, the ISO must shut down one of cities A,B,C,... To simplify computations, cities demand the same load. Within city A, say, there are n users, each demanding 1 unit of load and having valuations (VOLL) v_i, which are private information. These users can either be price-sensitive, industrial users or LSEs serving price-insensitive users. Let the ISO shut down the city with the lowest total declared willingness to pay. That is, city A is served if and only if

$$\hat{V}_A \equiv \sum_{i \in A} \hat{\mathrm{v}}_i \geq \hat{V}$$

where $\hat{V}$ is the lowest total declared willingness to pay among other cities. City A then pays $\hat{V}$. The problem then boils down to a standard public good problem (the cost of getting the public good is $\hat{V}$-possibly unknown to members of city A, but this does not matter as this value is revealed through the aggregate bids in other cities).

In particular, use can be made of Clarke-Groves mechanisms : Member i of city i pays

[25] See Green-Laffont (1979a,b) for the general theory of public goods.

$$\begin{cases} \hat{V} - \sum_{j \neq i} \hat{\mathrm{v}}_i & \text{if} \quad \hat{\mathrm{v}}_i + \sum_{j \neq i} \hat{\mathrm{v}}_j \geq \hat{V} \\ 0 & \text{otherwise.} \end{cases}$$

Telling the truth ($\hat{\mathrm{v}}_i = \mathrm{v}_i$) is then a dominant strategy. [The Clark-Groves mechanism does not balance the ISO's budget, but a variant of it (the d'Aspremont-Gerard Varet (1979) scheme) does so in expectation.]

Besides transaction costs, there is under retail competition a major snag with such zonal voting mechanisms. While large industrial users' willingness to pay for reliability is not distorted by competition for the final consumer,[26] competing retailers' profit in a given zone depends only on the *relative* quality of their offer as compared with their competitors'. A retailer that bids for reliability increases the quality of service to its retail consumers, but it also increases its rivals' quality of service by the same amount, bringing no extra profit. This is best seen when considering the following timing: First, LSEs bid for reliability ($\widehat{v}_k^z$ for LSE k in zone z). Second, given the resulting reliability in each zone z, they compete for retail consumers. Given that they make no profit at stage 2, LSEs aim at minimizing expenditure at state 1 (they have de facto willingness to pay $v_k^z = 0$ in reference to our previous discussion).

Proposition 8 *Zonal rationing implies that the demand for rationing in a given zone is an aggregated demand.*

(i) In the absence of transaction costs, the constrained optimum can be obtained through standard public goods mechanisms if consumers are (non-competing) industrial users and retail consumers served by a monopoly distributor.

(ii) By contrast, the elicitation of consumers' willingness to pay for non-

[26] Unless two large industrial users both compete on the product market and produce in the same zone.

interruptibility is problematic under retail competition. In particular, if LSEs bid for reliability and then compete for retail consumers, no information can be obtained from LSEs concerning the consumers' demand for non-interruptibility.

7 Conclusion

In our companion paper (Joskow-Tirole 2004) we derive the optimal prices and investment program when there is state contingent demand, at least some consumers do not react to real time prices, but their LSE can choose any level of rationing it prefers contingent on real time prices. In this model consumers are identical, possibly up to a proportionality factor, and therefore all have the same load profile. We then derive the competitive equilibrium under these assumptions when there are competing LSEs that can offer two-part tariffs. This leads to a proposition that extends the standard welfare theorem to price-insensitive consumers and rationing; this proposition serves as an important *benchmark* for evaluating a number of non-market obligations and regulatory mechanisms:

The second best optimum (given the presence of price-insensitive consumers) can be implemented by an equilibrium with retail and generation (wholesale) competition provided that:

(a) The real time wholesale price accurately reflects the social opportunity cost of generation.

(b) Rationing, if any, is orderly, and makes efficient use of available generation.

(c) LSEs face the real time wholesale price for the aggregate consumption of the retail customers for whom they are responsible.

(d) *Consumers who can react fully to the real time price are not rationed. Furthermore, the LSEs serving consumers who cannot fully react to the real time price can demand any level of rationing they prefer contingent on the real-time price.*

(e) *Consumers have the same load profile (they are identical up to a scale factor).*

The assumptions underlying this benchmark proposition are obviously very strong. Our companion paper examines the implications of relaxing assumptions (a) and (b). This paper focuses on retail competition and examines the implications of departures from assumptions (c), (d) and (e).

When retail consumers are on traditional meters which measure their aggregate consumption over relatively long time periods rather than in real time, neither retail consumers nor, under retail competition, the LSEs responsible for purchasing the power required to serve their demand face the real time wholesale prices associated with the power they consume from the system. We derive the Ramsey optimal two-part tariffs given consumer insensitivity to the real time price and show that when there is retail competition with load profiling the Ramsey optimal prices are not a competitive equilibrium. In particular, the competitive retail market equilibrium involves linear average wholesale cost pricing rather than more efficient two-part tariffs. We go on to examine competition between independent LSEs and the incumbent distributor which also has the responsibility to serve retail consumers who are not served by independent LSEs. We show that independent LSEs can enter profitably against the incumbent if the incumbent offers the Ramsey optimal prices and that requiring the incumbent distributor through an affiliate that is "ring-fenced" does not change its behavior.

We next examine cases where consumers have real time meters but are either unresponsive or only partially responsive to variations in real time prices. In general, the Ramsey optimum is achieved with retail competition when consumers are identical up to a scaling factor and are rational.

We then extend the analysis to non-scale heterogeneity. Remarkably, adverse selection and the concomitant competitive screening do not arise under either load profiling or RT metering provided that consumers are rational. By contrast, competitive screening arises and severely distorts the allocation under boundedly rational consumers.

We go on to examine the incentives competing LSEs face to install two types of advanced metering equipment when consumers initially have traditional meters. We find that there is more investment in meters that measure real time consumption than in the Ramsey optimum. However, given the inefficiencies introduced by the combination of load profiling and retail competition, investments in advanced metering are socially optimal.

Finally, we consider the effects of the inability of the system operator physically to cut off individual customer loads. Instead rationing must be done on a "zonal" basis, perhaps involving rationing of both price sensitive and price insensitive retail consumers. This physical constraint means that individual retail customers cannot obtain their preferred priority for rationing by the system operator. Given this constraint, the Ramsey social planner could turn to standard public goods mechanisms to determine the relative priorities of the different zones that are physically capable of being cut off by the system operator and use this information to establish a second-best priority cutoff schedule. By contrast, in the presence of retail competition no information can be obtained by LSEs concerning their consumers' demand for non-interruptibility because they would prefer to free ride on the other

LSEs serving consumers in the same zone.

References

[1] Borenstein, S., and S. Holland (2003a) "Investment Efficiency in Competitive Electricity Markets With and Without Time-Varying Retail Price," CSEM WP 106R.

[2] Borenstein, S., and S. Holland (2003b) "On the Efficiency of Competitive Electricity Markets With Time-Invariant Retail Prices," CSEM WP 116.

[3] Clarke, E. (1971) "Multipart Pricing of Public Goods," *Public Choice*, 2: 19–33.

[4] D'Aspremont, C., and L.A. Gerard Varet (1979) "Incentives and Incomplete Information," *Journal of Public Economics*, 11: 25– 45.

[5] Green, J., and J.J. Laffont (1979a) *Incentives in Public Decision Making*, Amsterdam: North Holland.

[6] —— , eds. (1979b) *Aggregation and Revelation of Preferences*, North-Holland.

[7] Groves, T. (1973) "Incentives in Teams," *Econometrica*, 41: 617–631.

[8] Joskow, P., and J. Tirole (2004) "Reliability and Competitive Electricity Markets" mimeo, MIT and IDEI.

[9] Kahneman, D., and A. Tversky (1973) "On the Psychology of Prediction," *Psychological Review*, 80: 237–251.

[10] Rochet, J.C., and L. Stole (2002) "Nonlinear Pricing with Random Participation," *Review of Economic Studies*, 69(1): 277–311.

[11] —— (2003) "The Economics of Multidimensional Screening," in M. Dewatripont, L. Hansen and S. Turnovsky eds., *Advances in Economic Theory*. Cambridge University Press.

[12] Rothschild, M., and J. Stiglitz (1976) "Equilibrium in Competitive Insurance markets: An Essay in the Economics of Imperfect Information," *Quarterly Journal of Economics*, 90: 629–650.