16 Nonlinear pricing

Raghuram Iyengar and Sunil Gupta

Abstract
A nonlinear pricing schedule refers to any pricing structure where the total charges payable by customers are not proportional to the quantity of their consumed services. We begin the chapter with a discussion of the broad applicability of nonlinear pricing schemes. We note that the primary factor for the use of such schemes is the heterogeneity of the customer base. Such heterogeneity of preferences leads customers to choose different pricing plans based on their expected demand. We describe past analytical and empirical research. Past analytical work is categorized based on whether it is in a monopoly setting or a more general oligopoly context. Most past research has found two-part tariffs to be optimal in many settings. More recent research has begun to investigate the limits of such optimality and when a more general pricing scheme can be optimal. In the summary of empirical research on multi-part tariffs, we note that while nonlinear pricing schemes are popular, any analysis of demand under such schemes is nontrivial. One important reason is the two-way relationship between price and consumption in multi-part tariffs – the pricing scheme influences consumption and the level of consumption determines the applicable per-unit price. We describe how researchers have addressed this and other such issues and then show a modeling framework that integrates all the issues. We end by discussing empirical generalizations, which also suggest some promising areas for future research.

1. Introduction

A nonlinear pricing schedule refers to any pricing structure where the total charges payable by customers are not proportional to the quantity of their consumed services. The most common form is quantity discount for the purchase of large volumes. Several other forms of such pricing schemes exist across different industries. The following examples show the ubiquitous nature of this pricing strategy.

1. *Telecommunications* Most long-distance providers charge customers based on a combination of fixed fees (for access to the service) and per-minute price for each minute of a long-distance call. Wireless companies also charge customers in a similar manner for consumption of minutes but typically include some free minutes of consumption, along with a service plan.
2. *Consumer packaged goods* Quantity discounts are common in the consumer packaged goods industry. Typically, the per-unit price declines with package size. For instance, a recent search on Netgrocer.com showed that an 8 oz can of original B & M baked beans cost $1.39, which translates to $0.17/oz. A 16 oz can of the same baked beans cost $2.19, which is $0.14/oz. Some past research such as Nason and Della Bitta (1983) shows that consumers expect such quantity discounts.
3. *Electricity and water supply* Utility companies also offer quantity discounts. For instance, higher levels of consumption cost less for each kilowatt of consumption. In addition, energy rates for business users are different from those for residential users. Business users also incur varying rates based on peak versus off-peak electricity consumption.

4. *Business-to-business transactions* Many businesses offer quantity discounts to their customers. For instance in the electricity industry, customers purchasing large quantities of power have a high utilization as well. A quantity discount acknowledges the lower cost of idle capacity for such customers. Similar instances occur in the newspaper advertising industry, where businesses that advertise with a high frequency get charged at a lower rate per advertisement. See Dolan (1987) for a detailed discussion of various aspects of quantity discounts.
5. *Magazine subscriptions* Most magazines offer a lower rate for a two- or three-year subscription compared to the one-year subscription rate.

These examples show that nonlinear pricing takes many different forms. The purpose of this chapter is to summarize the research on nonlinear pricing. In Section 2, we explain the different kinds of nonlinear pricing schemes and discuss why such pricing schemes are used. Section 3 discusses the relevant managerial decisions for implementing such schemes. This is followed by a discussion in Section 4 on the theoretical findings on nonlinear pricing. In Section 5, we focus on empirical studies. Section 6 concludes the chapter.

2. Nonlinear pricing schemes – applications and motivation

Nonlinear pricing can be broadly classified in two categories – increasing block and decreasing block. In an increasing block pricing scheme, the marginal (per-unit) prices increase with quantity, whereas in a decreasing block scheme the marginal prices decrease with quantity. Figure 16.1 shows a few examples of increasing and decreasing block tariffs.

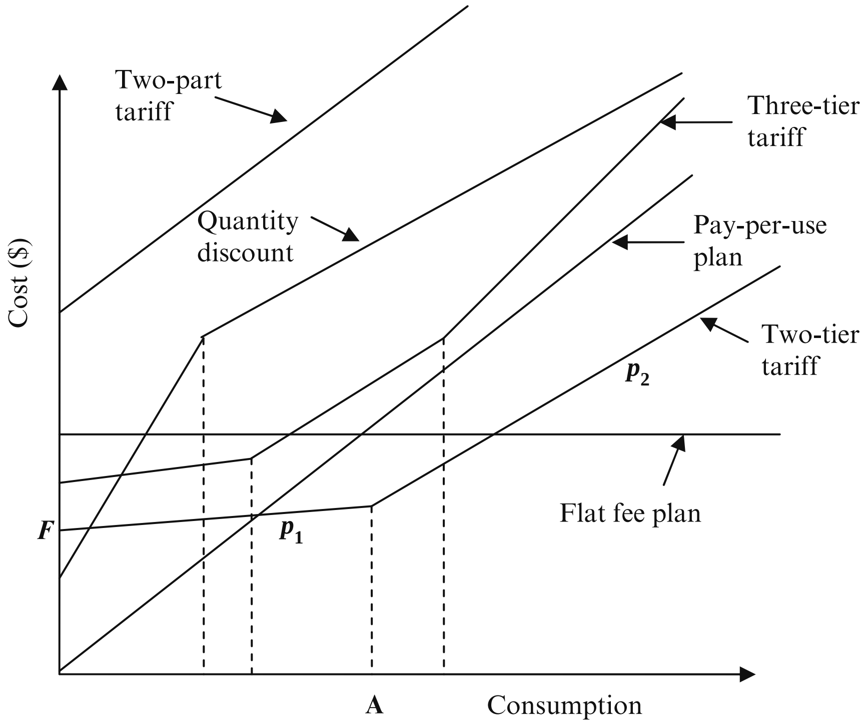
An increasing block tariff promotes conservation by penalizing excess consumption of units. A recent application of a multi-tier increasing block tariff for conservation is the electricity tariff in California. After the financial crisis in 2001, the California Public Utilities Commission imposed a new five-tier increasing block structure (see Reiss and White, 2005, p. 875 for more details). The new pricing scheme was implemented to encourage energy conservation. It was also expected to raise supplementary revenue for the state. Some evidence suggests that there was indeed a significant reduction in electricity consumption in 2001 as compared to the year before (Goldman et al., 2002).

A typical example of a decreasing block tariff is a quantity discount. For instance, Table 16.1 shows the rates that the *New York Times* charges in various categories (NYTimes Advertising Rates, 2008). Note that the rates decrease as the frequency of advertisement increases. This is essentially a mechanism for price discrimination – the advertisers who will commit to placing ads several times a year will get a cheaper rate than those customers who place only a one-time ad.

2.1 Reasons for nonlinear pricing

There are several reasons for firms to adopt a nonlinear pricing scheme. Here we discuss a few of the salient ones. See Wilson (1993) for a more detailed discussion.

1. *Price discrimination* Heterogeneity among customers is the primary reason to implement a nonlinear pricing scheme. This pricing structure can be thought of as a menu of quantities and corresponding charges. Each customer is expected to self-select the quantity–charge combination that is most appealing to him. As there is demand heterogeneity among customers, customers buy their ideal total quantity



Note: In the figure, the intercept on the vertical axis is the fixed fee associated with a pricing scheme while the slopes are the per-unit (marginal) prices. For the two-tier tariff, F refers to the access fee, p_1 and p_2 are per-unit (marginal) prices and A is the kink point when the per-unit price changes from p_1 to p_2.

Figure 16.1 Examples of nonlinear pricing schemes

Table 16.1 Advertising rates in the New York Times *for different categories*

Frequency (times/year)	Line rates ($)		
	Computer Services	Healthy Living	Home/Garden Guide
13	37.00	38.00	37.25
26	36.50	35.25	36.75
52	34.75	34.75	36.25

Source: *New York Times* website. See http://www.nytadvertising.com/was/ATWWeb/ProcessorAction.do.

based on how the per-unit rates vary with each incremental unit. Table 16.2 shows the wireless service plans offered by Verizon in the Philadelphia region. Note that these plans are an example of a two-tier (or three-part) tariff scheme.

Table 16.2 shows that there is significant variation in the number of free minutes among plans and thus can appeal to a wide customer base. In addition, plans are designed to offer a quantity discount to heavy users.

Table 16.2 Verizon wireless plans within Philadelphia, PA

Plans	Monthly access fee ($)	Overage rate ($/min)	Free minutes per month
1	39.99	0.45	450
2	59.99	0.40	900
3	79.99	0.35	1350
4	99.99	0.25	2000
5	149.99	0.25	4000
6	199.99	0.20	6000

Source: Verizon wireless website. See www.verizonwireless.com.

2. *Cost considerations* Decreasing block pricing schemes such as quantity discounts offer incentives for customers to stockpile and transfer the inventory of units from the firm to the customer. If the inventory cost for a firm is high, then such discounts offer a way of reducing its costs. Wilson (1993, pp. 15–16) gives an example from the electric utilities industry. In that industry, customers purchasing large quantities of power have a high utilization as well. A quantity discount acknowledges the lower cost of idle capacity for such customers.

   The pricing scheme within the package delivery industry provides another illustration of where the pricing scheme reflects cost considerations. Federal Express charges different rates depending on the weight of package and speed of delivery. Figure 16.2 shows the shipping charges for delivering a package from San Francisco to New York. These shipping charges increase with the weight of the package and the speed of delivery.
3. *Competitive pressures* Competitive pressures lead firms to use innovative nonlinear pricing schemes to entice customers. For instance, frequent flier miles began with each airline trying to acquire and retain business customers. Similarly, in the package delivery industry, many competitors of Federal Express such as UPS offer competitive nonlinear pricing schemes to draw customers. Figure 16.3 shows the package delivery charges from UPS for the same route (i.e. from San Francisco to New York).

A comparison of the UPS and Federal Express rates shows that they are similar, although the latter's prices are marginally lower. It is interesting to note that Federal Express also offers more alternatives – this can help customers to discriminate between companies even more. This suggests that the optimal design of a portfolio of nonlinear pricing plans involves the choice of number of plans as well as the pricing scheme for each plan.

3. Managerial decisions

The following example from long-distance telecommunications will provide a concrete context for the relevant decisions that a manager needs to make to set up a nonlinear pricing scheme.

Long-distance service providers typically price calling plans using a combination of

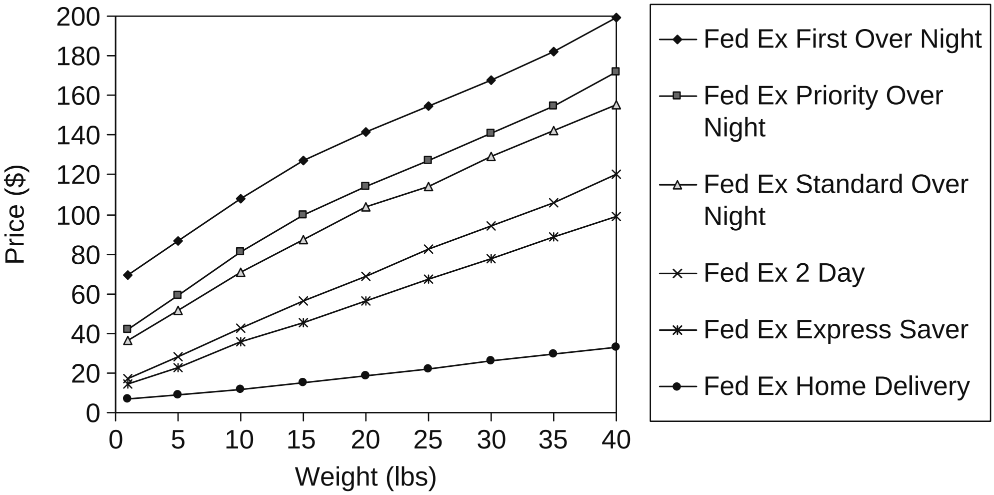


Figure 16.2 Federal Express delivery charges for shipping a package from San Francisco to New York

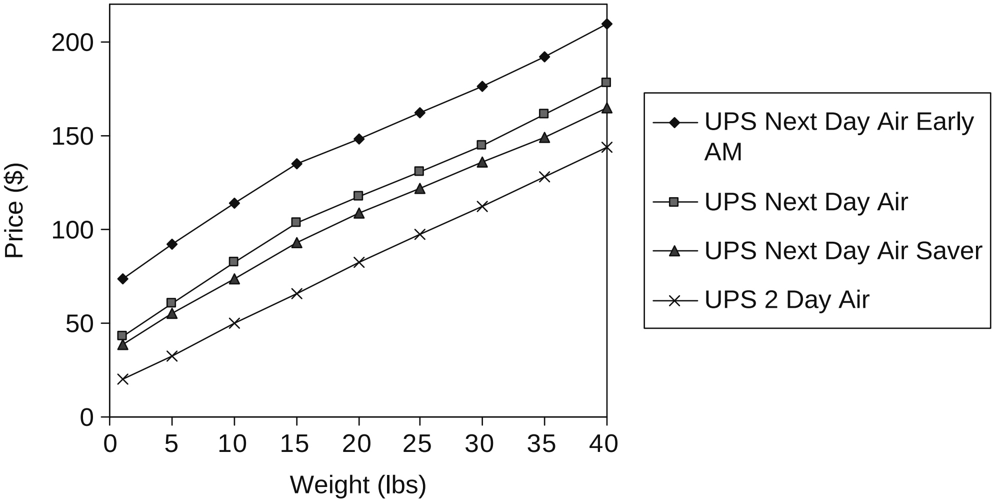


Figure 16.3 UPS delivery charges for shipping a package from San Francisco to New York

fixed fees (for access to the service) and per-minute price for each minute of a long-distance call. For instance, within New York State, Verizon offers several different callings plans. Table 6.3 illustrates these long-distance calling plans.

The table shows that there is some variation among the offered plans. For instance, the Timeless Plan has a fixed fee of $2.00 per month and a 10 c/minute rate for any consumption of long-distance minutes. This type of plan is termed a ‘two-part tariff’, with

Table 16.3 Verizon long-distance plans for New York State

Plan	Type of pricing plan	Monthly fee ($)	Detail of per-minute pricing
Timeless Plan	Two-part tariff	2.00	State-to-state and in-state calls: 10c/minute
E-Values	Two-part tariff	2.50	State-to-state and in-state calls: 10c/minute weekdays 7c/minute weekends
TalkTime 30	Three-part tariff	5.00	First 30 minutes free. Unused minutes do not carry over. State-to-state and in-state calls: 10c/minute after 30 minutes.
Verizon Five Cents Package Plan	Two-part tariff	6.00	State-to-state calls: 5c/minute In-state calls: 7c /minute
Verizon Freedom Value	Flat fee plan	34.99–39.99	Free

Source: Verizon website. See http://www22.verizon.com/Residential/Phone/Long+Distance/Long+Distance.htm.

the access fee and the per-minute price forming the two parts. Both Verizon Five Cents and E-Values have a similar structure but charge different prices for in-state and state-to-state calls. The remaining two plans (TalkTime 30 and Verizon Freedom Value) have a slightly different structure.

The Verizon Freedom Value Plan has an access fee ($34.99–$39.99) and any usage of long-distance minutes is free. Such type of plan is termed a ‘flat fee’ plan. Finally, the TalkTime 30 has three distinct components – an access fee ($5.00), per-minute rate (10 c/minute) and free minutes (30 minutes). Such a tariff is termed a ‘three-part tariff’. Another popular term for this pricing scheme is a ‘two-tier increasing block’ tariff. Here, the term two-tier refers to the fact that there are two consumption regions based on different per-minute prices – region 1, when the consumption is less than 30 minutes, has a zero per-minute price and region 2, when the consumption is greater than 30 minutes, has a per-minute price of 10 c/minute. The term ‘increasing block’ signifies that the per-minute price in region 2 (10 c/minute) is greater than the per-minute price in region 1 (0 c/minute). Readers can immediately see that a two-tier increasing block tariff can be extended to a pricing scheme that has multiple tiers, which can be either increasing or decreasing block.

This example shows that nonlinear pricing schemes appear in many different forms – at one extreme, there is the special case of a flat fee plan and, on the other, there are multi-tier tariffs. Such a wide spectrum of plans can enable Verizon to appeal to different types of customers. When the pricing scheme involves a flat fee or in case of a two-part tariff, a relatively higher monthly access fee combined with a lower per-minute charge, heavy users are more likely to sign up for that plan. In contrast, light users will prefer the pricing scheme that has a relatively lower monthly access fee but a higher per-minute charge. This example also highlights the key managerial questions that have to be answered prior to designing a nonlinear pricing scheme. We show these decisions in Figure 16.4. There are three broad sets of decisions:

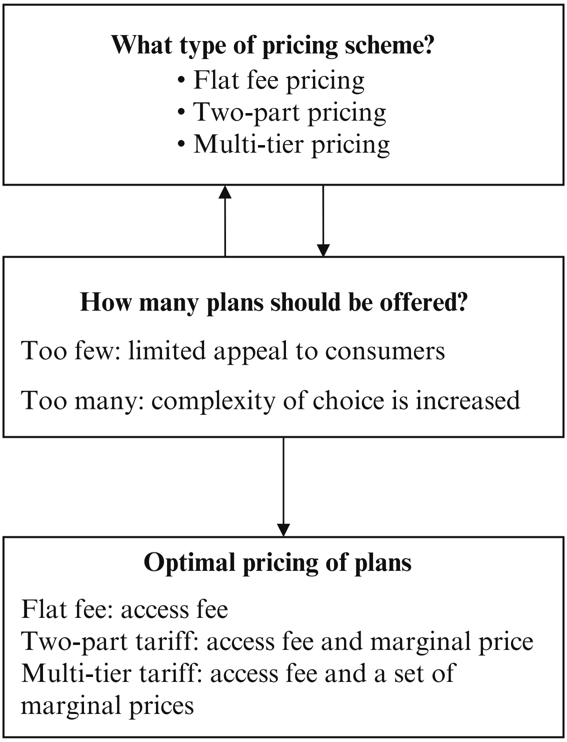

What type of pricing scheme?
• Flat fee pricing
• Two-part pricing
• Multi-tier pricing
How many plans should be offered?
Too few: limited appeal to consumers
Too many: complexity of choice is increased
Optimal pricing of plans
Flat fee: access fee
Two-part tariff: access fee and marginal price
Multi-tier tariff: access fee and a set of marginal prices

Figure 16.4 Managerial questions for implementing a nonlinear pricing scheme

1. *Type of pricing schemes* A typical portfolio of plans can have a flat fee, two-part tariff and even a few multi-tier tariffs. Much analytical work has investigated the optimality of two-part tariffs (Schmalensee, 1981; Stole, 1995; Armstrong and Vickers, 2001; Rochet and Stole, 2002). Are such two-part tariffs optimal in every circumstance or does the presence of competition and customer heterogeneity affect the optimality of a pricing scheme? Similar questions can be asked about multi-part tariffs.
2. *Number of plans* One of the primary motivations of nonlinear pricing is consumer heterogeneity, and thus offering too few plans limits its appeal to a wide range of customers. At the same time, past research suggests that increasing the number of plans might not be the answer either (Iyengar and Lepper, 2000; Iyengar et al., 2004). This line of work suggests that consumers are less motivated to make a decision if there are too many alternatives. The optimal number of plans, which would differ from one context to another, will then emerge from modeling the tradeoff between a firm's desire to offer many alternatives to appeal to the heterogeneous customer base and consumers' motivation to process all the information. In addition, as Figure 16.4 shows, the two decisions, i.e. the number of plans and type of pricing scheme for each plan, are interlinked.
3. *Optimal pricing of plans* Given a set of plans, a firm has to choose the access fees and marginal prices for each of these plans. These decisions have to consider the impact

of pricing structure on consumers' choice, consumption and retention. The presence of competition (see the earlier example of Fed Ex© and UPS©) can further complicate the situation.

Next, we discuss an example that shows how a firm designed a nonlinear pricing scheme.

3.1 Illustrative Example Deutsche Bahn AG

We discuss how Deutsche Bahn AG, the German railroad corporation, implemented a two-part tariff pricing scheme and also highlight the type of data collection and analysis required for designing such a scheme. This example is adapted from Dolan and Simon (1996, p. 164), where it is discussed in much greater detail.

Duetsche Bahn AG faced stiff competition from the automobile industry. It charged DM 0.36 per kilometer for first-class rail travel and DM 0.24 per kilometer for second-class travel. Compared to these prices, the typical gasoline price in Germany was about DM 0.15 per kilometer. Thus it was cheaper for everyone to drive and indeed most people did perceive the prices for rail travel to be too high. In addition, the company also did not price-discriminate in any other way among its customers. For instance, an obvious price segmentation strategy is based on frequency of travel, with heavy and light users being charged at a different rate. It is the possibility of implementing such usage-based price discrimination that led to the concept of BahnCard – a card that would have an annual fee and, once purchased, would lead to discounted trips. Such a pricing scheme is a two-part tariff as there is a fixed fee for access to the card and then a per-kilometer charge for any travel. Further, the two-part tariff scheme of the card would be designed such that it can be a viable alternative to attract people away from just driving to their destination. Intuitively, it would be the heavy users who will be drawn towards such a card.

On route to designing the pricing plan, the management of the railroad corporation struggled with several key questions:

(a) What percentage discount over the regular per-kilometer rate should be granted to BahnCard buyers?
(b) What should be the price of the BahnCard?
(c) How should the price be varied by class and special groups such as elderly and students?

The answers to these questions were critical to optimally designing the pricing plan and required extensive data collection from customers and potential customers of the railroad system. This data collection, in the form of responses to a conjoint design, measured the willingness to pay for varying levels of discounts. In addition, a model was developed to simulate the effects of the different pricing structures on customer segments and thereafter to estimate optimal pricing. This model took into account various tradeoffs, such as that a low price for the card may sell a high volume but the overall revenue may be negative as otherwise the full-paying heavy usage segment will pay a lower price. On the other hand, a high price for access to the card will deter many potential customers and even current customers might not increase their usage.

The analysis resulted in the set of optimal prices for both the fixed fee (access to the card) and the marginal price (percentage discount per kilometer) of the two-part tariff. The discount was set at 50 percent, i.e. the per-kilometer rate for first-class travel was DM 0.18 and for second-class travel, at DM 0.12. The fixed fee for the BahnCard for first- (second-) class travel was determined to be DM 440 (220). Finally, for elderly and the students, the card was offered at half the regular price.

We can analyze the attractiveness of this pricing scheme from the viewpoint of a second-class traveler. If the customer purchases a BahnCard, then he pays an initial fee of DM 220 and receives a rate of DM 0.12 per kilometer. Thus, for the first 100 kilometers, the customer pays a total of DM 232 (= DM 220 + 0.12*100). This translates to a rate of DM 2.32 per kilometer. If the customer did not purchase a BahnCard, he would be charged at the uniform rate of DM 0.24 per kilometer. At this rate, for the first 100 kilometers, he would pay only DM 24. The break-even point between paying the uniform rate and buying the BahnCard, and getting the discount rate occurs at around 1833 kilometers. If the customer is going to travel more than 1833 km annually, then it would be cheaper for him to purchase the BahnCard. Next, we compare the cost for a BahnCard customer with his cost for driving to his destination. As mentioned before, the typical gasoline charge was about DM 0.15 per kilometer. In this case, if the customer does not buy the BahnCard, then it would never be economical to travel by train. However, after purchasing the BahnCard, he receives a discounted per-kilometer rate that is lower than the per-kilometer rate for driving. The break-even point between driving and train travel occurs around 7333 km. If the customer is going to travel more than 7333 km annually, then it will cheaper for him to purchase the BahnCard.

Since its introduction in 1993, BahnCard has been a spectacular success. In 2004, there were about 3.2 million BahnCards sold, giving Deutsche Bahn AG an overall revenue of $450 million.

4. Theoretical research

Analytical work has focused on the issue of optimality of certain nonlinear pricing schemes under different market conditions such as monopoly and oligopoly. We begin with some broad findings applicable in monopoly settings.

4.1 Monopoly

In a classic paper, Oi (1971) addressed the following question: as an owner of Disneyland, should you charge a high entry (fixed) fee and give the individual rides for free or should you let people come in for free but charge a high price per ride (marginal price)? These two alternatives represent two extremes: either charge a flat fee for entry or a per-ride rate. Oi considered the different roles played by the entry fee and price per ride. He noted that if the monopolist desired to have all consumers in the marketplace be interested in its product, then the entry fee has to be equal to the smallest of consumer surpluses. Next, as the marginal price and entry fee together determine the demand and the overall profit, there is an implicit relationship between the two prices. He showed that a two-part tariff (as opposed to a flat fee or a per-ride rate) will allow a monopolist to be both efficient in allocation and profit maximizing. The allocation efficiency comes from setting the usage price close to the marginal cost and the profit maximization occurs by using the access (or fixed) fee to extract all or most consumer surplus. In addition, the resulting pricing scheme can be such that a few consumers might be left out of the market (i.e. the entry fee is higher than the minimum of consumer surplus). This reduction in market coverage is compensated by a lower per-ride fee and the subsequent increase in demand for rides from the rest of the market.

In later work, Schmalensee (1981) and Varian (1985) have extended this analysis for situations where the monopolist can price-discriminate and investigated how it changes the welfare implications. Welfare change is the sum of monopoly profits and consumer surplus changes. They found that there is an increase in welfare from a simple monopoly to a price-discriminating monopoly only if the total quantity produced increases. In another extension, Rochet and Stole (2002) showed that even with random participation constraints, the optimal nonlinear pricing scheme takes the form of a two-part tariff.

Recent work has investigated the conditions that can alter the optimal combination of the fixed fee and marginal price in a two-part tariff. Essegaier et al. (2002) consider the dual roles of capacity constraints and usage heterogeneity in the customer base for optimal pricing of access services (e.g. services such as AOL, sports clubs, resorts and cable TV services). They make the following modeling assumptions: there are two consumer segments in the market – heavy users, who account for a fraction α of the market and use d_h units of capacity, and the rest $(1-\alpha)$ are light users who use d_l $(d_l < d_h)$ units of capacity. These usage rates are assumed to be independent of price. Thus the maximum usage rate (assuming the number of customers in the market is normalized to 1) is given by $\bar{d} = \alpha d_h + (1-\alpha)d_l$. This is the maximum capacity that is required to service the entire market. For any given fee (f) and usage price (p), light users pay $P_l = f + pd_l$ and heavy users pay $P_h = f + pd_h$. In addition, they model customer heterogeneity in preference by using the Hotelling line – a consumer who is located at x $(0 \le x \le 1)$ has a linear transportation cost of tx to access the monopolist's service, where t is the unit transportation cost. In addition, V is the reservation price for the service (which is assumed to be the same for the two segments).

With these assumptions, they show that in the case of no capacity constraints, a monopolist will charge a flat fee such that it can cover the entire market. This flat fee price is $f = V - t$. The more interesting case arises when there are capacity constraints. The following constrained maximization problem captures the managerial decision:

$$\begin{aligned}&\underset{(f,p)}{\text{Max}}\ (1-\alpha)x_l(f+pd_l) + \alpha x_h(f+pd_h),\\ &\text{subject to } 0 \le x_1 \le 1\,, 0 \le x_h \le 1, \qquad (16.1)\\ &\text{and } (1-\alpha)x_ld_l + \alpha x_hd_h \le K.\end{aligned}$$

Here, K is the capacity of the provider which satisfies, $0 \le K \le \bar{d}$ and x_l is $(V - f - - pd_l)/t$, which is the location of marginal light users who are just indifferent between buying and not buying. Similarly, x_h is $(V - f - pd_h)/t$, which is the location of heavy users who are just indifferent between buying and not buying. The above maximization problem can be used to calculate what the optimal f and p should be as the capacity K changes. Essegaier et al. perform such an analysis and find that the two pricing components (f, p) should be negatively correlated. The flat fee is an effective way of extracting surplus from light users whereas the heavy users are more sensitive to the usage rate.

Thus, when customers have different usage rates, the pricing policy determines the customer mix that will be present and how much of the constrained capacity will be used. See Oren et al. (1985) and Scotchmer (1985) for other research that relates nonlinear pricing with capacity constraints.

An important question is whether firms should have a fixed fee and other nonlinear pricing plans together in their portfolio of offered plans. Sundararajan (2004) offers some guidelines in this regard. He analyzed a scenario where a firm associated with information goods offered both a fixed fee and a usage-based pricing plan under incomplete information. He found that if there are transaction costs associated with administering the usage-based pricing scheme, then offering a fixed fee pricing scheme (in addition to the usage-based scheme) is always profit improving. In fact, there may be situations (such as an information market in its early stages with a high concentration of low-usage customers) wherein a pure fixed fee pricing is optimal. What about the optimality of other types of nonlinear pricing schedules within a monopolistic setting? In a recent work, Masuda and Whang (2006) show that a portfolio comprising special forms of three-part tariff plans wherein, upon payment of a fixed fee, consumers receive certain units of the service for free and then are charged on a per-unit rate delivers as good a performance as any other nonlinear pricing schedule. Such special forms of three-part tariff are commonly used in the wireless telecommunications industry.

The examples described so far have considered a firm selling only a single product. What happens if the firm sells multiple products? Is a two-part tariff still optimal under some conditions? Armstrong (1999) attacked such a problem with a model that assumed consumers had multiple latent preference parameters, which might or might not be correlated across the products. He finds that if the preference parameters are independently distributed across products, the almost optimal tariff is a two-part tariff. If, however, there is a correlation in the preferences across products, the almost optimal tariff can be implemented as a menu of two-part tariffs. Thus a correlation of consumers' preferences induces a change in the overall optimal pricing scheme. See other work such as Mirman and Sibley (1980) and Wilson (1991) for other examples of optimal multiproduct pricing.

In this section, we have described only a small fraction of the enormous amount of research that has been done in monopoly settings. See Wilson (1993) for a more detailed discussion of such work.

4.2 Oligopoly

For oligopoly settings, researchers have tried to ascertain whether an increase in competition changes the structure of offered nonlinear pricing schemes. The typical modeling framework in such settings has both vertical and horizontal differentiation – the horizontal component captures the preferences of consumers across competitors while the vertical component captures differences in quality (Stole, 1995; Villas-Boas and Schmidt-Bohr, 1999; Armstrong and Vickers, 2001; Ellison, 2005). Stole (1995) showed that as competition increases, the quality distortion (i.e. the classic result that a monopolist will distort the quality level of its offered products to extract higher profits) decreases. Other work (Rochet and Stole, 2002; Armstrong and Vickers, 2001) have also found a similar result. In addition, both Rochet and Stole and Armstrong and Vickers show that, with some simplifying conditions such as full market coverage, the nearly optimal pricing scheme is again a two-part tariff scheme. One salient aspect of research in oligopoly settings is the rapid increase in mathematical complexity, which constrains researchers from obtaining simple closed-form solutions.

While the two-part tariff scheme can be nearly optimal under many conditions, several firms use more complex pricing schemes. Are such schemes optimal under any circumstance? The recent work of Jensen (2006) provides some direction, albeit in a much simpler duopoly setting. Jensen shows that implementation of simple two-part tariffs may not be a feasible strategy as the optimal nonlinear tariff exhibits a convexity for lower quantities. She shows that an optimal outcome can be implemented if firms use a tariff with inclusive consumption, i.e. a two-tier tariff where consumption on the first tier is free. This is exactly the type of pricing scheme used in wireless services. Such a finding clearly points to some future research that can investigate the implementation of other, more complex, pricing schemes.

5. Empirical research

While theoretical work has addressed the optimality of nonlinear pricing schemes under different conditions, the other two issues – the number of plans and the determination of optimal access fee and marginal prices – are empirically driven (see Section 3). Some researchers have begun to address these latter two questions and we describe such work in this section. To a large extent, however, empirical researchers have been concerned with several critical intermediate steps in modeling demand under nonlinear pricing schemes. Table 16.4 shows a summary of various studies in chronological order. In the table, we also indicate the key issue that a study considered and its main findings. Here we discuss a few of these studies in more detail within the broader framework of key issues.

5.1 Simultaneity of price and consumption

Services typically charge based on some form of a multi-part tariff. Such multi-part pricing induces a two-way dependence of price and consumption – the price influences consumption while the level of consumption depends on the prices charged by a provider. This two-way dependence occurs in many contexts. Examples are utilities such as electricity and water supply (Taylor, 1975; Nordin, 1976; Hausman et al., 1979; Billings and Agthe, 1980; Hewitt and Hanemann, 1995; Reiss and White, 2005), landline telephone services (Park et al., 1983; Train et al., 1987; Kling and Van der Ploeg, 1990; Kridel et al., 1993; Miravete, 2002; Danaher, 2002; Narayanan et al., 2007) and cellular phone (Miravete and Roller, 2004; Miravete, 2007; Iyengar et al., 2007a).

Research on addressing this simultaneity has its roots in labor economics (Hall, 1973; Rosen, 1976; Burtless and Hausman, 1978; Wales and Woodland, 1979; Hausman, 1985; Blomquist, 1996; Moffitt, 1990; Van Soest, 1995; Van Soest et al., 2002). Labor economists are concerned with the prediction of changes in the labor supply when a new tax structure is imposed on people. The early work on labor supply (Hall, 1973) used an ordinary least squares (OLS) approach with hours of work as a dependent variable and the applicable federal income tax rate as an explanatory variable. While OLS is attractive because of its simplicity, it is clearly not a viable option for this application because of the endogeneity of tax rate. When such endogeneity is present, researchers have typically used an instrumental variables (IV) approach (Hausman and Wise, 1976; Hausman et al., 1979). The biggest issue with the IV approach is that in practice it is often difficult to find

Table 16.4 Summary of empirical research on nonlinear pricing

Authors	Objective	Data	Model	Key Findings
Hall (1973)	Establish a relationship between the nonlinearity in tax schedules and number of hours that people work	Survey of Economic Opportunity (SEO) – hourly wage rates, personal characteristics of respondents	Ordinary least squares (OLS)	Effects of demographics, such as age, gender, race on number of work hours
Taylor (1975)	Survey of econometric literature on demand for electricity	Description of several studies	–	Appropriate modeling of demand under nonlinear pricing schemes involves an inclusion of marginal and average prices
Burtless and Hausman (1978)	Demand model with a nonlinear budget set arising from changes in tax rates	Consumer-level work hours, tax rates, wages, non-wage compensation and personal characteristics	Utility-based economic model that allows for maximization under a nonlinear budget	Much lower wage elasticities from tax rate changes than previous reduced-form approaches
Hausman et al. (1979)	Forecast consumer-level electricity usage	Household electricity consumption data	Economic demand model using budget constraints	Electricity usage under both time-of-day and declining block rate are predicted well
Park et al. (1983)	Calculate price elasticity for local telephone calls	Number of calls and minutes of local calls	Heteroskedastic and autocorrelated regression	Very small (about 0.1 or less) price elasticities for both calls and minutes
Dubin and McFadden (1984)	Analysis of residential electricity appliance holdings and consumption	Household-level applicance and electricity consumption data	Discrete/continuous demand model	Estimating demand using OLS without modeling appliance choice leads to an overestimation of the elasticity of demand
Train et al. (1987)	Forecast plan choice and demand for local telephone service	Number and average duration of local calls for a sample of customers	Nested logit	Households respond to a price change by changing their calling patterns more than their calling plans

Table 16.4 (continued)

Authors	Objective	Data	Model	Key Findings
Nunes (2000)	How consumers anticipate product usage during purchase deliberation and how it affects the choice of flat fee versus measured plan	Experimental data/ health club visitation records	NBD model	Consumers employ a break-even model – if expected usage is greater than the break-even, they choose a flat fee plan, otherwise they choose a measured plan
Danaher (2002)	Empirically derive a revenue-maximizing strategy for a two-part tariff with customer defection	Monthly usage and retention data	Regression model which also accounts for attrition	Access and usage prices have different relative effects on demand and attrition
Miravete (2002)	Analyze choice of plan and demand for local telephone service with consumer uncertainty	Monthly usage data on number and duration of calls	Demand model with a distinction between *ex ante* and *ex post* consumer types	There is evidence for *ex ante* and *ex post* asymmetry of information. This has implications for optimal design of plans
Lambrecht and Skiera (2007)	Understand the antecedents of flat fee and pay-per-use bias	Customer transaction data and survey data from an Internet service provider	Binomial logit model	Underestimation of usage is a major reason for pay-per-use bias. In addition, flat fee bias does not significantly increase customer defection
Allenby et al. (2004)	Model consumer choice for packaged goods and account for discrete quantities and quantity discounts	Scanner panel data	Discrete/continuous demand model	Model provides a valid measure of utility to assess changes in consumer welfare with assortment changes

Lambrecht et al. (2007)	Analyze the effect of consumer uncertainty on choice among three-part tariff plans	Customer transaction data from an Internet service provider	Discrete/continuous model	Demand uncertainty decreases consumer surplus and increases provider revenue. Access fee is the main driver of tariff choice
Iyengar et al. (2007a)	Incorporate consumer uncertainty on both quality and usage and model choice among three-part tariffs	Consumer-level monthly usage from a wireless service provider	Discrete/continuous model	Consumer learning can be a win–win for both consumers and provider. There is a 35% increase in customer lifetime value with learning than without
Narayanan et al. (2007)	Incorporate consumers' usage uncertainty and model choice between flat fee and measured plan	Consumer-level monthly usage for local telephone service	Discrete/continuous model	Consumers learn about their usage rapidly when they are on a measured plan and learn very slowly when on a fixed plan
Iyengar et al. (2007b)	Incorporate consumer uncertainty in expected usage and model choice among three-part tariffs	Choice-based conjoint	Utility-based economic model, with an underlying latent usage, in the presence of a nonlinear budget	Utility-based discrete choice model with inferred usage predicts significantly better than a traditional conjoint model

proper instruments and justify their use. Given the deficiency of the IV approach, other methods based on the selectivity bias literature (Heckman, 1979) have been developed (Heckman and MaCurdy, 1981; Reiss and White, 2005).

In a seminal paper, Burtless and Hausman (1978) suggested a technique, which combined theory with econometrics, to address this problem. In a pricing context, an application of this technique involves maximizing a specified utility function subject to the constraints imposed by the pricing scheme. With suitable assumptions on the utility function (quasi-concavity) and under increasing block pricing schemes, such maximization can yield a unique optimal solution. The actual consumption is then modeled as a deviation from this optimal solution. Thus it is not the observed consumption that results from an optimization but rather depends on the optimal consumption, which in turn is influenced by the pricing scheme. Burtless and Hausman termed the deviation between the optimal consumption and actual consumption as the 'optimization error'. A detailed explanation of all past research can be found elsewhere (Hausman, 1985; Moffitt, 1990).

Note that uniqueness of the optimal solution requires the presence of an increasing block pricing scheme. This is because these schemes translate to convex constraints and the maximization of a quasi-concave utility function subject to such constraints has a unique optimum (Hausman, 1985). This uniqueness is not ensured if the pricing scheme is decreasing block (e.g. a quantity discount). In such a case, multiple optima might exist. Thus the utility function will have to be directly evaluated to calculate the overall optimum. See Allenby et al. (2004) for such analysis where they evaluate the effect of quantity discounts on overall demand.

5.2 Endogenous choice and consumption decisions

In many service settings, consumers typically choose from a portfolio of nonlinear pricing tariffs. Thus they not only consume under a nonlinear pricing tariff but also choose that tariff (Dubin and McFadden, 1984; Train et al., 1987; Narayanan et al., 2007). For example, in a wireless service context, consumers choose a calling plan and then decide how many minutes to consume under that chosen plan. Such a process suggests two salient points. One, there is a temporal difference between the two decisions. Two, the choice and consumption decisions are endogenous (Hanemann, 1984; Chiang, 1991; Chintagunta, 1993).

Early research had modeled these two decisions as simultaneous. For instance, Train et al. (1987) used a nested logit model to captures households' choices among local telephone options and the relationship between the choice and the number and average duration of local calls. Here, they assume that choice and usage are simultaneous decisions. Similarly, Dubin and McFadden (1984) model the demand of consumer durables and the use of electricity. Here too, they assume that the two decisions are contemporaneous.

More recent research has focused on how to capture the intertemporal nature of the choice and consumption decisions. For instance, Miravete (2002) investigates how consumers choose between a flat fee and a measured tariff for local telephone service and then consume under the chosen tariff. He models the time lag and any uncertainty in consumers that arises by distinguishing between *ex ante* and *ex post* consumer types. A consumer knows only her *ex ante* type when she makes a choice among the different plans. After making the choice, she receives a shock, which alters her *ex ante* type to the *ex post* type. It is the *ex post* type that in turn influences the subsequent usage decision. This difference between the *ex ante* type and the *ex post* type captures any change in the information set of consumers due to the sequential nature of the decisions. Specifically, Miravete assumes the following relationship between the *ex ante* and the *ex post* type:

$$\theta = \theta_1\theta_2, \tag{16.2}$$

where θ is the consumer's *ex post* type, θ_1 is the *ex ante* type (known to the consumer at the tariff choice stage) and θ_2 is the shock. Thus the distribution of the *ex post* type is composed of the distribution of the *ex ante* type and the shock.

For model tractability, he makes the following distributional assumptions:

$$\theta_1 \sim Beta\left(1, \frac{1}{\lambda_1}\right). \tag{16.3}$$

and

$$\theta_2 \sim Beta\left(1 + \frac{1}{\lambda_1}, \frac{1}{\lambda} - \frac{1}{\lambda_1}\right). \tag{16.3}$$

With these assumptions, the consumer's *ex post* type has a Beta distribution as well:

$$\theta \sim Beta\left(1, \frac{1}{\lambda}\right). \tag{16.4}$$

With these distributional assumptions, these consumer types are similar to probabilities. The demand function for the telephone service is dependent on the *ex post* type and is specified as follows:

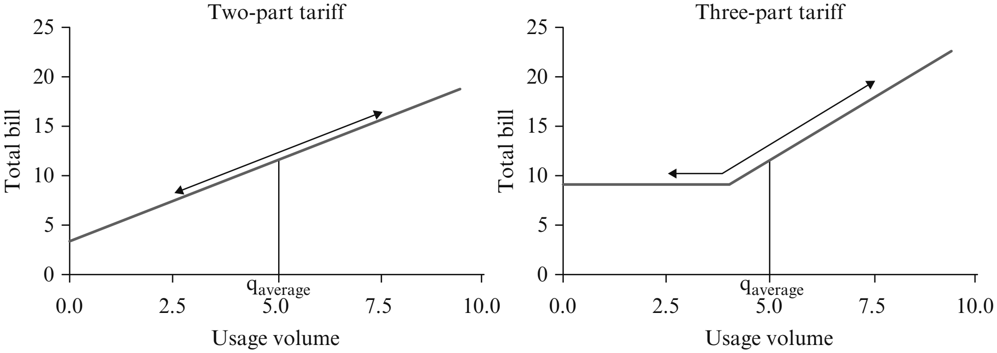
$$x(p, \theta) = \theta_0 + \theta - p, \tag{16.6}$$

where the parameter θ_0 is a parameter large enough to ensure that the demand is always positive and p is the per-minute price. This demand function, together with the distributional assumptions on the *ex post* type, then help Miravete test several hypotheses about how uncertainty plays a role in the sequential decision-making nature of the problem.

A different means for capturing this sequential nature of consumer decisions comes from extending the Burtless and Hausman model to incorporate the choice decision. The intuition is that consumers ascertain the optimal consumption under each available option, evaluate the utility of the different options with that option-specific optimal consumption and then choose the alternative that provides the highest utility. Subsequent to plan choice, consumers' actual consumption deviates from their optimal consumption due to optimization error. Thus the earlier decision of plan choice is influenced by optimal consumption and not the actual consumption. See Section 5.4 for an illustration of this modeling framework.

5.3 Usage uncertainty and learning

The sequential nature of decisions indicates that the information set of consumers could differ from when they are making a choice among different alternatives to when they



Source: Lambrecht et al. (2007), p. 11.

Figure 16.5 Symmetric deviations of usage under a two-part tariff and three-part tariff scheme

are consuming under a chosen plan. Further, if they have the opportunity to engage in repeated choice and usage decisions, their information set might alter over time as they 'learn' and resolve the uncertainty about their own usage patterns.

Lambrecht et al. (2007) use a simple example to show how such usage uncertainty can affect consumer choice. They consider symmetric distributions of usage under a two-part tariff and a three-part tariff. Figure 16.5 shows these deviations. The figure shows that usage deviations under a two-part tariff leave the expected bill unaffected, i.e. the expected bill is the same with low or high levels of uncertainty in usage. This is not so under a three-part tariff – under such pricing schemes, the higher the uncertainty in usage given the same level of mean usage, the higher is the overall bill. This clearly suggests that, under a three-part tariff and more complex multi-part tariffs, consumers' usage expectation can influence their choice of service plan.

Several researchers have found evidence to support this hypothesis (Nunes, 2000; Lemon et al., 2002; Lambrecht and Skiera, 2006). For instance, Nunes (2000) explores the cognitive process of how people anticipate service usage and how they integrate their expectations of usage to choose between a flat fee plan and a measured (pay-per-use) plan. He proposes that consumers calculate a break-even number and then see whether the break-even implies a choice of flat fee plan or a measured plan. Similarly, Lemon et al. (2002) show that consumer expectations of future usage influence their decision to stay with or leave a service provider.

Other researchers have quantitatively investigated consumers' usage uncertainty and learning using sophisticated models that incorporate Bayesian updating. For instance, Goettler and Clay (2007) capture consumer uncertainty and learning about the quality of an online retailer. Similarly, Narayanan et al. (2007) analyze data from an experiment conducted by South Central Bell. In this experiment, people had a choice between a flat rate pricing scheme and a two-part tariff. They find that consumer learning is very rapid when consumers are on the two-part tariff scheme but is very low while on the flat fee plan. Specifically, they make the following modeling assumption for the conditional indirect utility function for consumer *i*, plan *j* and time *t*:

$$V_{it}^{j} = (y_i - f^j) + \frac{\theta_{it}}{\beta}\exp(-\beta p_t^j). \tag{16.7}$$

Here, y_i is the income, θ_{it} is the consumer-specific and time-specific type (similar in spirit to the consumer type proposed by Miravete (2002)), f^j and p_t^j are the access fees and per-unit usage price and the parameter $-\beta$ is the price coefficient.

In addition, Narayanan et al. decompose the type parameter (θ_{it}) in the following manner:

$$\ln(\theta_{it}) = \alpha_i + \gamma Z_{it} + \eta_{it} + \nu_{it} \tag{16.8}$$

Here, the first component α_i is consumer specific but time invariant, the term $(\gamma Z_{it} + \eta_{it})$ captures the component observed by the consumer at the time of plan choice and finally, the shock ν_{it} is unobservable to the consumer during plan choice but is known at the time of usage decision. This framework captures the sequential nature of choice and consumption decisions. To capture learning, Narayanan et al. assume that consumers have beliefs over the parameter α_i, and these beliefs get updated as they observe their choices and the consumption signal.

Note that the above model is developed for a choice between a flat fee and a two-part tariff scheme. It is not straightforward to extend it to a setting where the pricing scheme has multiple tiers. Recently, Iyengar et al. (2007) developed a model that captured consumer learning and uncertainty within the context of more general pricing schemes. They found that consumer learning can lead to a win–win situation for both consumers and the firm – consumers leave fewer minutes on the table while the firm sees an increase in overall customer lifetime value (CLV). In particular, they estimated that there is about a 35 percent increase in CLV (about $75) in the presence of consumer learning. The key driver of this difference is the change in the retention rate with and without consumer learning.

Such quantitative models shed light on how different aspects of the pricing scheme and past choice and consumption decisions can affect consumers' information set and thereby influence their future decisions. While such work provides a direction, there are still many unresolved issues. For instance, within service settings, all models of consumer learning assume that each month's usage gives a signal to the consumer to better understand their own consumption pattern. However, there is research in a scanner data context that suggests that consumers have thresholds of insensitivity (Han et al., 2001). It is certainly plausible to assume that this might be the case within service contexts as well, i.e. perhaps only usage signals that are either above or below some threshold (which could be a function of how many free minutes are associated with the plan) have the potential to affect consumer learning. Such questions have much managerial significance given that consumer uncertainty and learning can affect their decision to defect from a service provider and thereby impact their overall lifetime value.

Thus far, we have given examples of how different researchers have addressed each of the issues associated with modeling consumer decisions under nonlinear pricing schemes. Next, we illustrate an integrated modeling framework that captures all three issues. See Iyengar et al. (2007) for more details. For this example, we use the context of wireless services.

5.4 Integrated modeling framework – example from wireless services

Consider a wireless service that has a two-tier increasing block pricing structure characterized by a fixed fee and two marginal prices. This scheme was graphically shown in Figure 16.1. Suppose *F* represents the access price for the service and the applicable marginal price is p_1 for consuming an additional unit before the kink is less than the marginal price, p_2, for consuming after the kink.

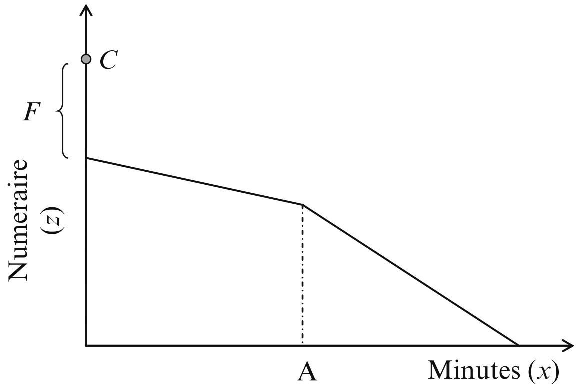
When consumers choose a wireless service, they do not make this decision in isolation from their other consumption decisions. At any point in time, they have several consumption opportunities and they allocate their income among these opportunities. This tradeoff across goods can be appropriately represented using a budget set representation. Such a budget set corresponding to an increasing two-tier pricing scheme is shown in Figure 16.6. The vertical axis in the figure corresponds to the consumption of the outside good (z) and the horizontal axis corresponds to the consumption of units of the service (x).

Figure 16.6 shows that the two-tier increasing block pricing structure of the service results in a piecewise linear budget set with a kink point (A). A consumer who subscribes to the service faces a convex budget set, and her income (*I*) is lowered by the sum of the access fee (*F*) and the variable charges for any consumed service. If, however, she does not subscribe to the service, then the entire income is used for consuming the outside good. If the marginal price of the outside good is normalized to 1 (numeraire), then the following equations represent the piecewise budget set.

$$p_1 x + z \leq I - F \quad \text{if} \quad x > 0 \quad \text{and} \quad x \leq A \tag{16.9}$$

$$p_2(x - A) + z \leq I - F - p_1 A \quad \text{if} \quad x > A \tag{16.10}$$

In the wireless communications industry, a restricted form of such a two-tier increasing block pricing scheme, where p_1 is 0, is widely used. Therefore the consumption of an addi-



Note: *F* refers to the access fee, A is the kink point (free minutes) and *C* is the total income used for consuming the outside good if the consumer does not subscribe to the plan.

Figure 16.6 A budget set representation of a two-tier increasing block pricing scheme

tional minute before the kink point is costless. Next, we specify the utility that a consumer receives when he/she uses the wireless service.

Utility function Let U_{ijt} be the direct utility function for a consumer I for consuming x_{ijt} minutes under a plan j and a quantity z_{ijt} of the numeraire commodity during period t. We specify U_{ijt} as

$$U_{ijt}(x_{ijt}, z_{ijt}) = \hat{a}_{ij} + \acute{a}_{i1}x_{ijt} + \acute{a}_{i2}z_{ijt} + \acute{a}_{i3}x^2_{ijt} + \mathring{a}_{ijt}. \qquad (16.11)$$

The terms $\hat{a}_{ij}$, $\acute{a}_{i1}$, $\acute{a}_{i2}$ and $\acute{a}_{i3}$ are individual-level parameters[1] and the random choice errors are contained in $\mathring{a}_{ijt}$. We assume that this choice error is double exponential.

The optimal consumption, x^*, which maximizes the direct utility in Equation (16.11) subject to the non-linear pricing constraints imposed by plan j, can be written as follows:

$$\begin{aligned}
&\underset{x}{\text{Max}}\ U_{ijt}(x, z(x))\\
&\text{subject to Constraint I: } p_{1j}x + z = I_i - F_j, \quad \text{if } 0 < x \le A_j,\\
&\text{Constraint II: } p_{2j}(x - A_j) + z = I_i - F_j - p_{1j}A_j, \quad \text{if } A_j < x < B. \qquad (16.12)
\end{aligned}$$

To ensure a unique solution to the above maximization problem, the utility function should be quasi-concave. This requires the Slutsky constraints – $\acute{a}_{i2} > 0$ and $\acute{a}_{i3} < 0$ on the parameters of the utility function. For a quasi-concave utility function and a convex budget set, the unique optimal solution x^* can be at an interior point (between 0 and A_j or between A_j and B) or one of the end points – 0, A_j and B. The two candidates for an interior optimal solution can be found by maximizing the utility function subject to the two linear constraints. The first-order conditions yield the following two interior candidate optima:

$$x^{candopt,I} = \frac{\acute{a}_{i2}p_{1j} - \acute{a}_{i1}}{2\acute{a}_{i3}},$$

$$x^{candopt,II} = \frac{\acute{a}_{i2}p_{2j} - \acute{a}_{i1}}{2\acute{a}_{i3}}, \qquad (16.13)$$

In the above equations, $x^{candopt,I}$ ($x^{candopt,II}$) refers to the candidate optimal consumption when the utility function is maximized with Constraint I (Constraint II).

Given the uniqueness of the solution, at most one of the two candidates will be attainable, i.e. will lie in the consumption interval where its applied constraint holds. As Constraint I holds for any positive consumption less than A_j minutes, even though $x^{candopt,I}$ can lie anywhere on the real line, it is attainable only if it lies between 0 and A_j minutes. Similarly, $x^{candopt,II}$ is attainable only if it lies between A_j minutes and B. It is, however, possible that none of two candidates for an interior solution is attainable. Then, one of end points (0, A_j or B) might be chosen. These cases are mutually exclusive and, together with any possible

[1] The term $\hat{a}_{ij}$ represents an individual and plan-specific intercept. The parameter $\acute{a}_{i1}$ represents the main effect of consumption of minutes and $\acute{a}_{i2}$ represents the effect of consuming a unit of the numeraire. The term $\acute{a}_{i3}$ captures the effect of differential marginal impact of consuming an additional minute.

interior solution, form an exhaustive solution set, i.e. $x^* \in \{0, A_j, B, x^{candopt,I}, x^{candopt,II}\}$. We denote this optimal quantity for consumer *i*, plan *j* and time *t* by x^*_{ijt}.

Let the actual demand under plan *j* for consumer *i* at time *t* be x^{act}_{ijt}, then the optimal demand is related to the actual demand in the following manner:

$$x^{act}_{ijt} = x^*_{ijt} + \eta_{ijt}. \tag{16.14}$$

Here, the demand error, η_{ijt}, is assumed to be normally distributed with a mean 0 and variance δ^2. Thus the actual demand is a function of the optimal demand, which in turn is dependent on the budget constraints imposed by the pricing scheme. Equation (16.14) can then be used to determine the likelihood of consuming a certain number of minutes under a given plan.

Note that we developed this model for a scenario where consumers were facing an increasing block pricing scheme. As discussed earlier, such a scheme results in a convex budget set, and together with a quasi-concave utility function, we obtain a unique optimal quantity. This uniqueness is not ensured if the pricing scheme is decreasing block (e.g. a quantity discount). In such a case, multiple optima might exist and the algorithm for finding the optima (see equation 16.12 and the following discussion) will not be applicable. Thus the utility function will have to be directly evaluated to calculate the overall optimum. See Allenby et al. (2004) for such analysis where they evaluate the effect of quantity discounts on overall demand.

In addition, the above example shows that the Burtless and Hausman model primarily investigated demand under a nonlinear budget set. In several service contexts, however, such a model captures only one part of consumers' decisions. For example, in the wireless service context, consumers choose a calling plan among several alternatives and then consume under the chosen plan. Next, we describe how the above model can be extended to include the choice decision.

5.4.1 Inclusion of choice decision To incorporate the choice decision within the above framework, we calculate the optimal consumption associated with every plan. Thus, for every service plan *k* ($k = 1\ldots J$), let the optimal consumption be x^*_{ikt}. Next, we determine the utility corresponding to this optimal consumption. This is the maximum utility that consumer *i* will receive if he or she chooses alternative *k*. Let the systematic component be denoted by V_{ikt}. Thus

$$U^{\max}_{ikt}(x^*_{ikt}) = V_{ikt} + \mathring{a}_{ikt}. \tag{16.15}$$

Recall that we assumed that the choice error is double exponential distributed. This assumption gives the familiar logit expression for the probability of choice:

$$P_{ijt} = \frac{e^{V_{ijt}}}{\sum_k e^{V_{ikt}}} \tag{16.16}$$

Equations (16.14) and (16.16) together give the likelihood of choosing plan *j* and consuming x^{act}_{ijt} minutes. In this model, the choice and consumption decisions are related via the optimal quantity, which in turn is determined by maximizing the utility function subject to the budget constraints. Thus both consumer decisions stem from a single utility function. In addition, the choice decision occurs before the consumption decision and is influenced by optimal consumption.

Note that so far in this framework, we have assumed that consumers are completely certain of their optimal consumption under the different plans. Next, we show a way in which such uncertainty can be incorporated within the model.

5.4.2 Consumption uncertainty If consumers have uncertainty in their consumption, then it renders the utility function stochastic. In such situations, consumers will use expected utility for making any decisions. This can be represented as follows:

$$\begin{aligned} EU_{ijt} &= E^{t}_{usage}[g(x_{ijt}, z_{ijt})] + \hat{a}_{ij} + \mathring{a}_{ijt}, \\ g(x_{ijt}, z_{ijt}) &= \acute{a}_{i1}x_{ijt} + \acute{a}_{i2}z_{ijt} + \acute{a}_{i3}x^{2}_{ijt}. \end{aligned} \tag{16.17}$$

Here, EU_{ijt} refers to the expected utility for consumer i and plan j and the term $E^{t}_{usage}[g(x_{ijt}, z_{ijt})]$ is the expectation with respect to a consumer's beliefs about his/her own usage. For each plan j we can assume an individual-specific belief distribution denoted by $f^{usage}_{ijt}(x)$. We subscript this belief distribution by time 't' to denote that it might be changing over time due to consumer learning. Different assumptions made for this belief distribution can investigate its sensitivity on the findings.

Thus, using the quantity belief distribution and the plan-specific budget constraints, the component, $E^{t}_{usage}[.]$, can be computed. The budget constraints for the plan impose a relationship between the consumed minutes (x_{ijt}) and the numeraire (z_{ijt}) as shown in equations (16.9 and 16.10). For example, if Constraint I holds, then $z_{ijt} = I_i - F_j - p_{1j}x_{ijt}$. Similarly, if Constraint II holds, then $z_{ijt} = I_i - F_j - p_{1j}A_j - p_{2j}(x_{ijt} - A_j)$. In other words, we can rewrite $g(x_{ijt}, z_{ijt})$ as a function of x_{ijt} only. Let $g(x_{ijt}, z_{ijt})$ be denoted by $h_1(x_{ijt})$ if $x_{ijt} \leq A_j$ and by $h_2(x_{ijt})$ if $x_{ijt} > A_j$. The quantity expectation is as follows:

$$E^{t}_{usage}[g(x_{ijt}, z_{ijt})] = \int_{0}^{A_j} h_1(x)f^{usage}_{ijt}(x)dx + \int_{Aj}^{\infty} h_2(x)f^{usage}_{ijt}(x)dx. \tag{16.18}$$

This expected quantity can be re-inserted in equation (16.17) to give the overall utility function. As before, if we continue to assume that the choice errors are double exponential distributed, then we can write the probability of choice for a plan with the familiar logit expression. This probability expression now would incorporate the effect of consumption uncertainty on plan choice. This completes our integrated modeling framework.

5.5 Key empirical results

Several empirical studies have focused on how consumers behave under nonlinear pricing schemes and then capture how the different components of a multi-part pricing scheme affect their behavior. Here, we summarize some key empirical results.

5.5.1 Flat fee bias A robust finding across many empirical studies is that many consumers prefer a tariff with a flat fee even though their overall expense will be lower on a pay-per-use plan (Kling and Van der Ploeg, 1990; Kridel et al., 1993; Nunes, 2000; Lambrecht and Skiera, 2006). This is referred to as the 'flat fee' bias. For instance, within the context of long-distance telephone service, Kridel et al. (1993) had found that 65 percent of consumers showed a flat fee bias. Similarly, in an application involving the use of an Internet service, Lambrecht and Skiera (2006) find that about 48 percent of consumers show a flat rate bias.

Lambrecht and Skiera (2006) also systematically consider the various causes for this bias and suggest that there are four reasons for its existence: insurance effect, taxi meter effect, convenience effect and overestimation effect. Insurance effect refers to the notion that consumers might want to choose a flat fee option as they want to 'insure' against future variation in their usage. The taxi meter effect captures the fact that many consumers can find their use of the service less enjoyable if they are paying by the minute. The term 'convenience effect' points to consumers choosing a status quo tariff to minimize any mental hassle associated with calculating the expected cost under the different available alternatives. Finally, the overestimation effect refers to the empirical finding that consumers can overestimate their demand, thereby biasing their choice towards a plan with a flat fee. In their study, Lambrecht and Skiera find that the insurance, taxi meter and overestimation effects account for the flat fee bias. Clearly, the level of consumers' usage uncertainty can moderate which of the four factors will have an influence on his/her choice decision.

5.5.2 Differential effect of access fee/marginal price A second empirical generalization is that the different components of a pricing scheme indeed have a differential impact on customer behavior. We discuss two aspects: price elasticity and the use of the multi-part tariff for discrimination.

1. *Price elasticity* Several studies across different contexts have investigated the price elasticity of different components of a multi-part pricing scheme. They have typically found price elasticity ranging from 0.1 to 1.0. Danaher (2002) describes a market experiment for a new telecommunication product (like a wireless service) in which the pricing scheme (a two-part tariff) was systematically manipulated. Consumers had to make a decision whether to continue using the product and if so, how much to use it. In that context, he found that both access fee and marginal price elasticity to be lower than 1.0. Within wireless services, Reiss and White (2007) also find that the mean price elasticity is less than one (1.00) and estimate it to be −0.44. Two studies in the context of local telephone service find very similar numbers – Park et al. (1983) and Train et al. (1987) found the price elasticity to be between 0.1 and 1.0. See Manfrim and Da Silva (2007) for a summary of estimated price elasticity across several different studies.
2. *Price discrimination* Iyengar (2007) reports that changes in access fee have a much larger impact on customer lifetime value (CLV) as compared to that from changes in marginal price. He analyzed consumers' choice among four wireless service plans and their decision to leave the service provider. Each of these plans had a three-part tariff structure – access fee, associated free minutes and a per-minute rate for any consumed minutes beyond the free minutes. Table 16.5 shows the details of the pricing scheme for the four plans. After estimating the model parameters, he then

Table 16.5 Elasticity of customer lifetime value with increase or decrease in prices

Plan	Access fee ($)	Free minutes	Per-minute rate ($/min)	Access fee		Per-minute rate	
				up	down	up	down
1	30	200	0.40	−1.18	1.08	–0.02	0.09
2	35	300	0.40	−0.09	0.09	–0.06	0.08
3	40	350	0.40	−0.48	0.25	–0.12	0.10
4	50	500	0.40	0.06	−0.09	–0.22	0.16

Source: Iyengar (2007).

performed simulation studies to capture consumers' choice and consumption decisions (which provide revenue to the service provider) and their decision to stay with or leave the provider (consumers' defection decision). He then combined the generated revenue and consumers' defection decision to determine their CLV. In addition, he calculated the elasticity of CLV with respect to both access fee and marginal price. In these simulations, he changed (either increased or decreased) the access fee and marginal prices of the four plans, one plan at a time. Table 16.5 shows the results of the simulations.

The table shows that, in general, a price decrease for a plan leads to a higher CLV than that from an equivalent price increase. A price increase for a plan results in higher average revenue per user (ARPU) but negatively affects retention. In contrast, a price decrease for a plan enhances retention but lowers the revenue. The CLV results suggest that an increase in retention is more effective for increasing the CLV than an increase in the revenue. He also finds that for all plans but Plan 4, the elasticity of CLV with respect to the access price for a plan is higher than with respect to its marginal price. Thus service providers can affect the CLV more by changing the access fee than by altering the marginal prices.

An analysis of the effects of changing the access price on the CLV shows that a decrease in the access price for Plan 1 has the highest effect. This effect on CLV can be decomposed into the effect on revenue and retention. Table 16.6 shows this decomposition.

The table shows that the primary contributor for this result is an increase in retention of the 'light users' on Plan 1. Interestingly, he finds that an increase in the access price for Plan 4 leads to a higher CLV than that arising from a price decrease. This result can be explained based on the tradeoff between the ARPU and retention. The table shows that for a change in the access price of Plan 4, the ARPU is more elastic than retention is. Hence the increase in the ARPU due to an increase in the access fee dominates the decrease in retention and thereby yields a higher CLV than that of the base case scenario. An analysis of the effects of changing the marginal price on the CLV reveals that an increase in the marginal price for Plan 4 has the highest effect. This result is due to the increase in the defection rate of 'heavy users' on Plan 4. These consumers have a high consumption of minutes and can only respond to a price increase by defecting since downgrading to lower plans is not attractive. In contrast, a decrease in the marginal price for Plan 4 generates an incentive for

Table 16.6 Elasticity of ARPU and retention with increase or decrease in prices

Plan	Elasticity of ARPU				Elasticity of retention			
	Access fee		Per-minute rate		Access fee		Per-minute rate	
	up	down	up	down	up	down	up	down
1	0.23	−0.24	0.16	−0.07	−0.68	0.58	−0.08	0.07
2	0.11	−0.12	0.02	−0.01	−0.10	0.10	−0.03	0.04
3	0.24	−0.27	0.01	−0.01	−0.34	0.24	−0.05	0.04
4	0.10	−0.09	0.01	−0.02	−0.03	0.01	−0.11	0.08

Source: Iyengar (2007).

these heavy users to stay longer with the company. These findings suggest that the different components of a multi-part pricing scheme can be effectively used for price discrimination.

Iyengar et al. (2007b) provide additional evidence in support of the differing effect of access fee and marginal prices on consumers' choice decisions. With data from a choice-based conjoint task using multi-part tariffs, they build an economics-based model to investigate how changes in the pricing scheme of plans affect its probability of choice. They find that changes in access fee affect the plan choice probability in a way that differs both qualitatively and quantitatively from those by changes in the marginal prices. Specifically, they find that above a certain threshold, an increase in marginal price of plan does not have any effect on the consumer choice decision. In contrast, any increase in access fee of a plan always reduces the probability of choice of that plan.

Iyengar et al. also address questions regarding optimal (profit-maximizing) values of access fee and marginal price for the available plans. They use individual-level parameter estimates, e.g. price sensitivity, to account for customer heterogeneity and calculate the value of access fee and marginal prices for a portfolio of plans, which would lead to maximum overall profit. Such an analysis combines economic theory with customer behavior under such a pricing structure to yield profit-maximizing values for the various components of the pricing scheme.

In summary, these findings suggest that components of a pricing scheme can have a systematically differential impact on customer behavior. It is only recently that researchers have started investigating such effects, which suggests that this area holds much promise for future investigations.

6. Conclusions

In this chapter, we discussed several aspects of nonlinear (or multi-part) pricing. Such pricing schemes are very common in the service industry. We began the chapter by discussing several reasons for the use of such schemes and noted that the primary factor is the heterogeneity of the customer base. Such heterogeneity of preferences leads customers to choose different pricing plans based on their expected demand.

Next, we discussed findings from analytical work on nonlinear pricing. Here, we categorized past research based on whether it was in a monopoly setting or a more general oligopoly context. Most past research has found that two-part tariffs are optimal in many settings. Researchers have now begun to investigate the limits of optimality of two-part tariffs and when a more general pricing scheme can be optimal.

Thereafter, we summarized the past work on empirical research on multi-part tariffs. We noted that while nonlinear pricing schemes are popular, any analysis of demand under such schemes is nontrivial. A primary reason is that within multi-tier pricing schemes, there is a two-way relationship between price and consumption – the pricing scheme influences consumption and the level of consumption determines the applicable per-unit price. Two other issues are especially relevant within service contexts. First, the linkage between the choice of a service plan and usage under the chosen plan has to be appropriately specified. Two, there is a need to incorporate consumption uncertainty within any demand model. We discussed how researchers have addressed these issues and then showed a modeling framework that integrates all three issues. We ended by discussing some empirical generalizations, which also suggested some promising areas for future research.

References

Allenby, Greg M., Thomas S. Shively, Sha Yang and Mark J. Garratt (2004), 'A choice model for packaged goods: dealing with discrete quantities and quantity discounts', *Marketing Science*, **23** (1), 95–108.

Armstrong, Mark (1999), 'Price discrimination by a many-product firm', *Review of Economic Studies*, **66** (1), 151–68.

Armstrong, Mark and John Vickers (2001), 'Competitive price discrimination', *RAND Journal of Economics*, **32** (4), 579–605.

Billings, Bruce R. and Donald E. Agthe (1980), 'Price elasticities for water: a case of increasing block rates', *Land Economics*, **56**, 73–84.

Blomquist, N. Soren (1996), 'Estimation methods for male labor supply functions: how to take care of non-linear taxes', *Journal of Econometrics*, **70**, 383–405.

Burtless, Gary and Jerry Hausman (1978), 'The effects of taxation on labor supply: evaluating the Gary income tax maintenance experiment', *Journal of Political Economy*, **86**, 1103–30.

Chiang, Jeongwen (1991), 'A simultaneous approach to the whether, what and how much to buy questions', *Marketing Science*, **10** (4), 297–315.

Chintagunta, Pradeep (1993), 'Investigating purchase incidence, brand choice and purchase quantity decisions of households', *Marketing Science*, **12** (2), 184–208.

Danaher, Peter J. (2002), 'Optimal pricing of new subscription services: analysis of a market experiment', *Marketing Science*, **21** (2), 119–38.

Dolan, Robert J. (1987), 'Quantity discounts: managerial issues and research opportunities', *Marketing Science*, **6** (1), 1–22.

Dolan, Robert J. and Hermann Simon (1996), *Power Pricing*, New York: The Free Press.

Dubin, Jeffrey A. and Daniel L. McFadden (1984), 'An econometric analysis of residential electric appliance holdings and consumption', *Econometrica*, **52** (2), 345–62.

Ellison, Glenn (2005), 'A model of add-on pricing', *Quarterly Journal of Economics*, **120**, 1349–72.

Essegaier, Skander, Sunil Gupta and Z. John Zhang (2002), 'Pricing access services', *Marketing Science*, **21** (2), 139–59.

Goettler, Ron L. and Karen Clay (2007), 'Price discrimination with experience goods: sorting-induced biases and illusive surplus', Working Paper, Carnegie Mellon University.

Goldman, Charles A., Joseph H. Eto and Galen H. Barbose (2002), 'California customer load reductions during the electricity crisis: did they help to keep the lights on?', Ernest Orlando Lawrence Berkeley National Laboratory, Paper No: LBNL–49733.

Hall, Robert (1973), 'Wages, income and hours of work in the U.S. labor force', in Glen Cain and Harold Watts (eds), *Income Maintenance and Labor Supply*, Chicago, IL: Markham Press, pp. 102–62.

Han, Sangman, Sunil Gupta and Donald R. Lehmann (2001), 'Consumer price sensitivity and price thresholds', *Journal of Retailing*, **77** (4), 435–56.

Hanemann, Michael W. (1984), 'Discrete/continuous models of consumer demand', *Econometrica*, **52**, 541–61.

Hausman, Jerry (1985), 'The econometrics of non-linear budget sets', *Econometrica*, **53**, 1255–82.

Hausman, Jerry and David Wise (1976), 'The evaluation of results of truncated samples: the New Jersey income maintenance experiment', *Annals of Economic and Social Measurement*, **5**, 421–45.

Hausman, Jerry, M. Kinnucan and Daniel L. McFadden (1979), 'A two level electricity demand model: evaluation of the Connecticut time-of-day pricing test', *Journal of Econometrics*, **8**, 263–89.

Heckman, James J. (1979), 'Sample selection bias as a specification error', *Econometrica*, **47**, 153–67.

Heckman, James and Thomas MaCurdy (1981), 'New methods for estimating labor supply functions: a survey', in Ronald Ehrenberg (ed.), *Research in Labor Economics*, Vol. 4, Greenwich, CT: JAI Press, pp. 65–102.

Hewitt, Julie A. and W. Michael Hanemann (1995), 'A discrete/continuous choice approach to residential water demand under block rate pricing', *Land Economics*, **7**, 173–92.

Iyengar, Raghuram (2007), 'A structural demand model for wireless services', Working Paper, University of Pennsylvania.

Iyengar, Raghuram, Asim M. Ansari and Sunil Gupta (2007a), 'A model of consumer learning for service quality and usage', *Journal of Marketing Research*, **44** (4), 529–44.

Iyengar, Raghuram, Kamel Jedidi and Rajeev Kohli (2007b), 'A conjoint approach to multi-part pricing', *Journal of Marketing Research*, **45** (2), 195–210.

Iyengar, Sheena S. and Mark R. Lepper (2000), 'When choice is demotivating: can one desire too much of a good thing?', *Journal of Personality and Social Psychology*, **79**, 995–1006.

Iyengar, Sheena S., Wei Jeing and Gur Huberman (2004), 'How much choice is too much?: Determinants of individual contributions in 401K retirement plans', in O.S. Mitchell and S. Utkus (eds), *Pension Design and Structure: New Lessons from Behavioral Finance*, Oxford: Oxford University Press, pp. 83–95.

Jensen, Sissel (2006), 'Implementation of competitive nonlinear pricing: tariffs with inclusive consumption', *Review of Economic Design*, **10** (1), 9–29.

Kling, John P. and Stephen S. Van der Ploeg (1990), 'Estimating local telephone call elasticities with a model of stochastic class of services and usage choice', in A. de Fontenay, M. Shugard and D. Sibley (eds), *Telecommunications Demand Modeling*, Amsterdam: North Holland, pp. 119–36.

Kridel, Donald J., Dale E. Lehman and Dennis L. Weisman (1993), 'Option value, telecommunications demand and policy', in *Information Economics and Policy*, New York: Elsevier, pp. 125–44.

Lambrecht, Anja and Bernd Skiera (2006), 'Paying too much and being happy about it: existence, causes and consequences of tariff-choice biases', *Journal of Marketing Research*, **43** (2), 212–23.

Lambrecht, Anja, Katja Seim and Bernd Skiera (2007), 'Does uncertainty matter? Consumer behavior under three-part tariffs', *Marketing Science*, **6** (5), 698–710.

Lemon, Katherine N., Tiffany B. White and Russell S. Winer (2002), 'Dynamic customer relationship management: incorporating future considerations into the service retention decision', *Journal of Marketing*, **66**, 1–14.

Manfrim, Gustavo and Sergio Da Silva (2007), 'Estimating demand elasticities of fixed telephony in Brazil', *Economics Bulletin*, **12** (5), 1–9.

Masuda, Yasushi and Seungjin Whang (2006), 'On the optimality of fixed-up-to tariff for telecommunication service', *Information Systems Research*, **17** (3), 247–53.

Miravete, Eugenio J. (2002), 'Estimating demand for local telephone service with asymmetric information and optional calling plans', *The Review of Economic Studies*, **64**, 943–71.

Miravete, Eugenio J. (2007), 'Competing with menus of tariff options', Working Paper, University of Texas at Austin.

Miravete, Eugenio J. and Lars Frederick Roller (2004), 'Estimating markups under nonlinear pricing competition', *Journal of the European Economic Association*, **2**, 526–35.

Mirman, Leonard J. and David S. Sibley (1980), 'Optimal nonlinear prices for multiproduct monopolies', *The Bell Journal of Economics*, **11**, 659–70.

Moffitt, Robert (1990), 'The econometrics of kinked budget sets', *The Journal of Economic Perspectives*, **4** (2), 119–39.

Narayanan, Sridhar, Pradeep Chintagunta and Eugenio J. Miravete (2007), 'The role of self selection, usage uncertainty and learning in the demand for local telephone service', *Quantitative Marketing and Economics*, **5**, 1–34.

Nason, Robert W. and Albert J. Della Bitta (1983), 'The incidence and consumer perceptions of quantity surcharges', *Journal of Retailing*, **59** (2), 40–53.

Nordin, John A. (1976), 'A proposed modification of Taylor's demand analysis: comment', *Bell Journal of Economics*, **7**, 719–21.

Nunes, Joseph C. (2000), 'A cognitive model of people's usage estimation', *Journal of Marketing Research*, **37**, 397–409.

NYTimes Advertising Rates (2008), http://www.nytadvertising.com/was/files/others/05-3419-8_General Classified.pdf.

Oi, Walter Y. (1971), 'A Disneyland dilemma: two-part tariffs for a Mickey Mouse monopoly', *Quarterly Journal of Economics*, **85**, 77–96.

Oren, Shmuel S., Stephen S. Smith and Robert B. Wilson (1985), 'Capacity pricing', *Econometrica*, **53**, 545–66.

Park, Rolla E., Bruce M. Wetzel and Bridger M. Mitchell (1983), 'Price elasticities for local telephone calls', *Econometrica*, **51** (6), 1699–730.

Reiss, Peter C. and Matthew W. White (2005), 'Household electricity demand, revisited', *Review of Economic Studies*, **72** (3), 853–83.

Reiss, Peter C. and Matthew W. White (2007), 'Evaluating welfare with nonlinear prices', Working Paper, University of Pennsylvania.

Rochet, Jean-Charles and Lars A. Stole (2002), 'Nonlinear pricing with random participation', *Review of Economic Studies*, **69** (1), 277–311.

Rosen, Harvey (1976), 'Taxes in a labor supply model with joint wage-hours determination', *Econometrica*, **44**, 485–507.

Schmalensee, Richard (1981), 'Monopolistic two-part pricing arrangements', *Bell Journal of Economics*, **12**, 445–66.

Scotchmer, Suzanne (1985), 'Two-tier pricing of shared facilities in a free entry equilibrium', *RAND Journal of Economics*, **16**, 456–72.

Stole, Lars A. (1995), 'Nonlinear pricing and oligopoly', *Journal of Economic Management Strategy*, **4** (4), 529–62.

Sundararajan, Arun (2004), 'Nonlinear pricing of information goods', *Management Science*, **50** (12), 1660–73.

Taylor, Lester D. (1975), 'The demand for electricity: a survey', *Bell Journal of Economics*, **6** (1), 74–110.

Train, Kenneth E., Daniel L. McFadden and Moshe Ben-Akiva (1987), 'The demand for local telephone service: a fully discrete model of residential calling patterns and service choices', *RAND Journal of Economics*, **18** (1), 109–23.

Van Soest, Arthur (1995), 'Discrete choice models of family labor supply', *Journal of Human Resources*, **30**, 63–88.

Van Soest, Arthur, Marcel Das and Xiadong Gong (2002), 'A structural labor supply model with nonparametric preferences', *Journal of Econometrics*, **107** (1–2), 345–74.

Varian, Hal R. (1985), 'Price discrimination and social welfare', *American Economic Review*, **74** (4), 870–75.

Villas-Boas, Miguel and Udo Schmidt-Mohr (1999), 'Oligopoly with asymmetric information: differentiation in credit markets', *RAND Journal of Economics*, **30** (3), 375–96.

Wales, Terence and Anthony Woodland (1979), 'Labor supply and progressive taxes', *Review of Economic Studies*, **46**, 83–95.

Wilson, Robert B. (1991), 'Multiproduct tariffs', *Journal of Regulatory Economics*, **3**, 5–26.

Wilson, Robert B. (1993), *Nonlinear Pricing*, Oxford: Oxford University Press.