

# Value Creation in Service Delivery: Relating Market Segmentation, Incentives, and Operational Performance

Evrim D. Güneş

INSEAD, Boulevard de Constance, 77305 Fontainebleau, France, evrim-didem.gunes@insead.edu

O. Zeynep Aksin

College of Administrative Sciences and Economics, Koç University, Rumelifeneri Yolu,  
34450 Sariyer, Istanbul, Turkey, zaksin@ku.edu.tr

This paper studies service-delivery design in settings where firms engage in value-creation activities that have the objective of generating additional revenue from customer interactions. The paper provides a general modelling framework to analyze the ties between market segmentation decisions, incentives, and process performance in such service-delivery systems. The firm is modelled as a single-server queue, in a principal-agent framework. Customers have different value-generation potentials whose realizations are observed by the server but not by the manager of the firm. The manager determines a market segmentation scheme given an overall customer value-generation profile, which divides customers into two groups (high and low), and also determines a service level for each segment. The server decides which of the two available service levels (high and low) to provide for each customer, given a compensation scheme offered by the manager. The optimal market segmentation decision, optimal service-level choice, and a set of optimal linear incentive contracts that enable their implementation are characterized. The robustness of these strategies is explored with respect to model parameters and assumptions. It is shown that a market segmentation scheme that combines revenue generation concerns with their process implications is essential for success. Characteristics of appropriate incentive schemes are identified.

*Key words:* call centers; cross-selling; incentives; queueing; marketing-operations interface

*History:* Received: May 5, 2003; accepted: May 3, 2004. This paper was with the authors 2 months for 3 revisions.

## 1. Introduction

In a wide range of industries today, market share growth is no longer significant, and growth is driven mainly by increasing the profitability of existing customers. This increase in profitability is pursued not only by improving efficiency measures but also by persuading existing customers to spend more money with the firm (Fisher 2001). Value-creation initiatives such as cross-selling are one form of achieving this aim. Today, these initiatives are part of a firm's customer relationship management (CRM) strategy and are supported by a panopoly of IT systems. The global market for CRM systems, service, and technology is estimated to be around \$25 billion (Benjamin 2001).

Increasingly, however, companies report failures of CRM-related initiatives. Among some of the

most-cited reasons for these failures are failure to integrate it to the back-office operations and failure to train and motivate the staff. Indeed, as noted in *The Economist* (2001), "Durable customer relations are partly about clever technology. Mainly, they require relentless attention to detail: good products, prompt service, well-trained staff with the power to do a little extra when they judge it right to do so" (p. 9). Service employees have an important role in determining customer needs and acting accordingly, because they are the ones who interact with the customer during a service encounter. Even in environments in which good customer information exists and automated prompts guide servers' effort to enhance customer profitability, the ultimate decision of the service level to be offered to each customer rests with the

server. In assessing the potential profitability of offering a customer value-added services, the server uses public information such as past buying behaviour, as well as private information such as what the customer says during the phone call. In this type of a setting, appropriate incentive design is essential in ensuring a match between servers' performance and the service provider's desires.

This paper focuses on value-creation strategies such as cross-selling or add-on sales. Our model captures the deepening of an existing relationship with a customer, where depth is characterized by additional revenue per transaction and not by additional transactions. In this setting, we explore the following questions: When is it worth undertaking a value-creation initiative such as cross-selling? What customer characteristics induce a desire to spend effort on value generation? Is it better to increase the profitability of all customers by uniformly targeting all customers of a firm, or should a company pursue a service-level differentiation strategy? What server-incentive schemes should be used to implement the desired actions of each strategy? How should these incentive schemes address the tension between creating value and providing fast and efficient service? Are value-generation-related incentives (such as sales incentives), alone, enough to achieve the desired strategies?

These questions are answered using a principal-agent model. Profitability characteristics of the customers are assumed to be given. The firm is modeled as a queue. The principal, or the manager, has a choice between the strategies labeled as remaining a cost center, targeting all customers, or pursuing service-level differentiation. Whereas remaining a cost center—i.e., not pursuing any additional value-creation activity—is the status quo option, targeting all customers requires expending additional effort on all customers, and service-level differentiation requires additional effort for a segment of the customer base. The key trade-off the manager is facing in making this decision is between value or revenue generated and costs. Revenues are determined by the value-creation effort and customer profitability characteristics known by the server. Costs are in the form of incentive payments necessary to induce the desired actions by the servers, as well as the systemwide cost due to the congestion effect of the

additional effort expended for value creation. Using this framework, we show under what conditions each strategy is preferred and what type of incentive scheme is necessary to ensure its implementation. In the first part of the analysis, it is assumed that the segments (high and low) for the service-level differentiation strategy are given exogenously, and both the manager and the server take these as given. The sensitivity of the results to this market segmentation decision are then explored, which leads to the second part of the analysis, where the segmentation choice is a decision variable for both the server and the manager.

The remaining parts of the paper are organized as follows. Relevant literature is reviewed in §2. The model is introduced in §3. We analyze the resulting principal-agent problem in §4, and characterize optimal strategies and contracts for given customer profitability and customer segmentation choice. The sensitivity of these results to the customer-segmentation decision is explored in §5.1. The optimal market segmentation choice and a contract that allows its implementation are characterized in §5.2. The paper concludes with a discussion of the main results in §6.

## 2. Literature Review

Any value-creation strategy requires an understanding of the relationship between customer needs and service offerings, and how these generate value. The fact that different needs can require different offerings and thus generate different profits for the firm is the basic premise that motivates a vast literature in marketing on market segmentation. Proliferation of direct and interactive forms of communication in recent years has brought concepts such as one-to-one marketing and relationship marketing to the forefront, leading to a stream of literature that focuses on estimating customer profitability. These papers typically focus on value estimation and ignore costs, despite the need to the contrary (Foster et al. 1996). The papers that do consider costs typically include only the marketing costs incurred for a customer in their profitability estimates (Mulhern 1999), or they assume a fixed service cost element ignoring the interaction between service-level and operational costs (Berger and Nasr 1998). For example, Niraj et al. (2001) explicitly include the supply chain costs in their model of customer

profitability analysis. However, they have an activity-based cost accounting model, which allocates the costs after they are incurred rather than considering the operational costs explicitly before making the service-level decision. This type of analysis is classified as *retrospective* by Storbacka (1997), because it is based on historical data. In contrast to this approach, the *prospective* approach considers the fact that customer profitability can be changed or influenced through the service provider's actions.

The approach in this paper can be viewed as being closer to the *prospective* analysis described in Storbacka (1997). We assume that the likelihood of generating revenue from a customer depends on the level of service provided. Thus, customer profitability is determined by the likelihood of generating revenue from high-level service and the associated congestion cost of offering such high-level service to a particular customer. Although it might be possible to estimate profitability for individual customers, service levels are typically determined for a segment of customers rather than for individuals. Thus, we consider the case in which a market segmentation decision separates customers into groups, and customers in a group are assumed to have an average revenue generation potential, which can be derived from the prior on the distribution of the revenue for an individual customer. The manager determines the optimal level of service that should be provided to customers in each segment, given revenue generation probabilities and cost parameters. The simplest case with two segments is considered for the analysis in this paper. The choice between service levels is represented as a choice between performing a basic service task or a combined basic and extension task, where the latter represents a higher service level.

The impact of combining tasks in processes, in terms of its effect on congestion, has been extensively studied in the operations management literature, mainly considering the systems as cost-minimizing units. An important finding is the pooling result, which says that combining tasks decreases congestion. The importance of various human resource issues in assessing the performance of a pooled system has been discussed and incorporated in different settings (Loch 1998, Buzacott 1996, Pinker and Shumsky 2000, Powell 2000, Buzacott 2002). The interaction

among combining tasks, incentives, and value generation, that we consider herein, has not been addressed before.

Our model lies at the interface of the problems dealt with in marketing and operations management. Marketing research focuses on value generation, but because there is no explicit modeling of the operational side we cannot take these value data to generate action plans in terms of appropriate service levels. The operations management literature that deals with process design, however, focuses on costs and does not consider the value implications of various process designs. Akşin and Harker (1999) analyze the congestion effect of a particular value-creation initiative in call centers. The revenue generation from a customer is not explicitly modeled. Fridgeirsdottir and Chiu (2001) model a marketing effort decision analytically in a queueing setting. Although their analysis models value creation from a customer, it considers only direct marketing cost associated with this value-creation effort.

There is a huge literature that deals with incentive contracts and agency problems in economics, marketing, and (more recently) in operations management. Among the classical papers on agency theory, Grossman and Hart (1983) and Holmstrom and Milgrom (1991) assume a generic function for the output rather than using the models for the underlying operational system through which the effort leads to outcomes. In marketing, a stream of literature on salesforce compensation has started from models with deterministic output functions (Farley 1964) and evolved into agency theoretic models. Basu et al. (1985) present various salesforce compensation plans in a principal-agent framework. The assumption of constant marginal cost is common in this literature (see, for example, Lal and Srinivasan 1993). For a thorough review of the salesforce compensation literature the reader is referred to Coughlan (1993).

In the salesforce compensation context, our study provides a link between the incentive and operations problems by explicitly modeling the operational costs of pursuing this additional value as opposed to assuming a constant marginal cost of production. In the typical setting considered by the salesforce compensation literature, the server is a salesperson whose job description is "selling." In our model, we

consider service settings in which the primary role of the server is to provide service and the additional extension task can be considered as a sales activity. As such, the sales activity is an additional component of the server's job description.

In the operations management literature, there are some studies considering incentive effects in different operational settings. A good review of this literature can be found in Plambeck and Zenios (2000). Studies on queueing systems have more often focused on pricing issues and related customer incentives, as in the articles of Mendelson and Whang (1990), Bradford (1996), and Van Mieghem (2000). Examples of papers that consider server incentives in congestion-prone settings are Gilbert and Weng (1998) and Shumsky and Pinker (2003). The latter considers incentive issues in service contexts such as medical services or call centers, where there is a gatekeeper who makes an initial diagnosis of a customer's problem and then either solves it or refers it to a specialist. The effect of different contracts on the referral rate the gatekeeper chooses are investigated in an environment in which the gatekeeper has an ability (unknown to the firm) to deal with problems of varying difficulty. The incentive side of our model is similar in structure to the gatekeeping problem. However, we consider incentive problems stemming from the variance in processing times and customer identities, as opposed to the server's identity.

### 3. The Model

We model the provision of a service that can be offered at two different levels. The standard level requires no effort from the server and generates no revenue. On one hand, this represents the prevailing level of service if no value creation is sought by the server. On the other hand, if the server opts for the high level of service this requires effort and results in the possibility of generating revenues. Using this model, we analyze the service-level decision, which determines the customer segment for which the revenue generating high level of service is optimal. Corresponding incentive contracts are characterized. The firm is modeled as a profit-maximizing, single-server Markovian queue with unlimited waiting space.

**Customer Base and Value Generation.** Customers arrive according to a Poisson process of rate  $\lambda$ . There

are two customer types, high and low, which we label as  $H$  and  $L$ , respectively. The server can observe the customer type at the start of service and incurs no cost for diagnosing a customer's type. For any customer, the probability of being a high type is  $q$ , and the probability of being a low type is  $(1 - q)$ . Thus, the parameter  $q$  determines the size of the high-type segment. The revenue generation potentials depend on the type of the customer and the service level offered. The probability of generating revenue  $R$  by offering a high level of service is  $p_H$  for the high-type customers and  $p_L$  for the low-type customers. We also make the assumption that  $p_H > p_L$ .

We illustrate in §5.1 how these parameters relate to the market segmentation decision of the firm. Until then, these parameters are taken to be given. The basic model, where these parameters are taken as given, can be seen as representing the case of a functional organization, in which customer-related information and any segmentation decision is taken by the marketing function and not questioned elsewhere. Thus, the manager and the server in the operations function take these as given. This assumption is relaxed in §5.2, where both the manager and the server are more sophisticated. The manager determines an optimal market segmentation scheme, which in turn determines the parameters  $q$ ,  $p_H$ , and  $p_L$ . The server might not accept this segmentation scheme, unless he is offered the appropriate incentives to do so. This latter setting represents the case of a more integrated organization, in which both manager and server have an understanding of the entire process rather than just a functional view.

**Service Process and Costs.** There are two service levels that can be offered to the customer: *standard* or *extended* service. Server effort is represented by the binary variable denoted by  $e_H \in \{0, 1\}$  and  $e_L \in \{0, 1\}$  for high-type and low-type customers, respectively.  $e_H = 0$  or  $e_L = 0$  represents the case with no effort and  $e_H = 1$  or  $e_L = 1$  the case in which the server exerts effort. Standard service does not require any effort from the server, so the effort is 0. It generates no extra revenue, hence we normalize the revenue in this case to 0. The service time for this type of service is exponentially distributed with rate  $\mu$ . The second level, *extended* service, where a service extension is

provided, can be interpreted as additional personalized attention, or a cross-sell attempt. This extension requires effort on the part of the server, so the effort is 1. As a result of the server's effort, a revenue of  $R$  is generated with a fixed probability that depends on the type of customer being served, as explained before. This effort is also reflected in the time spent for the service. The service time is exponential with rate  $\mu - k$  ( $\mu > k > 0$ ) in this case, where  $k$  represents the content or complexity of the extension task. Expending effort is unpleasant for the server, so he has a disutility of  $C_S$  whenever  $e_H = 1$  or  $e_L = 1$ . Given the nature of the extension task, this implies that the server does not enjoy the sales activity. This represents the direct cost of providing high-level service to a customer.

In addition, there are indirect costs due to the congestion experienced by customers in the system. For any customer, the time spent in queue costs  $c$  per unit time to the firm. This parameter represents the importance of congestion for the firm. There is no standardized practice for quantifying waiting costs (Borst et al. 2004). If market research reveals that customer loss of goodwill due to waiting is not a major cost component, then phone line cost can be used as a proxy. Otherwise, a marketing research technique for eliciting customer preferences, like conjoint analysis, can be used to assess the value of different waiting times to the firm (see Pullman and Moore 1999, Newman 1984, and Verna and Thompson 1999, for example applications).

Note that in this model increasing effort results in a *decrease* in service rate, contrary to the more common assumption in the literature that increasing effort increases service rate (for e.g., Gilbert and Weng 1998, Kalai et al. 1992). An important implication related to this is that high effort *might not* be desirable because of this consequent decrease in service rate, which would decrease the profitability of the customer (i.e., revenues net of costs of serving that customer) due to the increase in costs.

**Information and Decision Structure.** We assume there is a manager (she) who wants to implement a policy  $\pi$ , which is defined as the effort levels provided for each customer type, i.e.,  $\pi = (e_H, e_L)$ . There is one server (he) who serves each arriving customer after observing his or her type. The manager observes

only the time spent for the customer and the revenue generated. She cannot observe the realization of customer type, nor the distribution from which the service time is drawn. Hence, she does not know if a certain outcome is the result of the server's effort choice or of chance. However, the server does not incur a cost for the waiting time of customers, and furthermore, he does not like expending effort. As a result, his decision might not be optimal for the firm if he is not compensated appropriately. The manager wants to ensure that the desired service levels are offered to each customer, which might depend on the customer type, so she needs to find an incentive scheme that would induce the best decision by the server in the presence of moral hazard (the effort is not observed) and private information (customer type for a given realization is not observed).

**Performance Measures.** There are two outcomes that are the results of a server's effort decision that contribute to the overall system performance: service time,  $x_1$ , and revenue generated,  $x_2$ . The manager decides on a compensation scheme and declares a policy,  $\pi = (e_H, e_L)$  that she wants the server to implement. Then the agent (the server) decides on the effort levels  $(e_H, e_L)$  that maximize his utility (compensation less disutility for effort). The effort decision is taken once and applied to all customers in a particular segment, i.e., the decision is not taken dynamically. We assume that the contract is linear in the two outcomes  $x_1$  and  $x_2$  and that both the principal and the agent are risk neutral, maximizing their expected linear utilities.

The manager's objective is to maximize profits, i.e., revenues minus the costs as payments to the server and the cost associated with congestion in the system. The first cost component for the manager is the compensation of the server,  $w$ . We define the payment scheme as

$$w = \alpha_1 x_1 + \alpha_2 x_2 \quad (1)$$

for any customer served, where  $x_1$  is the service time and  $x_2$  is the revenue generated for that particular customer.

We assume that the principal measures performance on an individual customer basis. For each customer served, the outcome measures (service time and revenue generated) are determined and the corresponding bonus amount is added to the server's

account. In §4.3 we explore how measuring performance on average outcomes rather than single customer realizations can change these results.

The optimal policy for the firm is determined by taking into account the revenue generation potentials of the two customer types, and the additional costs for extended service. These costs include the direct cost of effort by the server, and the indirect cost of extra congestion in the system. A customer is said to be profitable if the revenue generated from him or her exceeds these costs. To avoid trivial cases, we assume that the direct costs of providing extension to the low types is less than the expected revenues, i.e.,  $Rp_L > C_S$ , so that when only the direct costs are considered it is profitable to provide the high level of service to the low-type customers. This makes the problem more interesting, and also allows us to illustrate the effect of the indirect cost of service extension.

#### 4. Model Analysis

The optimal policy analysis is first done for a given market segmentation scheme, i.e., considering the values  $q$ ,  $p_H$ , and  $p_L$  as parameters. Recall that this represents the case of a functional organization. In §5, we discuss the consequences of changing these parameters when customer segments are redefined.

To analyze the optimal contracts, we will use the two-stage procedure suggested by Grossman and Hart (1983). This approach is simply to break up the principal's problem into a computation of costs and benefits for different actions taken by the agent. For each policy  $\pi$ , we consider the incentive scheme that minimizes the expected cost of getting the agent to choose effort levels stipulated by that policy, and then select the policy with maximum profit for the manager.

The first cost component for the manager's objective function, compensation, is defined by

$$E[x_1] = \frac{1}{\mu} + \left( q \frac{e_H k}{\mu(\mu - k)} + (1 - q) \frac{e_L k}{\mu(\mu - k)} \right), \quad (2)$$

and

$$x_2 = \begin{cases} Re_H & \text{with probability } p_H, \\ & \text{if customer is high type} \\ Re_L & \text{with probability } p_L, \\ & \text{if customer is low type} \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

The second cost component is the congestion cost, measured by the average waiting time in the queue. Note that the waiting time for any customer type depends not only on the service rate chosen for that type but also on the choice for the other type. This is because of the effect of service time mean and variance on the waiting time of a given customer. More precisely, when the same service rate is chosen for all customers, the service time is drawn from an exponential distribution, and the queue is an  $M|M|1$  queue; whereas if different rates are chosen for the two types, the queue is an  $M|G|1$  queue with hyperexponential service times.

The expected queueing time for a customer can be found using the Pollaczek-Khintchine formula,

$$W(\pi) = \frac{\lambda E[x_1^2]}{2(1 - \lambda E[x_1])},$$

where  $\pi = (e_H, e_L)$ ,  $x_1$  is the service time and  $\lambda$  is the arrival rate. We can write the expected cost of waiting in line for a customer as  $T(\pi) = c \cdot W(\pi)$ .

The objective function for the manager, for given effort levels and compensation rates, is the long-run average profit rate, which can be written as

$$\begin{aligned} E[\Pi^P(\alpha_1, \alpha_2)] &= \lambda q \left[ (1 - \alpha_2) Rp_H e_H - \alpha_1 \frac{e_H k}{\mu(\mu - k)} \right] \\ &\quad + \lambda (1 - q) \left[ (1 - \alpha_2) Rp_L e_L - \alpha_1 \frac{e_L k}{\mu(\mu - k)} \right] \\ &\quad - \alpha_1 \lambda \frac{1}{\mu} - \lambda T(\pi). \end{aligned}$$

The agent (the server) observes the customer type at each service starts and decides on an effort level that maximizes his expected utility, which is the expected wage minus the cost of effort. That is, he solves two separate problems for the two customer types

$$\begin{aligned} E[\Pi^A(\alpha_1, \alpha_2) | H] &= \alpha_2 Rp_H e_H + \alpha_1 \left( \frac{1}{\mu} + \frac{e_H k}{\mu(\mu - k)} \right) - C_S e_H, \quad (4) \end{aligned}$$

$$E[\Pi^A(\alpha_1, \alpha_2) | L] = \alpha_2 Rp_L e_L + \alpha_1 \left( \frac{1}{\mu} + \frac{e_L k}{\mu(\mu - k)} \right) - C_S e_L. \quad (5)$$

Equation (4) is the expected utility when the customer is of high type, and (5) is the expected utility when the

customer is of low type. Because a policy is defined as the effort levels chosen for both customer types, there are four possible policies that the principal and the agent can choose.

- (1) (0, 0): Standard: no effort for any customer.
- (2) (1, 0): Differentiation: effort only for high-type customers.
- (3) (1, 1): Extension: effort for all customers.
- (4) (0, 1): Reverse differentiation: effort only for low-type customers.

Analyzing the incentives of the agent, we can show that policy (0, 1) can be dropped from the analysis.

**PROPOSITION 1.** *Offering extended service only for the low-type customers (policy (0, 1)) is never optimal for the agent.*

**PROOF.** It is easy to see that the agent never prefers this policy if  $p_H > p_L$ . If it is optimal for the agent to put effort for low-type customers, it must be optimal to do so for high-type customers as well, because the expected revenue from high type is higher. Hence, policy (0, 1) can never be optimal for the agent.  $\square$

To find the optimal policy for the manager, we maximize the long-run average profits for the three alternative policies. These can be written as follows, conditioned on the policy chosen:

$$E[\Pi^P(\alpha_1, \alpha_2) | (0, 0)] = -\lambda T(0, 0) - \alpha_1 \frac{\lambda}{\mu}, \quad (6)$$

$$\begin{aligned} E[\Pi^P(\alpha_1, \alpha_2) | (1, 0)] &= \lambda q \left[ (1 - \alpha_2) R p_H - \alpha_1 \frac{1}{(\mu - k)} \right] \\ &\quad - (1 - q) \alpha_1 \frac{\lambda}{\mu} - \lambda T(1, 0), \end{aligned} \quad (7)$$

$$\begin{aligned} E[\Pi^P(\alpha_1, \alpha_2) | (1, 1)] &= \lambda (1 - \alpha_2) [q R p_H + (1 - q) R p_L] \\ &\quad - \alpha_1 \frac{\lambda}{(\mu - k)} - \lambda T(1, 1). \end{aligned} \quad (8)$$

The *first-best solution* refers to the optimal solution, which can be achieved when there is no information asymmetry, i.e., when the customer types and the effort can be observed by the principal. In this case, because the customer types and effort levels are observed, the agent will be compensated for the effort he expends. Formally, the first-best contract is defined as a payment for each customer type, different from the contract definition given in (1), and it can be found solving the following program. Given a policy

$\pi = (e_H, e_L)$ , this program finds the first-best contract  $w^{FB} = (w_H, w_L)$ , which compensates the server  $w_H$  for each high-value customer and  $w_L$  for each low-value customer that is served. The first best contract maximizes the expected profits subject to the participation constraint of the agent, assuming the reservation utility for the agent is zero. Note that because the efforts are observed, there is no need for incentive compatibility constraints.

$$\begin{aligned} \max_{w_H, w_L} E[\Pi^{FB} | (e_H, e_L)] &= \lambda R(q p_H e_H + (1 - q) p_L e_L) \\ &\quad - \lambda T(\pi), \\ \text{s.t. } q(w_H - C_S e_H) + (1 - q)(w_L - C_S e_L) &\geq 0. \end{aligned}$$

In the optimal contract, the constraint inequality will be an equality. The first-best contract compensates the agent as much as his effort cost,  $C_S$ , whenever  $e_H = 1$  or  $e_L = 1$ , and pays 0 (the reservation wage), otherwise

$$w^{FB} = (w_H, w_L) = (C_S e_H, C_S e_L). \quad (9)$$

#### 4.1. Stage 1: Optimal Contracts for Each Policy

We can now solve for the optimal contract under each alternative policy, which constitutes the first stage of the analysis in the two-stage solution methodology. The complete optimization program for the differentiation policy, (1, 0), is shown below. For (0, 0) and (1, 1) the results are presented in the appendix, given the similarity of the analysis to the case for (1, 0). We use superscripts *FB* and \* to refer to the first-best solution and the optimal that a manager can achieve, respectively. In this optimization program, the set of  $(\alpha_1, \alpha_2)$  values that generates a set of feasible effort levels is defined by the agent's incentive compatibility constraints ICH and ICL. In addition, the expected utility that the agent gets from this contract should be at least as much as the reservation utility, which defines the outside option for the server. This is the individual rationality constraint (IR2), which defines the feasible set of contracts together with the incentive compatibility constraints. Finally, we have the constraint  $\alpha_2 \leq 1$  to ensure that the compensation for the revenue generated is not more than the revenue itself. Note that for any policy  $\pi$ , the congestion cost in the objective function,  $T(\pi)$ , is a constant. That is, the incentive contract design problem is not affected by the congestion measure, once a policy is given. The term  $T(\pi)$  plays a role only in the second-stage problem, where the optimal policy will be chosen.

$$\max_{\alpha_1, \alpha_2} E[\Pi^P(\alpha_1, \alpha_2)] = \lambda q \left[ (1 - \alpha_2) R p_H - \alpha_1 \frac{1}{(\mu - k)} \right] \\ - (1 - q) \alpha_1 \frac{\lambda}{\mu} - \lambda T(1, 0),$$

s.t.

$$q \left( \alpha_2 R p_H + \alpha_1 \frac{1}{\mu - k} - C_S \right) + (1 - q) \alpha_1 \frac{1}{\mu} \geq 0, \quad (\text{IR2})$$

$$\frac{\alpha_1}{\mu - k} + \alpha_2 R p_H - C_S \geq \frac{\alpha_1}{\mu}, \quad (\text{ICH})$$

$$\frac{\alpha_1}{\mu - k} + \alpha_2 R p_L - C_S \leq \frac{\alpha_1}{\mu}, \quad (\text{ICL})$$

$$\alpha_2 \leq 1,$$

$$\alpha_1^* \in \left[ -q C_S \frac{p_H - p_L}{\mu p_L + k q (p_H - p_L)}, 0 \right],$$

$$\alpha_2^* = \frac{C_S}{R p_H} - \alpha_1^* \frac{\mu - (1 - q)k}{\mu(\mu - k)q R p_H},$$

$$E[\Pi^*(\alpha_1, \alpha_2) | (1, 0)] = \lambda q R p_H - \lambda q C_S - \lambda T(1, 0).$$

First-best solution:

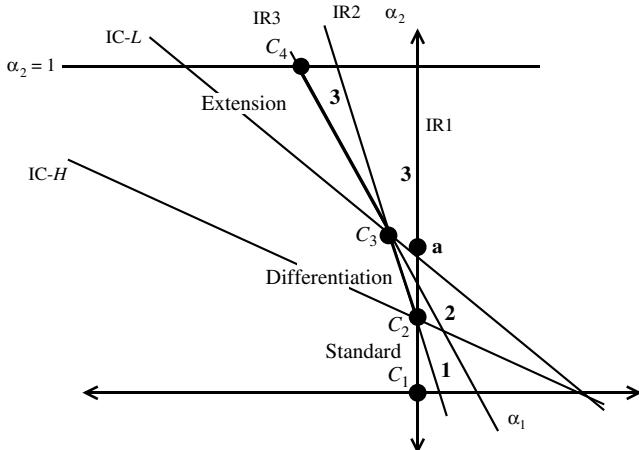
$$w^{FB} = (w_H, w_L) = (C_S, 0)$$

$$E[\Pi^{FB} | (1, 0)] = \lambda q R p_H - \lambda q C_S - \lambda T(1, 0).$$

The optimal contract rates for each policy can be seen graphically in Figure 1.

The numbered regions are the feasible regions for each policy considering the agent's incentive compatibility constraints; 1, 2, and 3 stand for a standard, differentiation, and extension policy, respectively. The

**Figure 1 Optimal Contracts and Sensitivity**



first-best solutions are achieved by the contracts that bind the individual rationality constraints for the policy, which are labelled by IR1, IR2, and IR3 for standard, differentiation, and extension policies, respectively. There are infinitely many contracts for each policy, in the feasible range determined by the incentive compatibility constraints of the agents and the constraint  $\alpha_2 \leq 1$ . Optimal contracts for each policy are designated by the thick lines, labeled by the policy name.

The contract rates at the end points seen in Figure 1 are as follows:

$$C_1 = (0, 0),$$

$$C_2 = \left( 0, \frac{C_S}{R p_H} \right),$$

$$C_3 = \left( -q C_S \frac{(p_H - p_L)(\mu - k)\mu}{\mu p_L - k(q p_H + (1 - q)p_L)}, \right.$$

$$\left. \frac{C_S}{R} \frac{(\mu - k)}{\mu p_L - k(q p_H + (1 - q)p_L)} \right),$$

$$C_4 = (-(R(q p_H + (1 - q)p_L) - C_S)(\mu - k), 1).$$

In all cases, the manager can achieve the first-best solution with the incentive contracts (because the agent is risk neutral). As a result, the optimal profits for the principal are the same as the first-best case.

Note that the compensation rates ( $\alpha_1, \alpha_2$ ) are not independent. The rate offered for service determines the rate that should be offered for revenue generated. Thus we find that incentive schemes that have only value-generation-related incentives (for example, sales-related incentives), commonly found in practice, are not always optimal. The trade-offs between the two performance metrics need to be taken into account explicitly for incentive contract designs.

For the differentiation policy,  $\alpha_1$  should be non-positive, and for the extension policy it is entirely in the negative region. The negative compensation rate implies a punishment for long service. This is used as a means to balance the commission paid for the revenue generated and to provide an incentive for the server to offer a shorter, rather than a longer (i.e. value-added), service. For the differentiation policy, the punishment for service time prevents the server from providing extended service to low-class customers. This is so because in the optimal contract

region (line segment  $[C_2, C_3]$ ) the expected commission rate,  $\alpha_2$ , is not high enough to compensate for the punishment associated with providing long service for low-class customers, which have expected revenue of  $Rp_L$ . On one hand, an increase in  $R$  and  $p_L$  would further reduce the commission rate to maintain the right incentives in place for differentiation. On the other hand, a high effort cost for the server ( $C_S$ ) increases the commission rate required to provide high-level service. In addition, for higher values of  $q$  and  $p_H$  the commission rate and the punishment rate increase. This implies—for everything else fixed—that a firm facing a high-end market will tend to offer stronger service time and sales incentives to implement the differentiation policy.

#### 4.2. Stage 2: Policy Selection

In the second stage, the optimal policy is determined, i.e., the strategic-level decisions are taken, given the best performance with each policy. This choice will be made (by the principal) to maximize net profits, found as revenues net of the cost of congestion and the compensation paid. In general, we can say that as long as the expected revenue generated is greater than the cost of effort, there will be a range of congestion cost values in which the value-generation policies, extension, or differentiation are optimal. Because there is a trade-off between the value generated and the congestion in the system, as the unit congestion cost increases there will be a switch to the policies that offer high-level service for a smaller portion of the customer base.

Even when the effort is costless for the server, we obtain results that confirm the above intuition. If  $C_S = 0$ , i.e., there is no cost associated with effort for service extension, then

$$\begin{aligned} E[\Pi^P(\alpha_1, \alpha_2) | (1, 0)] &\geq E[\Pi^P(\alpha_1, \alpha_2) | (1, 1)] \quad \text{iff} \\ c &\geq Rp_L \mu [\mu(\mu - k - \lambda) + k\lambda(1 - q)] \frac{(\mu - k - \lambda)}{\lambda k (2\mu - k - \lambda)}. \end{aligned}$$

In other words, if the unit congestion cost is higher than a critical value, differentiation is preferred to offering high-level service to everybody, regardless of the fact that it is costless for the server to offer high-level service. As a result, given a fixed market segmentation scheme, the optimal policy choice is characterized by the critical value of congestion cost

that trades off the extra revenue from extension to a customer segment with the extra load it brings to the system. The derivation of all the results are provided in the appendix.

**PROPOSITION 2.** *The critical value of unit congestion cost such that for  $c \geq c^*$  differentiation is preferred to extension is given by*

$$c^* = \frac{\mu}{\lambda} \frac{(\mu - k - \lambda)}{(2\mu - k - \lambda)} (Rp_L - C_S) \left[ \frac{(\mu - \lambda)(\mu - k)}{k} - \lambda q \right].$$

*Similarly, the critical value of unit congestion cost such that for  $c \geq c^{**}$  the standard policy is preferred to differentiation is*

$$c^{**} = \frac{(\mu - k)}{\lambda} \frac{(\mu - \lambda)}{(2\mu - \lambda - k)} (Rp_H - C_S) \cdot \left[ \frac{(\mu - \lambda)(\mu - k)}{k} - \lambda q \right].$$

When there is only one possible market segmentation scheme—i.e., parameters  $q$ ,  $p_H$ , and  $p_L$  are fixed—the optimal strategy the firm follows is found by comparing the unit congestion cost  $c$  with the critical values  $c^*$  and  $c^{**}$ . There are three cases.

*Case I ( $c > c^{**}$ ).* The firm is not in a value-creation environment, so extension is not suitable for any market segment. The firm prefers that the operation remains a cost center.

*Case II ( $c^{**} > c > c^*$ ).* The firm can apply relationship management or value-creation strategies, but not to its entire customer base. The only customers that are worth spending effort on are the high-type customers. The firm opts for service-level differentiation.

*Case III ( $c \leq c^*$ ).* All customer segments are profitable and extension is worthwhile for all. The firm chooses to target all of its customers for additional value creation.

Given capacity ( $\mu$ ), profiles of the segments ( $q, p_H$ ), complexity of the extension of service and the abilities of the servers ( $k$ ), the revenues versus direct costs of extension ( $R, C_S$ ), and the congestion averseness of the firm ( $c$ ), the firm will be in one of the above regions, and the region dictates the best policy to implement. The value of  $c/(Rp - C_S)$ , i.e., the ratio of the cost of a customer waiting in line one unit of time to the expected revenue generated from a customer, affects the policy regions' relative sizes. The comparative statics results stated below show the effect of the

parameters on the policy choice. By taking derivatives with respect to appropriate problem parameters, we have the following:

PROPOSITION 3.

$$\frac{\partial c^*}{\partial q} < 0, \quad \frac{\partial c^*}{\partial C_S} < 0 \quad \text{and} \quad \frac{\partial c^*}{\partial p_L} > 0;$$

also

$$\frac{\partial c^{**}}{\partial q} < 0, \quad \frac{\partial c^{**}}{\partial C_S} < 0 \quad \text{and} \quad \frac{\partial c^{**}}{\partial p_H} > 0.$$

The range between the two critical values,  $(c^*, c^{**})$ , determines the attractiveness of the differentiation policy for the firm. Differentiation becomes more attractive as  $c^*$  decreases (by increasing the proportion of high-type customers and the cost of effort, or decreasing the revenue potential from low types), or  $c^{**}$  increases (by decreasing the proportion of high-type customers and the cost of effort, or increasing the success probability for high-type customers). Note that both threshold values,  $c^*$  and  $c^{**}$ , decrease with the size of the high-class segment ( $q$ ). With a bigger high-class segment, the differentiation policy might increase the congestion too much compared with standard policy, so  $c^{**}$  decreases. Similarly, a big high-class segment implies a small low-class segment, in which case the additional revenue generation potential of the extension policy is small compared with the differentiation policy; therefore  $c^*$  decreases as well, favoring the differentiation policy. This implies that the differentiation policy will be favored by a bigger high-class segment size in low congestion cost or high capacity environments, whereas it will be less worthwhile in high congestion cost or low capacity environments.

We next analyze incentive contract sensitivity to assumptions made in the analysis. In §5.1, we explore the implications of changing the market segmentation decision on the policy choice.

#### 4.3. Alternative Compensation Schemes and Incentive Contract Sensitivity

In this section we explore the possibility of control mechanisms other than linear contracts, such as full monitoring via payment over average outcome values, and we discuss their implementation.

In the model considered in this paper, the incentive contract is designed such that payments are made

after each customer is served, according to the outcomes observed for that single customer. The payment per customer scheme is motivated by what is observed in practice for sales-related incentives. Note, however, that the model also assumes that the principal can observe average waiting times, and that in theory she can use this information to monitor the actions of the agent indirectly. In that case, knowing the service and demand parameters and observing the average waiting times, the principal can calculate the realized value of  $q$ , i.e., the proportion of customers that the server has spent effort on, and pay exactly  $\lambda q C_S$  to compensate for the effort cost of the server. This can be seen as a monitoring plan as studied by Joseph and Thevaranjan (1998). Given a risk-neutral agent, both this monitoring scheme and the earlier proposed linear contract result in no loss of efficiency. Thus, monitoring does not improve the profits for the principal.

A monitoring contract such as the one explained above is an alternative solution to the linear contract, but there are some practical concerns that need to be addressed in its implementation. For the monitoring contract to be applicable, the payments need to be made after a sufficient number of interactions, so that long-run averages can be observed. Furthermore, it should be appropriate to compensate the server on the basis of averages as opposed to a per customer basis. Depending on the scale of the demand process and the payment intervals, this might not be possible in some settings.

When some of the assumptions made in the analysis are relaxed, monitoring can become even more difficult to implement. Consider the case when there is a diagnosis cost for the server for each decision he makes about the customer type, or a case in which the service extension for the low-type customer takes a longer time than for the high type. In both of these settings monitoring only the average waiting time will not be enough; information on average revenue generation will also be required to ensure the proper actions by the server. Convincing a server that basing their compensation on expected successful cross-sells is fair might be more difficult to do than doing so for the waiting-time performance, given the possible volatility in customer revenue generation potentials and arrival patterns in shorter timeframes. To account

for the diagnosis cost, the linear incentive contract can be modified such that  $\alpha_2$  is adjusted up for the differentiation policy and down for the extension policy. When each customer type has a different extension task time, the linear incentive contract for the differentiation policy would be the same, whereas for the extension policy it can be used with modified  $(\alpha_1, \alpha_2)$ .

Finally, although we consider a single-server system, in practice, where there are several servers, it would not be appropriate to attribute the average waiting time to the performance of a single server. The server, being punished or rewarded for a performance measure that he does not perceive to be totally under his control, would not be able to get good guidance and motivation from such a contract.

A second alternative compensation plan is a profit-sharing contract, which is a classical solution suggested by the principal-agent literature for the risk-neutral agent case. This type of contract would be possible if the principal could charge the waiting time cost to the agent. In our analysis we assume this is not an option because of practical concerns, as explained in the previous paragraph.

This discussion illustrates that the linear contracts are not the unique solution to our model, though they provide a feasible and robust compensation mechanism under different assumptions. Next we investigate these contracts in more detail.

**Linear Contracts.** We briefly discuss the robustness (i.e., the ability to induce the desired action when some conditions change) of the linear incentive contracts to errors in parameter estimation and contract design, and the impact of deviating from the optimal contracts as a result of these errors. On one hand, the capacity of the system  $\mu$  and the revenues  $R$  are relatively easy to assess. On the other hand, the disutility of the agent for expending effort for the service extension,  $C_S$ , can be difficult to quantify exactly. The server himself might have an assessment of this cost but can miscommunicate this information to the manager. Moreover, evidence from call centers shows that this parameter can have a wide range. Servers differ in their preferences between generating extra revenue and providing fast and efficient service (Bettencourt and Gwinner 1996). Similarly, there can be errors in assessing the parameters  $q$ ,  $p_H$ , and  $p_L$ .

There are two types of deviations from an optimal contract. The first is failure to satisfy the individual rationality constraint as an equality, thereby paying the agent more than his reservation utility. The second is failure to satisfy the incentive compatibility constraints of the agent for the optimal policy, hence getting another policy implemented. The cost of this error would be either lost revenues (when the policy implemented is differentiation instead of extension) or increased congestion in the system (in case the policy implemented is extension instead of differentiation).

These potential deviations can be visualized in Figure 1. Assume the optimal policy is differentiation, and the contract chosen is the point  $C_2 = (0, C_S/Rp_H)$ . Let the estimated effort cost be  $\delta C_S$ ,  $\delta > 1$ . Then the contract offered moves up from the optimal point  $C_2$ . For  $\delta \leq p_H/p_L$ , the contract offered is in the region designated by Number 2, and this move would be a deviation of the first type as explained above. For  $\delta > p_H/p_L$ , however, the move would be to a point such as "a," and would constitute the second type of deviation. The latter would result in operational costs due to the increase in system congestion, in addition to the efficiency loss.

The ratio  $p_H/p_L$  is a measure of robustness of the contracts that induce the differentiation policy,  $(1, 0)$ . This fraction determines the allowable range for estimation error  $\delta$  of  $C_S$ . We see that as  $p_H$  and  $p_L$  get closer the feasible region for  $(1, 0)$  becomes smaller and the robustness of the contracts deteriorates. Intuitively, as the two market segments' characteristics differ more from each other, it becomes easier to differentiate between them and provide distinctly different incentives for the treatment of the two types of customers. In that case, small estimation errors in the contract parameters would not cause dramatic differences in the servers' behaviors.

When the optimal policy is extension, however, if the condition  $k/\mu < p_L/(qp_H + (1 - q)p_L)$  does not hold, then the linear contract suggests a punishment for the revenue generated and a positive compensation on the service time outcome. We consider a negative  $\alpha_2$  value to be infeasible in practice. Further intuition can be obtained by rewriting the condition as

$$\left( \frac{k}{\mu(\mu - k)} \right) / \frac{1}{\mu} < \frac{p_L}{q(p_H - p_L)}. \quad (10)$$

This condition is comparing cost and benefits for the extension policy: The ratio of “increase in service time by extension” and “service time for standard service” should be less than the ratio of “probability of revenue from low types” and “contribution of high types to success probability”; i.e., if the service time difference is huge ( $k$  is big) or if the low types’ revenue generation potential is very small compared with high types (not high enough to compensate the service time increase on the left-hand side), then the extension policy cannot be implemented via this linear contract.

In conclusion, environments with high  $p_H/p_L$  make the linear contracts more robust for implementation of the differentiation policy, whereas the extension policy becomes more difficult to implement by linear contracts in these environments. This suggests that the contract design should take into account the policy choice of the firm. For standard policy, no incentive is actually needed, so a flat salary would be enough. For the differentiation policy, a linear contract that gives incentives on only the revenue dimension would be appropriate. For the extension policy, the contract should include both of the two outcome dimensions. Moreover, the contract requires a punishment on service time. For some extreme cases, where  $k$  is too high relative to the capacity and the revenue generation potential, the linear contract would suggest a punishment on revenue, which may not be practical. In those cases, if possible, a monitoring contract could be used instead.

## 5. Market Segmentation Problem

In the previous sections, we analyzed the policy problem given two customer types and characterized the policy choice depending on the unit congestion cost. However, in reality the customer types are the result of the market segmentation decision of the firm.

To incorporate the market segmentation decision, the model is further developed as follows. We assume that for any customer, the probability of generating revenue  $R$  by offering high-level service is a realization  $\hat{p}$  of a random variable  $P$ , which we call the *probability of success*. Management knows the density function  $f(p)$  of this random variable but cannot observe the realized value for each customer. Given

the density of success probabilities  $f(p)$ , a segmentation scheme is determined first. This is done by dividing the customer base into two segments, using a critical probability value  $\theta$ . Namely, the customers with  $p > \theta$  are defined as the high-type customers, and the remaining (with  $p \leq \theta$ ) are defined as the low-type customers. Then the average representative success probabilities can be assessed for each segment, where

$$p_H = E[P | p > \theta], \quad (11)$$

$$p_L = E[P | p \leq \theta]. \quad (12)$$

Note that for any density  $f(\cdot)$ ,  $p_H > p_L$  holds. Similarly, for each arriving customer the probability of being a high type,  $q$ , is found as

$$q = P(\text{customer type} = H) = P(p > \theta) = 1 - F(\theta). \quad (13)$$

This parameter determines the size of the high-type segment. In this section, we first demonstrate the sensitivity of the policy choice and the profits to the market segmentation decision and then solve the optimal market segmentation problem.

### 5.1. Sensitivity to Market Segmentation

Determining a market segmentation scheme corresponds to the selection of a value for  $\theta$ . Each customer has a potential for generating a revenue  $R$ . However, unless the server chooses to undertake an extension task, this potential cannot be realized. The parameter  $k$  represents the content of this extension task and can be seen as a measure of customer needs. Thus, revenues are not generated unless customer needs are met. The higher these customer needs, the higher will be the  $k$  parameter, and as a result the higher the impact of extension on congestion and costs. In other words, the parameter  $k$  characterizes the operational impact of the value-creation activity on the system. Hence, the optimal choice of  $\theta$  can be seen as a market segmentation decision that takes into account both customer revenue generation potential and customer service needs. In this section, gains from optimizing  $\theta$  are illustrated by analyzing the case in which  $\theta$  is fixed first and then the policy decision is made, taking this  $\theta$  value as given.

We can characterize the policy choice as a function of these two key parameters: the additional load that service extension brings,  $k$ , and the threshold for the

minimum probability of success for a high-type customer,  $\theta$ . Using the results of the analysis in §4.2 for any given value of  $\theta$  policy choice can be defined by critical values of  $k$  as follows.

*Case I* ( $k \leq k^*(\theta)$ ). Extension policy is optimal.

*Case II* ( $k^*(\theta) < k < k^{**}(\theta)$ ). Differentiation policy is optimal.

*Case III* ( $k^{**}(\theta) \leq k$ ). Standard policy is optimal.

The following result provides some structural properties of the curves  $k^*(\theta)$  and  $k^{**}(\theta)$ . All proofs are in the appendix.

**PROPOSITION 4.**  $k^{**}(\theta) > k^*(\theta)$  for all  $\theta \in (0, 1)$ , and  $k^*(\theta)$  and  $k^{**}(\theta)$  are nondecreasing with  $\theta$ :

$$\frac{\partial k^*}{\partial \theta} \geq 0 \quad \frac{\partial k^{**}}{\partial \theta} \geq 0.$$

The result states that, for any given market segmentation decision, the maximum affordable load for the extension task is lower for the extension policy than it is for the differentiation policy. Moreover, as the high-type segment size decreases, i.e., as  $\theta$  is increased, extension tasks with higher complexity (i.e., higher  $k$ ) can be supported by the value-creation strategies, extension, and differentiation.

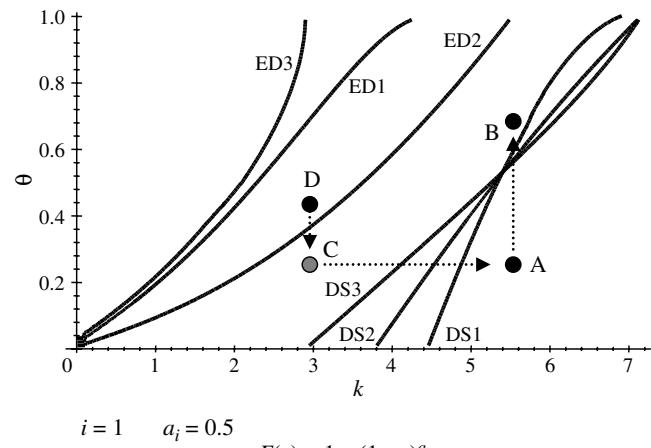
When the policy is standard, the profits are the same for all values of  $\theta$  (equivalently  $(q, p_H, p_L)$ ). Similarly, the profits are the same for all extension policies, because there is no market segmentation in practice and all customers receive the same service. This observation drives the following result, which points out the importance of optimally selecting  $\theta$ .

**REMARK 1.** For a given customer profile, a segmentation scheme  $\theta$  that leads to the selection of the differentiation policy as optimal ( $k^*(\theta) < k < k^{**}(\theta)$ ) by definition yields higher expected profits compared with a segmentation scheme  $\theta'$  that leads to the policies “extension” ( $k < k^*(\theta')$ ) or “standard” ( $k^{**}(\theta') < k$ ) as optimal.

This observation shows that making the segmentation decision cleverly would pay off, leading to higher profits. In other words, if we can make differentiation the optimal policy with a given customer profile through an appropriate choice of  $\theta$ , we are better off than if we target all customers or remain a cost center.

We illustrate the gains from optimizing the market segmentation decision  $\theta$  with a numerical example shown in Figure 2, where the curves  $k^*(\theta)$  and  $k^{**}(\theta)$

**Figure 2 Policy Choice for Different Customer Profiles**



$$\begin{array}{ll} i = 1 & a_i = 0.5 \\ i = 2 & a_i = 1 \\ i = 3 & a_i = 2 \end{array} \quad F(x) = 1 - (1-x)^{a_i} \quad R = 10 \quad C_S = 0.1 \quad c = 100 \quad \mu = 20 \quad \lambda = 10$$

are plotted for three success probability density functions. Note that all curves have a positive slope as indicated by Proposition 4. In this figure, the optimal policy is extension in the regions to the left of the lines  $ED_i$ , standard on the right of the lines  $DS_i$  and differentiation in between these two lines for each density function  $i$ .

The result stated in Remark 1 is illustrated by the move from point A to point B, and from D to C. With these moves into the differentiation region the profits are increased only by changing  $\theta$ . In addition, the effect of task complexity is illustrated by the horizontal moves. For example, whereas point C is in the differentiation region, when  $k$  increases we can move to a point such as A, where the standard policy is optimal and no value is generated. Then an optimal decision would be a move to a point such as B by increasing  $\theta$ .

To sum up, we have made three observations. First, the differentiation policy is potentially the most profitable policy. Second, the policy choice hinges on the market segmentation decision, so to achieve the maximum profits (using a differentiation policy if feasible),  $\theta$  should be chosen optimally. Finally, the optimal  $\theta$  choice is a function of  $k$ , representing the content or complexity of the extension task, and the success probability density function. Given these observations, the next question we address is: What is the optimal value for  $\theta$ , and how do we enforce it, given the private information of the server?

## 5.2. Market Segmentation Decision

Up to this point, the analysis is done taking the market segmentation variable  $\theta$  as a parameter. In this section,  $\theta$  is a decision variable, both for the principal (the manager) and the agent (the server). Recall that this represents the case of an integrated firm with a sophisticated manager (she) and server (he). In this setting,  $\theta = 0$  indicates choice of the extension policy,  $\theta = 1$  the standard policy, and  $0 < \theta < 1$  implies the differentiation policy. The manager declares her  $\theta$  decision. However, the server can choose another  $\theta$  without the manager observing it, given the private information he has about the customers and his objective of maximizing his own utility. Therefore, the contract  $(\alpha_1, \alpha_2)$  should be incentive compatible for the marginal customer (who has success probability  $\theta$ ) rather than the average customer for the particular segment, unlike the previous analysis of §4. There is only one incentive compatibility constraint (IC) that imposes indifference between offering standard and extended service for this marginal customer. The reservation utility for the agent is taken to be zero.

First, let us define

$$p_H(\theta) = \frac{\int_{\theta}^1 xf(x) dx}{\int_{\theta}^1 f(x) dx}, \quad q(\theta) = \int_{\theta}^1 f(x) dx,$$

$$W(\theta) = \frac{\lambda}{\mu} \left( \frac{((\mu - k)^2 + k(2\mu - k) \int_{\theta}^1 f(x) dx)}{((\mu - k)(\mu - \lambda) - \lambda k \int_{\theta}^1 f(x) dx)(\mu - k)} \right).$$

Then the optimization problem that the principal solves is as follows:

$$\begin{aligned} \max_{\theta, \alpha_1, \alpha_2} E[\Pi] &= \lambda(1 - \alpha_2)Rp_H(\theta)q(\theta) \\ &\quad - \lambda\alpha_1 \left( \frac{q(\theta)}{\mu - k} + \frac{(1 - q(\theta))}{\mu} \right) - \lambda c W(\theta), \\ \text{s.t.} \\ \alpha_2 R q(\theta) p_H(\theta) + \alpha_1 \left( \frac{q(\theta)}{\mu - k} + \frac{(1 - q(\theta))}{\mu} \right) - C_S q(\theta) &\geq 0, \quad (\text{IR}) \\ \alpha_2 R \theta + \alpha_1 \frac{1}{\mu - k} - C_S &= \alpha_1 \frac{1}{\mu}. \quad (\text{IC}) \end{aligned}$$

Because the efficient solution satisfies the individual rationality constraint as an equality, for any given

value of  $\theta$  the optimal contract  $(\alpha_1, \alpha_2)$  is at the intersection of the two constraints, (IR) and (IC), which is found as

$$\alpha_1 = -C_S q \frac{(p_H - \theta)(\mu - k)\mu}{-kq(p_H - \theta) + \theta(\mu - k)},$$

$$\alpha_2 = C_S \frac{\mu - k}{R(-kq(p_H - \theta) + \theta(\mu - k))}.$$

We see that there is a unique optimal (linear) contract for the optimal market segmentation scheme, as opposed to the infinitely many contracts in the case in which  $\theta$  was not a decision variable. Moreover, this contract requires a punishment for the service time component.

This optimal contract is a function of the variables determining the system capacity and the complexity of the value-creation task,  $\mu$  and  $k$ , as well as the variables defining the server disutility and customer characteristics,  $C_S$  and  $f(\cdot)$ . The comparative statics analysis shows the following result.

**PROPOSITION 5.** *Assume a market segmentation decision  $\theta$  and a segment size  $q$ . For two customer pools  $X, Y$  with densities of success probabilities  $f(p), g(p)$ , if  $X \geq_{st} Y$  for  $p \geq \omega \geq \theta$  (i.e.,  $1 - F(p) \geq 1 - G(p) \forall p \in [\omega, 1]$ ), then the magnitude of the contract rates for both task dimensions is higher for  $X$  than it is for  $Y$ , i.e.,  $\alpha_1^{(X)} < \alpha_1^{(Y)}$  and  $\alpha_2^{(X)} > \alpha_2^{(Y)}$ .*

This result compares two customer pools, one with more concentration on the high end than the other, and shows that firms operating in a high-end market should give more commission rate for the revenue and more punishment for the service time. In a high-end market, in comparison with a low-end market, a higher average success probability imposes high punishment in service time to balance for the commission. This, in turn, implies a higher commission rate for the revenue because for the marginal customer the success probability is the same for both markets (which is  $\theta$ ).

Having found the optimal  $(\alpha_1, \alpha_2)$ , the program reduces to finding the  $\theta$  that maximizes the objective value, given this contract. Proposition 6 characterizes the unique optimum for the market segmentation problem.

**PROPOSITION 6.** *Let  $f(x) > 0$  for all  $x \in [0, 1]$  and  $\theta'$  be the solution to*

$$(R\theta - C_S)((\mu - k)(\mu - \lambda) - \lambda k q(\theta))^2 \frac{1}{(2\mu - k - \lambda)\lambda k} - c = 0.$$

Then the optimal  $\theta = \theta^*$  is found as  $\theta^* = \theta'$  if  $0 \leq \theta' \leq 1$ ,  $\theta^* = 1$  if  $\theta' > 1$ .

The optimal cut-off point for the market segmentation problem,  $\theta^*$ , equates the expected marginal revenues from offering extended service to a customer with success probability  $\theta$  to the cost of expected congestion with that definition of the high-type segment. So for a customer with success probability  $p = \theta^*$ , we are indifferent between offering standard service and extended service. Then the long-run average profit rate would be as follows:

$$E[\Pi | \theta^*] = \lambda q(\theta^*)(p_H(\theta^*)R - C_S) - \lambda c W(\theta^*). \quad (14)$$

The optimality condition implies that for  $C_S > 0$ ,  $\theta = 0$  would never be optimal. In other words, it is never optimal to offer extended service to *all* customers. There is always some portion of customers for whom even the direct cost of effort would not be paid off if the extended service is offered. This implies that the strategy of targeting all customers is only possible in settings in which there is the type of server that has  $C_S = 0$ . In a setting where the value-creation activity being pursued is cross-selling, this implies that an optimized cross-selling initiative has to be targeted. On one hand, the common approach of attempting a cross-sell on all customers is clearly not supported when servers show the slightest disutility with respect to the sales activity. On the other hand, a very large unit congestion cost  $c$  or a very high utilization rate  $\lambda/\mu$  satisfies the following condition and makes  $\theta^* = 1$ :

$$(R - C_S) \frac{((\mu - k)(\mu - \lambda))^2}{(2\mu - k - \lambda)\lambda k} \leq c. \quad (15)$$

If the condition given in (15) is satisfied, then the standard policy becomes optimal.

Next, we provide some comparative statics results for the optimum.

**PROPOSITION 7.** (a) Optimal  $\theta$  increases with  $k$ , and  $\lambda/\mu$ . i.e.,  $\partial\theta^*/\partial k > 0$  and  $\partial\theta^*/\partial(\lambda/\mu) > 0$ .

(b) Given two customer pools  $X$ ,  $Y$  with densities of success probabilities  $f(p)$ ,  $g(p)$ , if  $X \geq_{st} Y$  (i.e.,  $1 - F(p) \geq 1 - G(p) \forall p \in [0, 1]$ ), then  $\theta_X^* \geq \theta_Y^*$ .

This proposition shows the environmental conditions that would affect the optimal market segmentation decision  $\theta^*$ . In the first part, we show that an

increase in the complexity of the extension task and an increase in system utilization would increase  $\theta^*$ . This is an intuitive result, given that both factors would increase the load on the system and thus would make the extended service more difficult to offer for the customers with low revenue-generation potentials.

The second part of the proposition shows the effect of the customer profile, represented by the success probability distributions. We compare two customer pools, one with more concentration on the high end than the other. This result shows that for the high-end markets the definition of a high-type customer will be upgraded, i.e., one would be more “picky” selecting the customers to offer service extension. The intuition behind this result comes from the issue of resource allocation among the customers. The customers with high success probabilities would be given the priority when allocating the scarce service time. When there are many customers at the high end, the targeted high-type market size is filled with high-end customers, and this results in selectiveness in defining customer types.

Another implication of this result is that if we consider a firm with constant capacity that could operate in two markets under the conditions stated in Proposition 7(b), the congestion level in the high-end market is expected to be higher compared with the low-end market. Increasing selectiveness increases the marginal revenues expected from a high-type customer, so the manager can afford to face the costs of higher congestion levels in the system. This result has implications on the customer experiences in different markets. For any customer with a given value of success probability  $p$ , being in a low-end market would be preferable to being in a high-end market for two reasons. First, the expected waiting time is lower. Second, the cut-off level of success probability,  $\theta$ , to receive a higher service level is lower, so that there is a higher chance to receive high-level service in a low-end market. Hence, we show the importance of one’s relative position in the population with respect to the service received.

## 6. Summary of Model

### Findings—Concluding Remarks

A stylized model that captures the key levers for value-creation strategy choice is presented and ana-

lyzed. The model is unique in that it combines value-creation and process-related issues, thereby providing a coherent framework to analyze sales initiatives such as cross-selling or service-level differentiation strategies. The analysis characterizes under what conditions a firm would choose to remain as it is or would choose to attempt value creation. The latter choice is further elaborated in terms of a choice between a differentiation strategy and one of “targeting all customers.” Value-creation strategies are favored by moderately loaded systems in which the extension task’s complexity does not push capacity utilization to a maximum, and the expected revenue from the extension task is high enough to compensate for the extra costs of congestion. When unit congestion cost is relatively high compared with revenues, service-level differentiation is preferred to targeting all customers.

For integrated firms that optimize  $\theta$ , service-level differentiation is the optimal choice. High congestion costs, high task complexity for the extension task, or high system utilization lead to the status quo choice where no value creation is pursued. Unless servers experience no disutility for the extension task, “targeting all customers” is never optimal, even if congestion cost is low. Firms that would like to pursue this strategy would need to select employees who enjoy the extension task, or provide additional automated support to their employees in an effort to reduce this disutility.

The analysis of market segmentation makes two important points. The first one is that a change in market segmentation ( $\theta$ ) can imply a change in policy choice. Thus we can talk about better  $\theta$  choices that lead to higher profitability. More specifically, we find that choosing and sustaining a service-level differentiation strategy might hinge on the market segmentation decision. Firms that can operate in the integrated mode—where, for example, marketing and operations jointly optimize  $\theta$ —are shown to be clearly better off in terms of achieving this profitability. The second point is that the optimal market segmentation decision depends on the distribution of revenue generation in the customer base. This, in turn, implies that actions that change the underlying customer profitability distribution shape such as—better-targeted sales, for example—can instigate a shift in policy. Thus, the second point emphasizes the need for a

prospective type of analysis as opposed to the currently prevailing retrospective analysis in customer value estimation.

The analysis herein uses the probability of revenue generation from a specific extension task as the measure of customer profitability. The parameter  $k$ , however, captures the difference in needs (in terms of service extension to create value) between the  $H$ - and  $L$ -type customers. As a result, the  $\theta$  choice discussed in §5.1 can be seen as a segmentation decision that combines value and customer needs concerns. Our analysis demonstrates how one can perform value coupled with needs-based segmentation, where the  $\theta$  choice allows one to make a trade-off between customer profitability and needs. This type of segmentation decision is in line with the recommendations in Carroll and Tadikonda (1997) and Giltner and Ciolfi (1999), who critique pure customer profitability—based segmentation schemes.

The market segmentation analysis clearly illustrates that even if firms have very reliable individual customer data and can effectively estimate individual customer value, unless the aggregation decision that determines the segmentation scheme is done correctly value cannot be maximized. As shown by the analysis, a correct aggregation decision needs to make the trade-offs between operational performance, the breadth of the value-creation activity (i.e., segment size), and the depth (i.e., profitability potential) of these types of activities. An organization that acts as what we labeled as the functional organization will not be able to reap all the benefits of a value-creation initiative. In other words, functioning as an integrated firm is essential to success.

Once a strategy choice is made, we illustrate how the desired strategy can be implemented through the design of appropriate incentive contracts. For the setting in which  $\theta$  is not optimized, a set of optimal linear incentive contracts is explicitly characterized for each strategy. We show that when the two tasks being considered have opposite performance effects (such as the standard service and extension activities are assumed to have in this analysis), then optimal linear contracts can involve punishments, i.e., disincentives, for one of the dimensions. Furthermore, we find that incentive payments for these two tasks (for example, service and sales) depend on each other, and providing incentives for only one dimension, as is frequently observed with

sales-based incentives, can lead to undesirable behavior. That is, incentive schemes which have only value-generation-related incentives (for example sales), are not always optimal. In the setting where we have a sophisticated manager and a sophisticated server who optimize the market segmentation decision, we show that there is a unique optimal linear contract. This contract is clearly a function of  $\theta$ , illustrating the close ties between market segmentation, process design, and incentives in value-creation initiatives.

The model illustrates that success in implementation of a service-level differentiation program depends on proper incentive contract design, which requires good parameter estimation. The policy implementation is particularly susceptible to misunderstanding employee preferences (characterized here by effort cost) and is not robust in settings where distinctly different customer segments cannot be formed. Although our analysis assumes that there are only two customer segments ( $H$  and  $L$ ), this last result suggests that as companies increase the number of segments that they define for their service-level differentiation strategies (for example, cases with three and five segments can be found in retail banking), implementations become less robust. For integrated firms, the uniqueness of the optimal incentive contract illustrates the high sensitivity of these types of value-creation initiatives to incentive design.

A final remark can be made for firms that plan to initiate cross-sell type of value-creation programs without reconfiguring their capacity. Considering the fact that a customer's experience (in terms of waiting time and service level in our setting) will impact their satisfaction, one can expect firms that have targeted a higher-end customer pool to experience more customer dissatisfaction associated with their value-creation initiatives if the service capacity is kept the same. This suggests that firms with relatively higher-end customer pools should be more careful in implementing programs such as cross-selling or add-on sales, and points out the importance of considering the capacity implications of these programs. Future research should focus on explicitly modeling the customer experience in value-creation initiatives.

### Acknowledgments

The authors gratefully acknowledge suggestions from two anonymous referees that significantly improved the paper.

They also thank Beril Toktay and Christoph Loch for helpful comments on an earlier draft of the paper.

### Appendix

#### Problem Formulations and Solutions for Incentive Contract Problems of Policies 1 (Standard) and 3 (Extension)

##### Policy 1: Standard.

$$\begin{aligned} \max_{(\alpha_1, \alpha_2)} E[\Pi^P(\alpha_1, \alpha_2)] &= -\alpha_1 \frac{\lambda}{\mu} - \lambda T(0, 0), \\ \text{s.t. } \frac{\alpha_1}{\mu} &\geq 0, \\ \frac{\alpha_1}{\mu - k} + \alpha_2 Rp_H - C_S &\leq \frac{\alpha_1}{\mu}, \\ \frac{\alpha_1}{\mu - k} + \alpha_2 Rp_L - C_S &\leq \frac{\alpha_1}{\mu}, \\ \alpha_2 &\leq 1. \end{aligned} \quad (\text{IR}) \quad (\text{ICH}) \quad (\text{ICL})$$

$\alpha_1^* = 0$	$\alpha_2^* \in [0, C_S/Rp_H]$
$w^{FB} = (w_H, w_L) = (0, 0)$	
$E[\Pi^*(\alpha_1, \alpha_2)   (0, 0)] = -\lambda T(0, 0) = E[\Pi^{FB}   (0, 0)]$	

##### Policy 3: Extension.

$$\begin{aligned} \max_{(\alpha_1, \alpha_2)} E[\Pi^P(\alpha_1, \alpha_2)] &= \lambda(1 - \alpha_2)R(qp_H + (1 - q)p_L) \\ &\quad - \alpha_1 \frac{\lambda}{\mu - k} - \lambda T(1, 1), \\ \text{s.t. } \frac{\alpha_1}{\mu - k} + \alpha_2 R(qp_H + (1 - q)p_L) - C_S &\geq 0, \\ \frac{\alpha_1}{\mu - k} + \alpha_2 Rp_H - C_S &\geq \frac{\alpha_1}{\mu}, \\ \frac{\alpha_1}{\mu - k} + \alpha_2 Rp_L - C_S &\geq \frac{\alpha_1}{\mu}, \\ \alpha_2 &\leq 1. \end{aligned} \quad (\text{IR}) \quad (\text{ICH}) \quad (\text{ICL})$$

$\alpha_1^* \in \left( -(R(qp_H + (1 - q)p_L) - C_S)(\mu - k), -qC_S \frac{(p_H - p_L)(\mu - k)\mu}{\mu p_L - k(qp_H + (1 - q)p_L)} \right)$ ,
$\alpha_2^* = \frac{C_S}{R(qp_H + (1 - q)p_L)} - \alpha_1^* \frac{1}{(\mu - k)R(qp_H + (1 - q)p_L)}$ ;
$w^{FB} = (w_H, w_L) = (C_S, C_S)$ ,
$E[\Pi^*(\alpha_1, \alpha_2)   (1, 1)] = qRp_H + (1 - q)Rp_L - C_S - T(1, 1) = E[\Pi^{FB}   (1, 1)]$ .

PROOF OF PROPOSITION 2. Differentiation is preferred to extension only if

$$qRp_H - qC_S - T(1, 0) \geq qRp_H + (1 - q)Rp_L - C_S - T(1, 1),$$

which is equivalent to

$$T(1, 1) - T(1, 0) \geq (1 - q)(Rp_L - C_S). \quad (16)$$

This condition compares the costs and profits obtained by offering extended service to low-type customers. Below, we rewrite condition (16) with some new notation. We also use  $T(\cdot) = cW(\cdot)$  as defined in the model. Let

$$\Delta W_L = W(1, 1) - W(1, 0), \quad \Delta T_L = T(1, 1) - T(1, 0),$$

$$\text{and } \Delta R_L = (1-q)(Rp_L - C_S).$$

Then we can rewrite (16) as  $\Delta T_L = c\Delta W_L \geq \Delta R_L$ , which is equivalent to

$$c \geq \frac{\Delta R_L}{\Delta W_L}$$

$$= c^* = [(1-q)(Rp_L - C_S)] / [(-2\mu + 2q\mu + k + \lambda - q\lambda - qk)\lambda]$$

$$\cdot (k/((\mu - \lambda - k)\mu(\mu^2 - \mu k - \lambda\mu + \lambda k - \lambda qk))].$$

After simplification, we get

$$c^* = \frac{\mu}{\lambda} \frac{(\mu - k - \lambda)}{(2\mu - k - \lambda)} (Rp_L - C_S) \left( \frac{(\mu - \lambda)(\mu - k)}{k} - \lambda q \right).$$

Selection between differentiation and standard policies is done in a similar way, this time comparing the costs and profits of offering extended service to the high-type customers. The condition to determine the critical unit congestion cost  $c^{**}$  is  $qRp_H - qC_S - T(1, 0) \geq -T(0, 0)$ , or equivalently  $qRp_H - qC_S \geq T(1, 0) - T(0, 0)$ . This inequality can be represented as a comparison of marginal gains ( $\Delta R_H$ ) and losses ( $\Delta T_H$ ), i.e.,  $\Delta R_H \geq \Delta T_H = c\Delta W_H$ .

Thus the breakeven value of unit congestion cost  $c^{**}$  that makes us indifferent between the two policies differentiation and standard is found as follows:

$$c^{**} = \frac{\Delta R_H}{\Delta T_H}$$

$$= [\{q(Rp_H - C_S)\} / [\lambda qk((2\mu - \lambda - k)) / ((\mu^2 - \mu k - \lambda\mu + \lambda k - \lambda qk)(\mu - k)(\mu - \lambda))],$$

$$c^{**} = \frac{(\mu - k)}{\lambda} \frac{(\mu - \lambda)}{(2\mu - \lambda - k)} (Rp_H - C_S)$$

$$\cdot \left[ \frac{(\mu - \lambda)(\mu - k)}{k} - \lambda q \right]. \quad \square$$

**PROOF OF PROPOSITION 4.** In this proof, we use the two basic assumptions stated below, which imply profitability of all customers considering direct costs only and stability for all policies.

$$(i) \quad Rp_L - C_S > 0. \quad (17)$$

$$(ii) \quad \frac{\lambda}{\mu - k} < 1. \quad (18)$$

Define  $k^*$  and  $k^{**}$  implicitly as follows:

$$\psi(\theta, k) = c^*(\theta, k) - c = 0 \quad \text{for } k = k^*,$$

$$\phi(\theta, k) = c^{**}(\theta, k^{**}) - c = 0 \quad \text{for } k = k^{**}.$$

First we show that  $\psi(\theta, k)$  and  $\phi(\theta, k)$  are strictly decreasing in  $k$ :

$$\begin{aligned} \frac{\partial}{\partial k} \psi(\theta, k) &= \mu^2 (Rp_L - C_S) \\ &\cdot (2\lambda^2 k - 4\mu\lambda k - 2\mu^3 + 2\mu^2 k + 5\lambda\mu^2 - 4\lambda^2 \mu \\ &+ \lambda^3 + \lambda qk^2) / (\lambda k^2 (-2\mu + k + \lambda)^2) \\ &= -(Rp_L - C_S) \left\{ \frac{\mu^2}{\lambda(2\mu - k - \lambda)^2} \right. \\ &\cdot \left( \frac{(\mu - \lambda)(\mu - k)}{k} - \lambda q \right) \left\} \right. \\ &- (Rp_L - C_S) \left\{ \frac{\mu^2(\mu - \lambda)}{\lambda k^2} \frac{(\mu - k - \lambda)}{(2\mu - k - \lambda)} \right\} < 0. \end{aligned} \quad (19)$$

It is easy to see that the above expression is negative because  $(Rp_L - C_S) > 0$  by (17); also, for both terms,

$$\frac{\mu^2}{\lambda(2\mu - k - \lambda)^2} \left( \frac{(\mu - \lambda)(\mu - k)}{k} - \lambda q \right) > 0 \quad \text{and}$$

$$\frac{\mu^2}{\lambda(2\mu - k - \lambda)^2} \left( \frac{(\mu - \lambda)(\mu - k)}{k} - \lambda q \right) > 0$$

hold by (18).

Similarly,  $\partial\phi(\theta, k)/\partial k < 0$  as the expression found below is negative by our assumptions.

$$\begin{aligned} \frac{\partial\phi(\theta, k)}{\partial k} &= ((-\mu + \lambda)(Rp_H - C_S)(\mu\lambda k^2 + \lambda^2 qk^2 + 2\mu^2 k\lambda \\ &+ 2\mu^4 - 2\mu^3 k - 3\mu^3 \lambda + \lambda^2 \mu^2 - \lambda^2 k^2 - \mu\lambda qk^2)) \\ &/ \lambda k^2 (-2\mu + k + \lambda)^2 \\ &= -(\mu - \lambda)(Rp_H - C_S) \\ &\cdot (\mu^2(2(\mu - k) - \lambda)(\mu - \lambda) + \lambda k^2(\mu - \lambda)(1 - q)) \\ &/ \lambda k^2 (-2\mu + k + \lambda)^2 < 0. \end{aligned} \quad (20)$$

Now we can show the results in Proposition 4.

Part I:  $k^{**} > k^*$  for any given  $\theta$ . This follows from the fact that  $\phi(\theta, k) > \psi(\theta, k)$  and both  $\phi(\theta, k)$  and  $\psi(\theta, k)$  are decreasing in  $k$  (as shown in (19) and (20)). Then the solution of  $\phi(\theta, k) = c$  must be greater than the solution of  $\psi(\theta, k) = c$ .

Part II:  $k^{**}$  and  $k^*$  are increasing in  $\theta$ . Recall that  $k^*$  is implicitly defined by the equation  $\psi(\theta, k) - c = 0$ , so we use implicit differentiation to find  $\partial k^*/\partial\theta$ :

$$\frac{\partial k^*}{\partial \theta} = -\frac{\partial(\psi(\theta, k) - c)}{\partial \theta} \left[ \frac{\partial(\psi(\theta, k) - c)}{k} \right]^{-1} \Big|_{k=k^*}. \quad (21)$$

The first part of (21) is found to have a  $(-)$  sign as shown below.

$$\begin{aligned} -\frac{\partial(\psi(\theta, k) - c)}{\partial \theta} &= -\frac{d}{d\theta} \psi(\theta, k) \\ &= -\left( \frac{\partial\psi(\theta, k)}{\partial q} \frac{\partial q}{\partial \theta} + \frac{\partial\psi(\theta, k)}{\partial p_L} \frac{\partial p_L}{\partial \theta} \right) < 0. \end{aligned} \quad (22)$$

To see that (22) holds, we check all the terms as follows.

(1) The first term of (22) is (+) because  $\partial\psi(\theta, k)/\partial q < 0$  (by Proposition 3) and  $\partial q/\partial\theta \leq 0$  by definition of  $q$ .

(2) The second term of (22) is (+) because  $\partial\psi(\theta, k)/\partial p_L > 0$  (by Proposition 3) and  $\partial p_L/\partial\theta \geq 0$  by definition of  $p_L$ .

The second part of (21) has a (−) sign, also, as shown in Equation (19). So,  $\partial(\psi(\theta, k) - c)/k < 0$ . Therefore,

$$\begin{aligned} \frac{\partial k^*}{\partial\theta} &= -\frac{\partial(\psi(\theta, k) - c)}{\partial\theta} \left[ \frac{\partial(\psi(\theta, k) - c)}{k} \right]^{-1} \Big|_{k=k^*} \\ &> 0 \quad \text{by Equations (19) and (22).} \end{aligned} \quad (23)$$

The same analysis is repeated to show that  $\partial k^{**}/\partial\theta > 0$ :

$$\frac{d}{d\theta}\phi(\theta, k) = \frac{\partial\phi(\theta, k)}{\partial q} \frac{\partial q}{\partial\theta} + \frac{\partial\psi(\theta, k)}{\partial p_H} \frac{\partial p_H}{\partial\theta} > 0 \quad (24)$$

$$\begin{aligned} \frac{\partial k^{**}}{\partial\theta} &= -\frac{\partial(\phi(\theta, k) - C_S)}{\partial\theta} \left[ \frac{\partial(\phi(\theta, k) - C_S)}{k} \right]^{-1} \Big|_{k=k^{**}} \\ &> 0 \quad \text{by Equations (20) and (24).} \quad \square. \end{aligned} \quad (25)$$

**PROOF OF PROPOSITION 5.** Given  $f(p), g(p), \theta$  such that

$$\begin{aligned} q &= \int_{\theta}^1 f(p) dp = \int_{\theta}^1 g(p) dp \quad \text{and} \quad \exists \omega, \text{ s.t.} \\ \int_{\omega}^1 vf(p) dp &\geq \int_{\omega}^1 g(p) dp, \quad \forall \omega \geq \theta; \\ \text{then} \quad (p_H(\theta))^X &\geq (p_H(\theta))^Y \end{aligned} \quad (26)$$

is true. Then, because

$$\begin{aligned} \frac{\partial\alpha_1}{\partial p_H} &= -C_S q (\mu - k)^2 \mu \frac{\theta}{(-kqp_H + kq\theta + \theta\mu - \theta k)^2} < 0, \\ \frac{\partial\alpha_2}{\partial p_H} &= C_S (\mu - k) k \frac{q}{R(-kqp_H + kq\theta + \theta\mu - \theta k)^2} > 0, \end{aligned}$$

$\alpha_1^{(X)} < \alpha_1^{(Y)}$  and  $\alpha_2^{(X)} > \alpha_2^{(Y)}$  follows from (26).  $\square$

**PROOF OF PROPOSITION 6.** The necessary condition for optimality is found as follows:

$$\begin{aligned} \frac{\partial}{\partial\theta}(E[\Pi]) &= f(\theta) \left( -(R\theta - C_S) \right. \\ &\quad \left. + \frac{(2\mu - k - \lambda)c\lambda k}{((\mu - k)(\mu - \lambda) - \lambda k \int_{\theta}^1 f(x) dx)^2} \right) = 0. \end{aligned}$$

For  $f(\theta) > 0$ , it reduces to the condition stated in Proposition 6. So  $\theta'$  is a local optimum. Moreover,

$$\frac{\partial}{\partial\theta}(E[\Pi]) \geq 0 \quad \text{for } \theta \leq \theta', \quad \text{and} \quad \frac{\partial}{\partial\theta}(E[\Pi]) \leq 0 \quad \text{for } \theta \geq \theta'.$$

Thus, we conclude that  $\theta'$  is a global maximum. If  $\theta' > 1$ , then  $\frac{\partial}{\partial\theta}(E[\Pi])|_{\theta=1} > 0$  and the optimum is at the boundary, i.e.,  $\theta^* = 1$ .  $\square$

**PROOF OF PROPOSITION 7.** For a given density  $f(\cdot)$  of success probabilities, define

$$\begin{aligned} \xi_f(\theta) &= (R\theta - C_S) \left( (\mu - k)(\mu - \lambda) - \lambda k \int_{\theta}^1 f(x) dx \right)^2 \\ &\quad \cdot \frac{1}{(2\mu - k - \lambda)\lambda k} - c = 0 \quad \text{at optimum.} \end{aligned} \quad (27)$$

We use the following four results in the proof:

$$\begin{aligned} \frac{\partial\xi_f(\cdot)}{\partial\theta} &= (((\mu - k)(\mu - \lambda) - \lambda k(1 - F(\theta))) (R((\mu - k)(\mu - \lambda) \\ &\quad - \lambda k(1 - F(\theta)) + 2\lambda k f(\theta) (R\theta - C_S))) / ((2\mu - k - \lambda)\lambda k) \\ &> 0, \end{aligned} \quad (28)$$

$$1 - F(\theta) \geq 1 - G(\theta) \Rightarrow \xi_f(\theta) \leq \xi_g(\theta), \quad (29)$$

$$\frac{\partial\xi_f(\cdot)}{\partial k} < 0 \quad (30)$$

as shown in the proof of Proposition 4

$$\frac{\partial\xi_f(\cdot)}{\partial(\lambda/\mu)} < 0. \quad (31)$$

(a) The results follow easily taking the derivatives using the implicit function theorem as follows:

$$\frac{\partial\theta^*}{\partial k} = -\frac{\partial\xi_f(\cdot)}{\partial k} \left[ \frac{\partial\xi_f(\cdot)}{\partial\theta} \right]_{\theta=\theta^*}^{-1} \geq 0 \quad \text{by (28) and (30).}$$

$$\frac{\partial\theta^*}{\partial(\lambda/\mu)} = -\frac{\partial\xi_f(\cdot)}{\partial(\lambda/\mu)} \left[ \frac{\partial\xi_f(\cdot)}{\partial\theta} \right]_{\theta=\theta^*}^{-1} \geq 0 \quad \text{by (28) and (31).}$$

(b) Given  $X, Y$  with densities  $f(x), g(x)$ , and  $X \geq_{st} Y$  defined as  $1 - F(\theta) \geq 1 - G(\theta)$  for all  $\theta \in [0, 1]$ , i.e.,  $\overline{F(\theta)} \geq \overline{G(\theta)}$   $\forall \theta \in [0, 1]$ . We can rewrite the optimality condition for any customer pool as follows:

$$\begin{aligned} \xi(\theta) &= (R\theta - C_S)((\mu - k)(\mu - \lambda) - \lambda k \overline{F(\theta)})^2 \\ &\quad \cdot \frac{1}{(2\mu - k - \lambda)\lambda k} - c = 0. \end{aligned} \quad (32)$$

If  $\theta_Y^*$  is optimal for  $Y$ ,  $\xi_g(\theta_Y^*) = 0$  by the optimality condition, stated in (32). We know that  $\xi_f(\theta) \leq \xi_g(\theta)$  (by (29)). Equivalently,

$$\xi_f(\theta_Y^*) < \xi_g(\theta_Y^*) = 0. \quad (33)$$

Then,  $\theta_X^* \geq \theta_Y^*$  because  $\theta$  should increase to make  $\xi_f(\theta_X^*) = 0$ , given the results  $\xi_f(\theta_Y^*) < 0$  (by (33)) and  $\partial\xi_f(\cdot)/\partial\theta > 0$  (by (28)).

The objective function for any  $\theta$  is then found as follows (let  $p_H(\theta) = p_H$ ,  $q(\theta) = q$  for ease in exposition):

$$E[\Pi|\theta] = (1 - \alpha_2) R p_H(\theta) q(\theta) - \alpha_1 (\mu - k q(\theta)) - c W(\theta)$$

$$= \left( 1 - \left( \frac{C_S}{R} \frac{\mu}{\mu\theta + kq(p_H - \theta)} \right) \right) R p_H q$$

$$\begin{aligned}
 & -\left(-qC_S \frac{p_H - \theta}{\theta\mu + kq(p_H - \theta)}\right)(\mu - kq) - cW(\theta) \\
 & = \frac{q\mu C_S p_H - q\theta\mu C_S + kq^2\theta C_S - kq^2 C_S p_H}{\theta\mu - kq\theta + kqp_H} - cW(\theta) \\
 & \quad + \frac{Rq\theta\mu p_H - q\mu C_S p_H - Rkq^2\theta p_H + Rkq^2 p_H^2}{\theta\mu - kq\theta + kqp_H} - cW(\theta) \\
 & = \frac{1}{\theta\mu - kq\theta + kqp_H} (\theta\mu - kq\theta + kqp_H) q(Rp_H - C_S) \\
 & \quad - cW(\theta) \\
 & = q(Rp_H - C_S) - cW(\theta).
 \end{aligned}$$

## References

- Akşin, O. Z., P. T. Harker. 1999. To sell or not to sell: Determining the trade-offs between service and sales in retail banking phone centers. *J. Service Res.* 2(1) 19–33.
- Andrews, B., H. Parsons. 1993. Establishing telephone-agent staffing levels through economic optimizations. *Interfaces* 23(2) 14–20.
- Basu, A. K., R. Lal, V. Srinivasan, R. Staelin. 1985. Salesforce compensation plans: An agency theoretic perspective. *Marketing Sci.* 4(4) 267–291.
- Benjamin, K. 2001. Customer relationship management: Building a strategy. Special report: Understanding CRM, *Financial Times*, <http://specials.fl.com>.
- Berger, P. D., N. I. Nasr. 1998. Customer lifetime value: Marketing models and applications. *J. Interactive Marketing* 12(1) 17–26.
- Bettencourt, L. A., K. Gwinner. 1996. Customization of the service experience: The role of the frontline employee. *Internat. J. Service Indust. Management* 7(2) 3–20.
- Borst, S., Mandelbaum, A., M. I. Reiman. 2004. Dimensioning large call centers. *Oper. Res.* 52(1) 17–34.
- Bradford, R. M. 1996. Pricing, routing, and incentive compatibility in multiserver queues. *Eur. J. Oper. Res.* 89(2) 226–236.
- Buzacott, J. A. 1996. Commonalities in reengineered business processes: Models and issues. *Management Sci.* 42(5) 768–782.
- Buzacott, J. A. 2002. The impact of worker differences on production system output. *Internat. J. Production Econom.* 78 37–44.
- Carroll, P., M. Tadikonda. 1997. Customer profitability: Irrelevant for decisions? *Banking Strategies* 75(6) 77–82.
- Coughlan, A. 1993. Salesforce compensation: A review of MS/OR advances. J. Eliashberg, G. L. Lilien, eds. *Handbooks in Operations Research and Management Science*, Vol. 5. Elsevier, Oxford, U.K., 611–652.
- Economist, The. 2001. Keeping the customers satisfied. 360(8230) 9–10 (July 14).
- Farley, J. U. 1964. An optimal plan for salesmen's compensation. *J. Marketing Res.* 1(2) 39–43.
- Fisher, A. 2001. Customer relationship management: The personal touch. Special report: Understanding CRM. *Financial Times* (November 26).
- Foster, G., M. Gupta, L. Sjoblom. 1996. Customer profitability analysis: Challenges and new directions. *J. Cost Management* 10(Spring) 5–17.
- Fridgeirsdóttir, K., S. Chiu. 2001. Pricing and marketing decisions in delay sensitive markets. Working paper, Department of Management Science and Engineering, Stanford University, Stanford, CA.
- Gilbert, S. M., Z. K. Weng. 1998. Incentive effects favor nonconsolidating queues in a service system: The principal-agent perspective. *Management Sci.* 44(12) 1662–1668.
- Giltner, R., R. Ciolfi. 1999. Rx for segmentation. *Banking Strategies* 75(6) 43–50.
- Grossman, S., O. Hart. 1983. An analysis of the principal-agent problem. *Econometrica* 51 7–45.
- Holmstrom, B., P. Milgrom. 1991. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *J. Law, Econom., Organ.* 7 24–53.
- Joseph, K., A. Thevarajan. 1998. Monitoring and incentives in sales organizations: An agency-theoretic perspective. *Marketing Sci.* 17(2) 107–123.
- Kalai, E., M. I. Kamien, M. Rubinovitch. 1992. Optimal service speeds in a competitive environment. *Management Sci.* 38(8) 1154–1163.
- Lal, R., V. Srinivasan. 1993. Compensation plans for single and multi-product salesforces: An application of the Holmstrom-Milgrom model. *Management Sci.* 39(7) 777–793.
- Loch, C. 1998. Operations management and reengineering. *Eur. Management J.* 16(3) 306–316.
- Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Oper. Res.* 38(5) 870–883.
- Mulhern, F. J. 1999. Customer profitability analysis: Measurement, concentration, and research directions. *J. Interactive Marketing* 13(1) 25–40.
- Newman, R. 1984. A conjoint analysis in outpatient clinic preferences. *J. Health Care Marketing* 4(1) 41–49.
- Niraj, R., M. Gupta, N. Chakravarthi. 2001. Customer profitability in a supply chain. *J. Marketing* 65 1–16.
- Pinker, E., R. Shumsky. 2000. The efficiency-quality trade-off of cross-trained workers. *Manufacturing Service Oper. Management* 2(1) 32–48.
- Plambeck, E., S. Zenios. 2000. Performance-based incentives in a dynamic principal-agent model. *Manufacturing Service Oper. Management* 2(3) 240–263.
- Powell, S. G. 2000. Specialization, teamwork, and production efficiency. *Internat. J. Production Econom.* 67 205–218.
- Pullman, M., W. Moore. 1999. Optimal service design: Integrating marketing and operations perspectives. *Internat. J. Service Indust. Management* 10(2) 239–260.
- Shumsky, R., E. Pinker. 2003. Gatekeepers and referrals in services. *Management Sci.* 49(7) 839–856.
- Storbacka, K. 1997. Segmentation based on customer profitability: A retrospective analysis of retail bank customer bases. *J. Marketing Management* 13 479–492.
- Van Mieghem, J. A. 2000. Price and service discrimination in queueing systems: Incentive compatibility of  $G\mu$  scheduling. *Management Sci.* 46(9) 1249–1267.
- Verna, R., G. Thompson. 1999. Managing service operations based on customer preferences. *Internat. J. Oper. Production Management* 19(9) 891–908.