A Service Provider’s Elicitation of Its Customers’ Demand Distributions by a Price Mechanism

Gregory Dobson and Phillip J. Lederer
Simon School, University of Rochester, Rochester, NY 14627, e-mail: dobson@simon.rochester.edu, lederer@simon.rochester.edu

Lawrence W. Robinson
Johnson School, Cornell University, Ithaca, NY 14853-6201, e-mail: lwr2@cornell.edu

ABSTRACT
In a service environment a service provider needs to determine the amount and kinds of capacity to meet customers’ needs over many periods. To make good decisions, she needs to know the probability distribution of her customers’ demand in each period. We study a situation in which customers’ demand for a given service is random in each period, but inelastic, or modeled well by this assumption, and cannot be delayed to the next period. This article presents a mechanism that allows a service provider to learn the distribution of a customer’s demand by offering him a set of contracts through which he can partially prepay for future service for a reduced cost for units of service based on anticipated needs. We describe the form of a set of contracts that will cause the customer to reveal his demand distribution as he minimizes his expected costs. To justify the effort of organizing and offering contracts, we present an application that demonstrates the cost savings to the service provider with better capacity planning using the truthfully elicited distribution. [Submitted: December 23, 2010. revisions: July 11, 2011, August 29, 2011. Accepted: September 1, 2011.]

Subject Areas:

INTRODUCTION
We study a situation in which a service provider must acquire capacity and then deliver service to a number of customers whose demands are uncertain in each period. Given fixed revenues, the service provider’s problem is to minimize the cost of capacity while insuring the customers’ supply. The structure of the problem is similar to that of inventory planning by a manufacturer for its customers. However, there are two reasons why service capacity planning is more critical than planning inventory. First, service firms cannot use safety stock to guard against uncertainty; only extra, safety, capacity can be employed. All things being equal, safety capacity is more costly than safety stock because undemanded safety stock
A Service Provider’s Elicitation

Q3

can be used in later periods, but service capacity disappears. Because of this effect, the accuracy of demand estimation becomes more important than when inventory can be held. Second, service providers may have a special need to know the entire demand distribution in order to plan capacity to minimize its cost in the environment of uncertain customer demand. Often a service provider can choose capacity from a number of alternatives, for example from a service technology with high acquisition costs and low use costs or from one with low acquisition costs and high use costs. The cost minimizing choices of quantity and type of capacity from alternative service technologies will be a mix of the technologies and depend upon the distribution of uncertain demand. Thus, a service provider’s choice of capacity mix will require more information about the aggregate demand distribution of her customers. In contrast, in the traditional inventory-planning context a newsvendor-choosing inventory needs to know one quantile of the distribution.

Typically, a customer is much better able to forecast his usage than the service provider. But an information asymmetry generally exists due to the contractual agreement between the parties. In short, the customer may not be motivated to truthfully communicate his needs as expressed by the best estimate of his probabilistic distribution of demand. Thus, for the two reasons outlined above, elicitation of the customer’s demand distribution is an important feature of planning service capacity.

The original inspiration for this problem came from a defense electronics firm, which produced a wide variety of integrated circuit (IC) boards for use in military two-way radios. In this electronics firm, various business units within the firm shared the “tech center,” which performed automated insertion and wave soldering services. The custom nature of the business generally prevented holding inventories of completed IC boards. The loads the business units placed on the tech center varied substantially. For example, as customer orders changed in mix or volume, demand for this firm’s two-way radios would vary. As revenues were not affected by its decisions, the challenge of the tech center was to plan capacity by considering the cost of acquiring capacity balanced against the ability to provide the business units’ immediate needs. The tech center managers chose capacity as a mix of different types of equipment. Managers chose among several types of equipment that ranged from nearly automatic (that was more expensive but required little labor and thus little as-used cost) to more manual equipment (that was less expensive but required more labor and thus more as-used cost). The business units were in a superior position to forecast needs, but due to organizational incentives between the center and business units, the business units were not generally motivated to truthfully communicate their needs. Why this is so becomes clear with two extreme examples. A customer paid only based on his estimate of demand and not on the actual usage, he would be inclined to underestimate demand in reports to the tech center. On the other hand if he paid only for actual usage but nothing on the estimate, he would be inclined to overestimate to ensure there was plenty of capacity when he needed it. Thus a second challenge for the tech center, and our focus in this article, is to elicit the best possible forecast of a business unit’s demand.

We now begin a more formal description of our problem by describing the sequence of events and decisions. For purposes of exposition, assume a planning
period for capacity over which capacity is held fixed, say 1 year. The planning period is divided into usage periods over which the customer demands and uses service capacity, say weeks. At the beginning of the planning period the service provider (the tech center) first elicits an estimate of each customer's demand distribution, then aggregates this probabilistic information into an estimate of total demand, and finally acquires physical capacity (machines and people) in a cost minimizing fashion to provide service to the customers (the business units).

In this article, the method that the service provider will use for eliciting the customer's demand distribution is a pricing mechanism that involves the customer purchasing multiple contracts for service at the beginning of the year. To simplify the presentation for the moment we assume that each contract is for one unit of service and that each week the customer either uses it for one unit of service or he does not use it at all. Each contract has associated with it an upfront price and a conditional or as-used price for the actual delivery of the unit of service during the week. The service provider will offer a variety of such contracts that differ on the amount paid up front versus the amount paid at use. For the customer to receive a unit of service he must have purchased a contract in advance. Because the customer is uncertain of his demand in any week but knows his demand distribution he can take advantage of these options to reduce his expected total fees paid for services. Intuitively, if it is fairly likely that his demand will exceed one unit, the customer will be willing to buy a contract for this first unit of demand where the fee is, say, mostly upfront and very small when used. The 100th unit of demand is much less likely to occur in any period and so he may prefer a contract that involves lower payment upfront and much more when used. The former contract would need to be cheaper than the latter if used most weeks, to induce the customer to buy it, because he is taking the risk of paying for it upfront. The service provider can construct these contracts so that when the customer makes these choices to minimize his expected cost, he will, as we will show, reveal information about his distribution.

This revealed information is useful to the cost minimizing service provider who chooses from multiple technologies for supplying internal service capacity. The optimal choice of the technologies bought will generally be a mix. The reason a mix is optimal can be intuitively explained. For the example of the tech center, if some portion of volume is steady from week to week the tech center could buy relatively sophisticated and automated equipment, and hire a skilled operator, who would be dedicated to the machine. The costs to have the capacity whether it is used or not would be high but the marginal cost to use that unit would be low (e.g., low scrap and thus low material costs and no additional labor). Similarly if some portion of volume were rarely demanded, the service provider could choose a labor-intensive technology, and thus the cost to have a unit of this capacity is low, assuming that manual labor is productively used elsewhere when volume is low, but the marginal cost to use that unit would be high. Thus, on a per unit basis, the service provider would have options that would have high costs of having capacity and low costs of using capacity and others that would have low costs of having capacity and high costs of using capacity. Driven by the uncertainty in customer demand, typically each of these options will be used in varying amounts.
Finally, for simplicity, we assume that there is a single stage to the process of providing the service because if there are multiple stages the analysis can be replicated for each resource. To assure that all demand can be met, we assume that there is an outside provider who the internal service provider can call on, typically at a high cost. On peak demand days she may do this, but typically she will acquire capacity to deliver most of the service to the customers using internal resources.

The remainder of the article is organized as follows. In the next section we discuss the literature related to the focus of this article, namely the elicitation of the unknown demand distribution. The following section introduces the mathematical model. Next, we show that the newsvendor-like problem that the service provider passes to her customer by constructing a finite set of contracts, can cause the customer to truthfully reveal a finite set of quantiles of his demand distribution and to truthfully report his actualized demand. It also presents a related mechanism that allows direct revelation of a customer’s quantiles, by asking him for his probability distribution. In the next section, we present an alternative representation of contract prices that will help elicit the customer’s entire probability distribution. This section demonstrates the generalization and proves the incentive compatibility of the new mechanism to elicit the entire demand distribution. The following section establishes the value of distribution elicitation. It shows for a range of scenarios how information about the demand distribution can be used in capacity planning and computes the cost savings created from knowledge of the probability distribution. It considers an example where the customer would have the incentive to collect additional information, which would increase the service provider’s profit. Next, two extensions are briefly considered: lost sales and multiple customers. Finally, we conclude with a discussion of the advantages and limitations of the model, and identify opportunities for future research.

LITERATURE REVIEW

Our modeling is most closely related to the literature in probability and decision analysis on the design of scoring rules. This literature focuses on how a principal can ex-ante induce an agent to accurately (and thus honestly) assess and report a probability distribution whose ex-post outcome is observable by the principal. Based upon the reported distribution and the observed outcome, the agent is paid a reward that is ex-ante maximized with truthful revelation. It is implicitly assumed that the incentive reward is sufficient encouragement for the agent to reveal what he knows. Thus, the ex-post use of the elicited distribution has no effect on the agent’s incentives. Our work differs from this general literature as we use pricing as the means of inducement, which generates a mechanism that is quite different from that found in the scoring rules literature. Another difference is that we assume that the ex-post outcome is not directly observable by the principal (here the service provider) who depends upon truthful revelation of the outcome by the agent (here the customer).

A recent survey of research about scoring rules is found in Gneiting and Raftery (2007). Scoring rules were first studied by Savage (1971) and applied to practical problems such as weather forecasting. A scoring rule that is maximized ex-ante only with the truthful distribution is called a proper scoring rule. The
following basic theorem by McCarthy (1956) and Savage (1971) states necessary and sufficient conditions which characterize categorical proper scoring rules.

**Theorem 1:** Suppose $P_m = \{(p_1, p_2, \ldots, p_m)\}$ such that $\sum_{i=1}^m p_i = 1$, a scoring rule $S$ is a mapping from an element of $P_m$ into $R^m_+$; with $S(p, i) \to R_+$; $p \in P_m, i \in \{1, \ldots, m\}$. Let $G(p) = \sum_{i=1}^m S(p, i) p_i$. Any scoring rule is proper if and only if $S(p, i) = G(p) - G'(p) \cdot p + G'_i(p)$ and $G(p)$ is a convex function in $P_m$, $G'(p)$ is the subgradient of $G$ at $p$ and $G'_i(p)$ is the $i$th component of the subgradient.

Theorem 1 in Gneiting and Raftery (2007) is an extension of this theorem to continuous probability distributions, which is technically complex as it is generalized to function spaces but follows the discrete case in spirit. Thus, any bounded convex function $G$ on $R^m_+$ can generate a proper scoring rule and the space of proper scoring rules is huge.

Because the space of proper scoring rules is so vast, the specific elicitation problem at hand (e.g., to elicit a categorical distribution, a dichotomous probability distribution, a continuous density, a continuous cumulative distribution, or a first and second moments estimate) suggests different scoring rules: quadratic (Brier, 1950), spherical (Gneiting & Raftery, 2007), logarithmic (Good, 1952), Beta (Buja, Stuetzle, & Shen, 2005), CPRS (Matheson & Winkler, 1976), Energy (Szekely, 2000), and based upon first and second moments only (Dawid & Sebastiani, 1999) to name only a few. Choosing among these rules for the best one for each application is largely an issue that can be informed empirically or by postprediction simulations (Gneiting & Raftery, 2007, pp. 373–374).

Matheson and Winkler (1976), Cervera and Munoz (1996), and Jose and Winkler (2009) explore distribution elicitation using newsvendor-like scoring rules. Their process (like ours) is quantile elicitation for a random variable, where an agent is given a cumulative probability and then reports a value of the random variable with the corresponding probability. The principal who wishes to find out the agent’s distribution sets up a lottery whose payoffs look like a newsvendor problem. It should be emphasized that the agent’s payoff is constructed to create a revelation incentive. The agent is not otherwise a true newsvendor with customers to serve. Indeed the distribution can be related to any uncertainty, not just demand.

Matheson and Winkler (1976) assume a principal needs information about a random variable, $X$, with continuous cumulative distribution function $P(x)$ that is held by an agent. The rule is constructed as the sum of an infinite number of single-quantile rules, as follows: The agent is asked to assess the $z = \frac{y-c}{s}$ th quantile of $X$. Suppose the agent replies “$y$.” If the realization of $X$ is $x$ the agent receives:

$$T(z, y, x, s, c) = \begin{cases} (s-c)x - c(y-x) & \text{if } x \leq y \\ (s-c)y & \text{if } x > y. \end{cases}$$

This payoff structure is that of a newsvendor: with the message “$y$,” the agent makes an upfront payment $cy$, and ex-post receives a payment from the principal of $s\min(x, y)$. The following story can be told to generate this payoff function: ex-ante of observing $X$, the agent can buy a quantity $y$ of some good at a unit price $c$ from the principal, and ex-post of observing $X$, the agent can sell units back to
the principal at a unit price $s$, up to the realized demand $x$ but limited by $y$. With this payoff structure, and assuming risk neutrality, the optimal action for the agent is to report the $z$th quantile, that is, $y = P^{-1}(z)$. Thus this scoring rule is a proper scoring rule, and will elicit the $z$th quantile. In summary, for the $z$th quantile, an artificial newsvendor type payoff structure is constructed. We say artificial because the agent is not buying and trading a real good.

The authors extend this result to an infinity of quantiles by considering all values of $z$ and asking the agent to report the entire probability distribution as a cumulative distribution function $R(\cdot)$ that may or may not be the truthful $P(\cdot)$. The expected payoff to the agent is defined by Matheson and Winkler as

$$ T \ast (R^{-1}) = \int_0^1 \int_0^\infty T(z, R^{-1}(z), x, s(z), c(z)) dP(x) dz, $$

where $s(z)$ and $c(z)$ must satisfy $z = \frac{s(z) - c(z)}{s(z)}$. (Note, the authors do not denote the functional dependence of $s$ and $c$ on $z$, but without such dependence the rule is not proper.)

This rule is the sum of the payoffs of an infinite number of artificial newsvendor-like problems (one for each $z$). The payoff structure is very different from that of any real newsvendor problem where a single inventory level is planned. In contrast, we model elicitation from a customer to discover many, if not all, quantiles. We show how a service provider can elicit the entire distribution of its customers’ by offering a menu of upfront and conditional prices, and letting customers’ choose as many of each as desired. The distribution elicitation problem using such contracts does not appear in the large scoring rules literature to the best of our knowledge.

Related research in economics studies mechanism design that assures truthful revelation. In our problem, a service provider designs a mechanism using price to elicit valuable information about customers’ needs. The mechanism design problem can be viewed as a classic principal-agent problem in which the service provider picks the price mechanism to maximize her profit, subject to the constraint that the agent is given incentive to report truthfully. As such, it can be analyzed as a game with incomplete information. Myerson (1979) shows that any game of incomplete information with a Nash equilibrium can be transformed into a mechanism where truthful revelation is a Nash equilibrium. Green and Lafont (1979) provide a survey of the mechanism design literature. In this literature, truthful revelation is not generally a dominant equilibrium. Our result is that one of the players (the service provider) has a dominant strategy using a well-defined price system that always induces the customer to truthfully reveal his probability distribution. However, there are costs associated with inducing information in the form of forgone profit. We show that our mechanism can reduce these losses to values arbitrarily close to zero.

Our article employs prices that have a two-price structure: an upfront and a conditional component, where customers are offered a range of plans. It is well recognized in the economics literature that a multi-part tariff is useful for segmenting markets, cost recovery by monopolies, and "golden handcuffs" to encourage customer loyalty. See for example, Oi (1971) which explains the use of two-part prices
to extract surplus from customers. Wilson (1993) contains a thorough overview
and discusses the profitable use of multi-part and other nonlinear pricing with
heterogeneous customers. Although we use very similar "two-part prices" we use
them to elicit probabilistic information. This is quite unlike the use in the former
literature.

MODEL

The sequence of events for the service provider and the customer are as follows.
The service provider requests information from the customer about his demand
distribution by offering a set of contracts. Next, the customer chooses contracts.
Based on this information the service provider decides on the capacity of each
technology to acquire, ensuring that she can meet the contractual obligations
of the contracts sold to the customer. Finally the customer privately observes his
demand and requests service using his chosen contracts. In this section, we simplify
the situation described in the introduction to a problem where a service provider
delivers service to a single customer for a single period. The extension to multiple
customers is discussed in a later section.

The customer wants to minimize his expected fee, including the reserve
price of all the quantities of contracts purchased and all the conditional or as-used
prices of all contracts used to satisfy demand. The provider wants to maximize her
expected profit, which, once the contracts are set, will be equivalent to minimizing
her costs, equal to the costs of acquiring the quantities of all the technologies and
the usage costs of all units used to satisfy demand.

Next, we model the environment and decisions of the customer and then the
decisions of the service provider. The customer’s demand is a random variable
$D$ with a distribution $\text{Prob}\{D \leq x\} = P(x) = 1 - Q(x)$, which is known to the
customer or costless for him to obtain, and it is unknown to the provider.

From a set of $n$ contracts, the customers must choose the quantity to buy
of each contract. A unit of a contract guarantees a unit of capacity to meet the
customer’s demand. The price structure for these contracts is as follows. There is a
reserve price $r_j$ to buy a unit of contract $j$ before a demand actualization is observed
and a price $u_j$ to use a previously reserved unit of contract $j$ to receive a unit of
service, so we call this the as-used or conditional price. Appendix B provides a
table of notation.

If a customer purchases a unit of contract $j$ then it is not necessary for the
service provider to reserve a physical unit of capacity for this individual customer.
Nonetheless we will occasionally refer to a unit of contract $j$ purchased by the
customer as a unit of (reserved) capacity of type $j$.

The customer can purchase any amount $y_j \geq 0$ of contract $j$, $j = 1, \ldots, n$. If
the demand allocated to these units is $d (\leq y_j)$ then the fee the customer pays is $r_j y_j + u_j d$.

We order the $(r_j, u_j)$ pairs so that the $r_j$’s are in nonincreasing order. If one
of the $u$’s is smaller than its predecessor, then the predecessor contract will never
be desired as the current contract will be cheaper and preferred in both reserve
price and as-used price. Such a predecessor contract may be eliminated because it is “dominated”, and we assume that all dominated contracts are removed. The
customer purchases contracts $1, \ldots, n-1$. We assume that contract $n$ has $r_n = 0$, so that the customer will always effectively purchase an infinite amount of this contract so long as this conditional price $u_n$ does not exceed the customer’s willingness to pay for a unit of service which we denote by $p$. These service contracts share some characteristics with the “real options” models, which are described by Dixit and Pindyck (1994, 1995) and Coy (1999), among others. For these papers, an up-front commitment is made that allows the option of undertaking some future activity if desirable or necessary. But our focus is somewhat different: a monopolist (the service provider) sets the prices for options and offers a myriad of contracts for the customer to choose from. Further, unlike financial options, the uncertainty in demand drives the decision to exercise the option, not fluctuations in the price of a security or commodity.

Let $x_j$ represent the sum of prepurchased amounts of all contracts $1, \ldots, j$. Let $y_j = x_j - x_{j-1}$ represent the amount of contract $j$ which is purchased. For notational convenience, we define $x_0 \equiv 0$ and let $X = (x_0, x_1, \ldots, x_n)$ be the vector of the contract quantities.

The service provider wishes to minimize the supply cost and has a choice of $m$ technologies to purchase to supply capacity. A unit of capacity of technology $i$ has associated with it two cost functions: a cost $h_i(z_i)$ for having $z_i$ units of output capacity using technology $i$, (i.e., the cost of purchasing it upfront before an actualization of demand is realized), and a cost $c_i(d)$, for using a quantity $d$ of capacity of technology $i$. The provider can purchase any quantity $z_i > 0$ of each technology $i, i = 1, \ldots, m$. If demand allocated to these units is $d (\leq z_i)$ then the total cost to the provider for these units is $h_i(z_i) + c_i(d)$. Because our goal is to demonstrate that the elicited information is valuable to the service provider, we study the simplest case, namely when the marginal cost of having and using capacity is constant. So, for the remainder of the article we will limit ourselves to the case that $h_i(z_i) = h_i z_i$ and $c_i(d) = c_i d$.

A much more complicated cost model could be used involving, for example, limited, discrete choices on quantities of capacity to purchase or other nonlinear costs. The calculations would be more complex, yet as long as the cost structure allows for savings for the service provider when there is uncertainty in demand, the service provider still has the potential to save by eliciting an accurate estimate of the demand distribution and using that information in solving the optimization.

With constant marginal cost, the service provider will adopt only nondominated technologies. Essentially, this means that each of the $m$ technologies being considered will be optimal for a different range of frequency of use. This nondominated property induces an order on the technologies, and we assume they are sequenced in this order, which implies that the $h_i$’s are decreasing and the $c_i$’s are increasing. Define $z_i$ to be the number of units of capacity of technology type $i$ purchased, with $Z \equiv (z_1, \ldots, z_m)$. Then the total costs incurred for having this capacity will be $H(Z) \equiv \sum_{i=1}^m h_i z_i$.

We assume a final technology $m$, $h_m = 0$, and $c_m \leq p$, where $p$ is the reservation price of the customers. This assumption allows the firm to handle arbitrarily large demand and to serve all demand profitably. This implies that there exists some technology, which might correspond to outsourcing. We assume that this technology does not require any upfront payment by the firm ($h_m = 0$), so
that she can deliver any number of units of service with this technology ($z_{m}^{*} = \infty$). Later we discuss the cases where $c_{m} > p$ so the service provider won’t be able to use it profitably, or serve all the market demand. In addition, we define $w_{i} \equiv \sum_{k=1}^{i} z_{k}$ and $W \equiv (w_{0}, w_{1}, \ldots, w_{m})$ where $w_{0} \equiv 0$ for convenience. Next we study the problem of truthful elicitation of the customers’ distribution.

ELICITING TRUTHFUL INFORMATION OF CUSTOMER QUANTILES

In this section we focus on the customer decisions, to purchase contracts and then after demand is actualized to allocate those contracts to demand. We describe how the pricing mechanism for services given by these contracts allows the service provider to infer some information about the true probability distribution of demand of the customer via quantile assessment. The mechanism also gives an incentive for the customer to report actual demand once it is actualized because contract prices are always less than or equal to the customer’s reservation price. In the next section we show how the entire distribution can be discovered.

Incentive compatibility in customer choice arises in many operations models, especially involving service contracting. Lederer and Li (1997) and Ha (1998) present service models with priority in which customers choose to pay higher prices for faster service.

The customer will report the true demand actualization because to not report some portion of demand would mean not receiving service and his willingness to pay for service, $p$, exceeds $u_{n}$. When the customer experiences a demand $d$, he will use the purchased contracts, sorted by increasing as-used (or conditional) price. Thus, he will first use $\min(x_{1}, d)$ units of contract 1, and more generally use $\min(x_{j}, d) - \min(x_{j-1}, d)$ units of each contract $j$. The total reserve fee $R(X)$ will be

$$R(X) = \sum_{j=1}^{n} r_{j}(x_{j} - x_{j-1}) = \sum_{j=1}^{n-1} (r_{j} - r_{j+1})x_{j},$$

because $r_{n} = 0$. The as-used fee $U(X, d)$ will be

$$U(X, d) = \sum_{j=1}^{n} u_{j}[\min(x_{j}, d) - \min(x_{j-1}, d)]$$

$$= \sum_{j=1}^{n} u_{j}[(d - x_{j})^{+} - (d - x_{j-1})^{+}]$$

$$= u_{1}d + \sum_{j=1}^{n-1} (u_{j+1} - u_{j})(d - x_{j})^{+}.$$
The expected total expenditure by the customer will be

\[ \Pi(X, D) \equiv R(X) + E(U(X, D)) \]

\[ = u_1 E(D) + \sum_{j=1}^{n-1} [(r_j - r_{j+1})x_j + (u_{j+1} - u_j)E(D - x_j)^+] \]

and the optimal reservations \( X^\ast \), which we will also denote by \( X(D) \) to indicate its dependence on \( D \), will satisfy

\[ q_j \equiv \frac{r_j - r_{j+1}}{u_{j+1} - u_j} = Q(x_j^\ast) \text{ for } j = 1, \ldots, n-1, \]  

(1)

with \( q_0 \equiv 1 \) and \( q_n \equiv 0 \). The ordering of the contracts ensures that the \( q_j \)'s are monotone and decreasing.

From this mix of purchased contracts, the service provider can infer some information about the customer’s probability distribution. In particular, as the service provider also knows the \( r_j \)'s and \( u_j \)'s, she can infer \( n-1 \) points, \((x_j^\ast, q_j)\), on the customer’s complementary cumulative demand distribution. Thus, use of this pricing mechanism will reveal information to the service provider that includes truthful reports of a finite number of quantiles of the probability distribution, and the demand actualization in the period. The customer’s demand will be truthfully reported because we assume that the price paid is always less than or equal to his reservation price.

**Example of Eliciting Quantiles**

To illustrate the idea, suppose that our customer knows his demand is uniformly distributed between 5 and 10 units, and that the firm offers three types of contracts:

\[ r_1 = \$2; \quad u_1 = \$3, \]
\[ r_2 = \$1; \quad u_2 = \$5, \]
\[ r_3 = \$0; \quad u_3 = \$10. \]

When the customer is offered these three contracts, he will reveal his \( \frac{r_1 - r_2}{u_2 - u_1} = 0.5 \), and \( \frac{r_2 - r_3}{u_3 - u_2} = 0.2 \) quantiles of the reverse c.d.f. Indeed the customer will purchase \( x_1 - x_0 = x_1 - 0 = y_1 = 7.5 \) units of contract 1; \( x_2 - x_1 = y_2 = 9 - 7.5 \) units of contract 2 and will buy all units of service in excess of 9 units at a price of \$10 per unit using contract 3.

In some situations, the unknown demand distribution may be drawn from a known class described by, say, \( n-1 \) unknown parameters. Thus, by offering the customer \( n \) contract types the customer’s revelation of \( n-1 \) quantiles will reveal the entire distribution.

As an alternative mechanism, the customer reports his probability distribution directly, rather than the quantities of contracts to be purchased. Suppose the service provider offers the same portfolio of contracts, given by the reserve price \( r_j \), and an as-used or conditional price \( u_j \) for \( j = 1, \ldots, n \). Suppose that rather than picking some vector of the contract quantities, \( X \), the customer reports a complementary
cumulative probability distribution $S(\cdot)$:

$$S(\cdot) \text{ nonincreasing, with}$$

$$S(0) = 1,$$

$$\lim_{x \to \infty} S(x) = 0.$$  \hspace{1cm} (2)

Given the report of $S(\cdot)$ the mix of capacity contracts purchased will be defined by

$$S(x_j) = q_j, \ j = 1, \ldots, n - 1.$$ \hspace{1cm} (3)

Note that the $q_j$’s are defined by Equation (1). We can show that if the customer chooses $S$ to minimize his expected cost, he will truthfully report the $n-1$ quantiles defined in Equation (3). This result follows because the cost minimizing choice of contracts corresponds to the solution to Equation (3), thus, truthful reporting of $S = Q$ will result in the customer’s fee-minimizing $X^*$. However, any $S$ that agrees with the $n-1$ quantiles $q_j, \ j = 1, \ldots, n - 1$ will also work. Next, we ask how the complete distribution can be truthfully elicited.

**ELICITING THE ENTIRE DISTRIBUTION**

**An Alternate Representation of Contract Prices**

To elicit the complete distribution we need to generalize the concept of a finite number of contracts to a continuum of contracts. To develop this concept we present an alternative representation of a contract. We begin by using the finite portfolio of contracts as an example and then demonstrate its general use. Consider an individual unit of capacity purchased under contract $j$ with prices $(r_j, u_j)$. Recall that we sequence the contracts that the customer purchased in order of their frequency of use. If this unit is the $x$th unit, it will be used only when all previous units are in use. Thus, the probability that the customer uses this unit is $\text{Prob}\{D \geq x\} = Q(x) = q$, and the expected fee for this unit of service is $r_j + u_j q$. We now define a new function $\pi(\cdot)$, which gives the least expected fee of using a unit of a contract for a fraction $q$ of the time through contracts $\{(r_j, u_j)\}_{1 \leq j \leq n}$. Let

$$\pi(q) = \min_{1 \leq j \leq n} \{r_j + u_j q\}. \hspace{1cm} (4)$$

Again we assume that the contracts are not dominated, so for a set of contracts the function $\pi(q)$ describes can also be equivalently written as

$$\pi(q) = r_j + u_j q \text{ if } q \in [q_j, q_{j-1}] \text{ for } j = 1, \ldots, n,$$ \hspace{1cm} (5)

where $q_0 = 1, q_j = (r_j - r_{j+1})/(u_{j+1} - u_j)$, and $q_n = 0$. The function $\pi(\cdot)$ is piecewise linear, so it is not differentiable at the $n-1$ points $\{q_j\}$. For convenience, define $\pi'(q_j)$ to be the limit from the left, $\pi'(q_j^-)$, for these points.

The function $\pi(q)$ is an alternative representation of the prices charged by the portfolio of contracts. We interpret the $\pi(q)$ function as follows. If a customer wishes to buy a unit of capacity corresponding to $q$, where one can think of $q$ as a parameter specifying the contract, then as before the customer pays in two parts:
\(\pi(q) - q\pi'(q)\) is the reserve price for this contract, and \(\pi'(q)\) is the as-used price of that contract, or the conditional price, if the unit is used. When the function \(\pi(q)\) is piecewise linear, the points \(q_j\) define the quantiles at which the contracts change their reserve and as-used charges. We now present an expression for the total reserve fees and the total conditional fees for a customer who reports his distribution as \(S(\cdot)\) and incurs a demand \(d\).

**Proposition 1:** Given the \(S(\cdot)\) defined from \(X\) as in Equation (3), \(\pi(\cdot)\) defined by Equation (4), and the actual demand \(d\), then

\[
R(X) = \int_0^\infty \left[ \pi(S(x)) - S(x)\pi'(S(x)) \right] dx
\]

\[
U(X, d) = \int_0^d \pi'(S(x)) dx.
\]

The proof of this and all the later results in this article appear in Appendix A.

The next subsection shows that these price contracts can induce the customer to truthfully reveal his entire distribution function.

**An Extension for General Concave Functions \(\pi(\cdot)\) and the Incentive Compatibility of the Pricing Mechanism**

In the previous section, we described how a customer would choose among \(n\) contracts by considering both the reserve and the as-used prices of contracts to determine the amount of each contract to buy, and that the customer’s choice can be described by a function \(S(\cdot)\), which has the form of a complimentary cumulative distribution. In the previous subsection we showed that the prices of the contracts can be described by a piecewise-linear increasing concave function \(\pi(q)\). Now we drop the restriction that \(\pi(q)\) be piecewise linear, and instead assume only that it is an increasing, strictly concave function with \(\pi(0) = 0\) and \(\pi'(0) = \pi_0 > 0\). We will show that this generalization can truthfully elicit the entire demand distribution from each customer, rather than a finite number of quantiles. The service provider offers each customer a “contract” function \(\pi(\cdot)\), and asks him to decide on contracts by specifying a decreasing function \(S(\cdot)\). One interpretation of this choice is that, rather than a discrete number of contracts, an infinite number of contracts are now available; each one of which is defined by some \(q \in [0,1]\). Any contract \(q\) will have a reserve price of \(r(q) = \pi(q) - q\pi'(q)\), which is the intercept of the tangent to \(\pi(\cdot)\) at \(q\), and an as-used (conditional) price of \(u(q) = \pi'(q)\), which is the slope of \(\pi(q)\). The customer submitting a schedule \(S(\cdot)\) will pay

\[
R(S) = \int_0^\infty \left[ \pi(S(x)) - S(x)\pi'(S(x)) \right] dx \tag{6}
\]

to reserve service and

\[
U(S, d) = \int_0^d \pi'(S(x)) dx \tag{7}
\]
to use that service if demand is $d$. Thus the customer’s expected fee under any such schedule $S(\cdot)$ will be

$$\Pi(S, D) = R(S) + E_D[U(S, D)] = R(S) + \int_0^\infty U(S, d) dP(d).$$

In this generalized setting, the selection of $S(\cdot)$ can be equivalently interpreted as the selection of a total capacity $w(q)$ for the subset of contracts $\{(r(t), u(t))|t \in [q, 1]\}$ and $w(q) = S^{-1}(q)$.

Given this situation, what $S(\cdot)$ should the customer specify in order to minimize his expected fees? This question is answered by the next proposition, which shows that reporting one’s true probability distribution is incentive compatible.

**Proposition 2:** Incentive Compatibility. Let $S$ denote the class of possible contract schedules, which will be functionally equivalent to the class of (left continuous) complementary cumulative tail distribution functions. Consider a customer with a demand of $D$ with a complementary c.d.f. of $Q(x)$. This customer would then minimize his expected fee by picking $S = Q$. In other words, the optimization

$$\Pi^* \equiv \min_{S \in S} \{\Pi(S) = R(S) + E[U(S, D)]\}$$

is solved by $S^* = Q$, and its optimal value $\Pi^*$ is

$$\Pi^* = \int_0^\infty \pi(Q(x)) dx.$$

As a result of Proposition 2, we know that contract $q$ will be used a fraction $q$ of the time and so we can now interpret $\pi(q)$ as the expected fee for a unit of contract $q$. The next proposition follows from Proposition 2.

**Proposition 3:** The expected payment $\Pi^*$, is a concave function of the complementary cumulative distribution function, $Q$.

This result is clear because for any fixed $x$, $\pi(\cdot)$ is a concave function of $Q(\cdot)$ and integration preserves concavity.

An important consequence is that the customer has incentive to reduce his own uncertainty about demand. Suppose the service provider and customer have a common prior on demand, but the customer observes a private signal about the distribution before he places his order. Because of Jensen’s inequality, the expected payment under a mixture of posteriors is greater than the weighted sum of the expected values of the posteriors. However, in many cases, observation of a signal is costly. Therefore, the customer values more refined information about his demand distribution and will invest in observing the signal if there is net value gained.

When the customer gains value by investing in more information then the service provider and the customer both expect the payment to be smaller than would be obtained without the signal using the prior. Therefore, if the customer has incentive to collect information privately, our pricing mechanism will induce a change in the service provider’s revenue. We explore this issue in the next section.

In the case that the customer’s probability distribution is parameterized by its mean and standard deviation we can present a stronger result that shows that
the customer’s fee is directly related to both these parameters. In particular, the fee paid is linearly and positively related to the standard deviation.

**Proposition 4:** Consider some given family of location/scale probability distributions for nonnegative demand, parameterized by $\mu$ and $\sigma$. Then for a fixed set of contracts $\pi(-)$, there exist constants $a > 0$ and $b > 0$ such that for all $(\mu, \sigma)$, $\Pi^{\ast}_{\mu, \sigma} = a\mu + b\sigma$. That is, all other things being equal, the expected fee is a linear function of $\mu$ and $\sigma$.

**Proposition 5:** The optimal expected fee paid by a customer, $\Pi^{\ast}$, is bounded both above and below as $\pi(1)E(D) \leq \Pi^{\ast} \leq \pi'(0)E(D)$.

Although we have shown a bound exists on the service provider’s profit, we next show that she can extract revenue almost equal to the upper bound.

**Proposition 6:** There exists a choice of $\pi(-)$ that allows the service provider to learn the entire demand distribution of the customer and obtain a revenue arbitrarily close to the maximum possible, $\pi_0E(D)$.

This section has shown that the customer generally has incentive to reduce his uncertainty about demand, but that reduced uncertainty causes the service provider’s revenue to fall. In the next section, we show that the service provider’s expected cost declines with better information. An example shows that there are situations where with better information the service provider’s expected cost declines more than the fall in the expected payment, and, at the same time, the customer’s fee declines more than the cost of obtaining a signal. Thus in special situations and under the pricing mechanism, both the service provider and the customer are better off with more information, and our mechanism is attractive to the service provider.

**VALUE OF INFORMATION AND THE INCENTIVE TO COLLECT INFORMATION**

In this section we discuss two issues: first, the value to the service provider of knowledge of the customers’ distribution of demand, and second, how the service provider’s choice of $\pi(-)$ can create incentives for the customer to collect information and thereby increase the service provider’s profit.

We next study the service provider’s production problem: how the elicited demand distribution is used by the firm to minimize the expected cost of serving the customer through technology mix decisions as described in the introduction. The optimal choice of capacity is generally a mix of technology types. We use a simple model with constant marginal costs but more complex problems with nonlinear costs could be analyzed and the service provider would still gain from elicitation of the demand distribution if it had these two marginal costs (for having and for using capacity). This problem of planning capacity mix in our simple context has been previously studied most prominently by Crew and Kleindorfer (1976), and
we follow their approach. Note that here we take the \( \pi(\cdot) \) function as given so minimizing the service provider’s cost is the same as maximizing her profit.

The analysis of capacity plans follows a newsvendor analysis. This newsvendor analysis is quite similar in form but quite different in interpretation to the study of probability distribution elicitation studied so far. Previously we studied the problem of demand distribution elicitation by the service provider from her customer(s), but here we study the planning of capacity using the already elicited distribution. We return to the single period problem. Now \( D \) is the total demand of all the customers.

Recall that \( z_i \) is the amount purchased of technology \( i \) and \( w_i \) represents the total amount purchased of technologies 1 through \( i \). (Recall that \( z^*_m = \infty \) and thus \( w^*_m = \infty \).) Then defining \( w_0 \equiv 0 \) and observing that \( z_i = w_i - w_{i-1} \) for \( i = 1, \ldots, m \), we have \( H(W) = \sum_{i=1}^m h_i(w_i - w_{i-1}) = \sum_{i=1}^{m-1} (h_i - h_{i+1})w_i \), since \( h_m = 0 \).

The expected total usage cost incurred depends on the distribution of demand. Let \( d \) be any realization of the random demand \( D \) whose c.d.f. and its complement are \( P(\cdot) \) and \( Q(\cdot) \), respectively. With \( H(W) \) effectively sunk after capacity is acquired, the service provider will use as much as possible of the purchased technologies with the lowest usage costs \( c_i \). So whenever the service provider experiences some demand \( d \), she will first use \( \min(w_1, d) \) units of capacity from technology 1. More generally, she will use

\[
\min(w_i, d) - \min(w_{i-1}, d)
\]

units of technology \( i \) to meet a demand of \( d \). Thus the total cost to use the capacity for any demand \( d \) will be

\[
C(W, d) = \sum_{i=1}^m c_i [\min(w_i, d) - \min(w_{i-1}, d)]
\]

\[
= \sum_{i=1}^m c_i [(d - w_{i-1})^+ - (d - w_i)^+].
\]

Because \( w_0 = 0 \) and \( w_m = \infty \), \( (d - w_0)^+ = d \) and \( (d - w_m)^+ = 0 \), we obtain

\[
C(W, d) = c_1d + \sum_{i=1}^{m-1} (c_{i+1} - c_i)(d - w_i)^+.
\]

Intuitively, the first term charges the cheapest variable cost \( (c_1) \) to every unit of demand. The summation then applies successive surcharges against the portion of demand that exceeds the capacity of each of the cheaper technologies. The total expected cost is

\[
\Gamma(W, D) = H(W) + E(C(W, D))
\]

\[
= c_1E(D) + \sum_{i=1}^{m-1} [(h_i - h_{i+1})w_i + (c_{i+1} - c_i)E(D - w_i)^+]. \quad (8)
\]
A Service Provider’s Elicitation

To find the optimal technology capacity-mix $W^*$, we set the partial derivatives of $\Gamma(W)$ equal to 0,

$$q_i \equiv \frac{h_i - h_{i+1}}{c_{i+1} - c_i} = Q(w^*_i) \text{ for } i = 1, \ldots, m - 1. \quad (9)$$

We define $q_0 = 1$ and $q_m = 0$ for consistency with the definition of $W$. The adoption of nondominated technologies implies that $\{q_i\}_{0 \leq i \leq m}$ is nonincreasing and so $\{w^*_i\}_{0 \leq i \leq m}$ is nondecreasing. For a distribution of demand $D$, let’s denote the vector of capacity choices $W(D)$. Analogous to $\pi(\cdot)$ we can define a cost function $\gamma(q) = \min_{i \leq j \leq m} [h_i + c_i q]$, which for a unit of capacity that will be used a fraction $q$ of the time, is the expected cost of using that unit of capacity.

If the firm does not have exact knowledge of the demand distribution, it will make its capacity mix choice based upon its prior distribution. Let $D'$ be the estimated random demand. Define $P'(\cdot)$ and $Q'(\cdot)$ to be the c.d.f. of demand and its complement. Then the firm will choose a capacity mix, $W(D')$, as given in Equation (9), with $Q'$ used in place of $Q$. The increased cost associated with the suboptimal capacity choices associated with $D'$, namely $W(D')$, is just the expected value of the difference in costs between $\Gamma(W(D), D)$ and $\Gamma(W(D), D)$ over all signals we would receive about $D$ through elicitation. This higher cost can be unboundedly large given a prior distribution that is highly skewed to the right.

In most instances, the optimal capacity mix vector, $W(D)$, obtained after eliciting the distribution of $D$ will differ from the capacity mix under limited information, $W(D')$, and thus the service provider will gain value by eliciting the demand distribution of the customer(s).

In the last section we showed that the customer has incentive to collect better demand information when the cost of information collection is less than the resulting fall in his expected fee. But a fall in expected fee reduces the service provider’s revenue. In this section, we have just shown that better information allows the service provider to reduce her cost. The question that remains is, “Can a price contract $\pi(\cdot)$ that induces information collection by the customer cause the service provider’s profit to rise because her cost falls more than her revenue falls?” In such a case both parties are better off with information collection. Because this situation does not always occur, we present an example that demonstrates the possibility.

Example in Which the Service Provider and the Customer Gain

To simplify this example we limit it to a single customer. We assume the service provider and the customer have a common prior distribution on the mean, $V$, of the customer’s demand distribution. We assume $V$ is distributed Uniform $[av, bv]$ and the customers demand distribution is also Uniform with mean $V$ and standard deviation $\sigma$, thus it has a distribution Uniform $[V - s\sqrt{V}, V + s\sqrt{V}]$. Given the mean is known and is equal to $v$, let $D_v$ denote the demand and $Q_v$ be the complementary cumulative distribution, namely, $Q_v(x) = (v + s\sqrt{V} - x)/(2s\sqrt{V})$ on $[v - s\sqrt{V}, v + s\sqrt{V}]$. The prior on the customer demand distribution is of course not uniform, and the complementary cumulative distribution, $Q'$ is given
Dobson, Lederer, and Robinson

Figure 1: Service provider’s and customer’s common complementary cumulative prior distribution of customer demand.

Note: For the example, we assume the underlying distribution of demand is uniform and derive the prior distribution of customer demand. The service provider knows that the standard deviation of the distribution is $s = 15$ and that the mean of this distribution has a prior density that is uniform $[150, 450]$.

by

$$Q'(x) = \int_{av}^{bv} \frac{1}{b_V - a_V} dv,$$

and an example is shown in Figure 1.

There are three technologies available to the service provider with $(h_1, c_1) = (6, 0), (h_2, c_2) = (1, 7)$, and $(h_3, c_3) = (0, 10)$. We assume the customer’s reservation price is $11$. For the contract price function we pick

$$\pi(q) = 11 \frac{(q + 1)^{1-\varepsilon} - 1}{1 - \varepsilon},$$

where $\varepsilon = 0.5$ and $\pi_0 = 11$, or a flat price of $11$.

Because the reservation price exceeds the service provider’s cost for any technology, the firm will make some profit on all capacity used. In Figure 2 we plot our flat price ($11q$), the contract price ($\pi(q)$), and the capacity cost as a function of $q$.

The service provider must decide whether to offer contracts using $\pi(\cdot)$ or offer a flat price. For the former, the customer must decide whether to collect information to update his prior. In this instance, the service provider knows the customer will update his prior if and only if his drop in expected fee, namely,

$$\int_{av}^{bv} \left( \Pi(Q, D_v) - \Pi(Q_v, D_v) \right) \frac{1}{b_V - a_V} dv$$
A Service Provider's Elicitation

Figure 2: For the example the graph shows the expected flat price, $11q$, the contract price, $\pi(q)$, and the expected capacity cost, $\gamma(q)$, for $0 \leq q \leq 1$.

exceeds his cost to collect the information. For now, we will assume that this holds, and that $\pi(\cdot)$ induces the customer to collect information.

If the service provider uses a flat price of $11$ per unit the customer has no incentive to collect information as its expected payment is independent of the information gleaned. Now the service provider purchases capacities $W(D')$ that minimize cost based on the prior, $D'$. Her expected profit is

$$
\int_{a_V}^{b_V} \left( 11E(D_v) - \Gamma(W(D'), D_v) \right) \frac{1}{b_V - a_V} dv.
$$

If the service provider decides to offer contracts given by $\pi(\cdot)$ then the customer does collect information. Now the service provider purchases capacities $W_v$ that minimize cost based on the posterior, $Q_v$. Her expected profit is now

$$
\int_{a_V}^{b_V} \left( \Pi(Q_v, D_v) - \Gamma(W(D_v), D_v) \right) \frac{1}{b_V - a_V} dv.
$$

If the difference in these two profits is negative the service provider will choose to offer the contracts $\pi(\cdot)$ and not the flat price.

We next present some computations where $\varepsilon = 0.5$ and we find values of $x \equiv (b_V - a_V)/2$ for which the customer collects information and the service provider increases her profit.

We calculate both the change in profit of the service provider and the change in fee for the customer. We use a uniform distribution on $[300-x, 300+x]$ for the prior distribution on the mean of the customer’s demand, and the actual customer distribution has a standard deviation of 15. We wish to see how results varied with $x$. The graphs of the absolute gain in the service provider’s profit and the customer’s savings appear in Figure 3. The percentage changes appear in Figure 4. One sees that the value of elicitation is substantial for the firm even when the uncertainty in
Figure 3: Service provider’s change in profit and customer’s change in fee when the customer collects additional information.

![Graph showing the change in profit and fee for the service provider and the customer.]

Note: The change in profit and the change in fee are shown for shifting ranges of the prior distribution on the mean, and increase as the spread, $x$, of the mean’s prior distribution is increased. If the customer’s change in fee exceeds the cost of collecting information then our mechanism can be implemented.

The demand is small (i.e., $V$ is uniform on $[250,350]$, and continues to grow more convex as $x$ increases). The customer does not gain as much. Yet, even a small percentage reduction in his fees could exceed the cost of collecting information and thus be enough to motivate the customer to provide an accurate estimate of his demand. If there is not much uncertainty in the mean of the distribution (i.e., $V$ has a narrow range), then the customer gains relatively little. The customer might be unwilling to do the work to provide a good estimate of the demand distribution.

Figure 4: Percentage change in service provider’s profit and the customer’s fee when the customer collects additional information.

![Graph showing percentage gain from improved estimate for the service provider and the customer.]

Note: The base for the service provider is the profit made with a flat price. The base for the customer is his or her fee when the service provider offers contracts $\pi(\cdot)$ but the customer does not collect information.
and so the service provider, recognizing this, would be unwilling to provide the
contracts and their associated discounts and instead use a flat price set to the
reservation value. The service provider fears losing substantial revenue without
gaining the associated reduction in costs. Clearly there are examples where the
mechanism described in this article will not make both parties better off.

One additional feature of this example is that the service provider can choose
the \( \pi(\cdot) \) to maximize her expected profit, understanding the customer use of \( \pi(\cdot) \)
in deciding to collect information or not. Technically speaking, the problem is
to choose \( \pi(\cdot) \) to maximize the expected service provider’s profit subject to the
constraint that the customer expects to reduce his expected fee, net of cost of
information collection, and that \( \pi(\cdot) \) remains concave. Such a computation depends
on the prior and posterior distributions assumed. This analysis is beyond the scope
of this article.

This section has demonstrated the net value of knowledge of the customer’s
demand distribution to the service provider for the purpose of planning capacity,
even when a substantial discount must be offered to the customer to induce his
cooperation to collect the requisite information, and this discount would result in
lost revenue for the service provider. This gain could justify the effort of offering
the contracts herein described. We now consider two extensions of the model.

EXTENSIONS

In this section, we consider two extensions of our model: lost sales and multiple
customers. If a technology, say outsourcing, with \( h_m = 0 \) does not exist then we
can conceptually think of defining this \( m \)th and last technology as “lost sales”
with \( h_m = 0 \) and \( c_m \) defined to be the reservation price of the customer, \( p \). We
now compute the fraction of time \( q_m - 1 \) from Equation (9), when demand is so
high, but infrequent, that the service provider would not be willing to actually
serve all the demand. Lost sales are captured by defining the \( \pi(\cdot) \) function to
have a slope of \( p \) in \([0, q_m - 1]\) which is then strictly concave and increasing from
there. When the customer picks a contract with \( q \leq q_m - 1 \) he receives no service
but at no cost to him. The service provider would be obligated to buy enough
capacity to handle all contracts purchased for contracts \( q \geq q_m - 1 \). The elicitation
would only yield information on the demand distribution for quantiles in \([q_m - 1, 1]\),
essentially eliciting a truncated distribution, yet that is all the service provider
needs to optimize her capacity decision.

To handle multiple customers the firm would require the customers’ indi-
vidual demand distributions in order to calculate the total demand distribution. If
customers’ demands are independent then the total demand distribution is easily
obtained. If there were an exogenous random variable, \( t \), observable by all parties
(e.g., a state of nature such as temperature, with distribution \( g(t) \), that tied together
the individual demands, in this case energy consumption), and the customers’ de-
mands are independent given the state of nature, then the service provider can set
state dependent price schedules using the ideas of this article and elicit the demand
distribution for each customer, \( l \), and temperature, \( t \), namely \( f_l(\cdot|t) \). The service
provider would then first compute the unconditional demand for each customer \( l \),
and aggregate these over the customers. This method
Dobson, Lederer, and Robinson

is closely related to the state-dependent pricing for electricity as modeled in Panzar and Sibley (1978).

SUMMARY AND DISCUSSION

This article has discussed the problem of eliciting the distribution function of customers’ demand. Because the purpose of eliciting the distribution function is for a service provider to make better decisions for the technology and capacity choice problem, we would like to conclude with a brief discussion of the advantages and the limitations of our model in real world applications.

We begin with some advantages. Our mechanism is built into the transactions of charging for service and so no external mechanism for elicitation is necessary. In order to receive service from the provider the customer must reveal his demand. To hide some portion of demand in a period would imply that no service is provided for the hidden demand. So, as long as the service provider keeps \( \pi'(0) \) below the customer’s willingness to pay for a unit of capacity there are no concerns about what actual demand is. The mechanism results in truthful revelations of the customer’s demand distribution so inefficiencies caused by underestimating or overestimating demand are avoided. The mechanism may appear to be complex, requiring the customer to solve an infinite number of trade-offs, yet, once both parties understand how and why it works it is simple to state and implement. The customer knows just to pass his best guess of the demand distribution. The reserve fee, computed from Equation (6), and the as-used fee each period, computed from Equation (7), is actually simple. The method works well with arbitrary distributions so the customer could simply report past empirical data, if he believes this is a fair representation of the future.

Next we list some limitations. First, in the case of a single customer, the service provider could simply construct the contracts to reveal the set of quantiles she needs to solve her technology choice problem. But when there are multiple customers, the service provider will need to convolute their distributions, which will require the entire distribution and will require some work depending on how the customers provide their distributions.

Second, we have assumed that the demand distribution was unaffected by the price the customer paid. But if we assume only that the expected demand is unaffected, then the extent of the concavity of \( \pi(\cdot) \) determines the discount, which will be larger for customers whose variance is less. Thus, there may be a motivation for the customer to modify his stream of demands to make them smoother (i.e., to reduce his variability and thus his cost). In this case, the distribution we would elicit would be this smoothed demand. Operationally, this adjustment is probably better for the system as a whole because variance in the demand is reduced, but determining the optimal concavity or discount that will induce the appropriate degree of smoothing is a topic for future research.

Third, following on this last observation, customers may see that it is possible to reduce their total cost by combining orders. When orders are combined, the concavity of the \( \pi(\cdot) \) function implies that total revenue to the service provider declines. This might reduce incentives of customers to acquire information about their own demand distribution.
Fourth, we note that our model is a single-period, non-repeated game. If there were opportunity to renegotiate at each period then the customer may not wish to reveal his entire distribution as we have shown, because it may leave him in a weaker bargaining position.

REFERENCES


**APPENDIX A: PROOFS**

**Proposition 1:** Given the $S(\cdot)$ defined from $X$ as in Equation (3), $\pi(\cdot)$ defined by Equation (4), and the actual demand is $d$, then

$$R(X) = \int_0^\infty \left[ \pi(S(x)) - S(x)\pi'(S(x)) \right] dx$$

$$U(X, d) = \int_0^d \pi'(S(x)) dx.$$  

**Proof:** Since $x_0 = 0$ and $x_n = \infty$, for $R(X)$ we have

$$\int_0^\infty \left[ \pi(S(x)) - S(x)\pi'(S(x)) \right] dx = \sum_{j=1}^n \int_{x_{j-1}}^{x_j} \left[ \pi(S(x)) - S(x)\pi'(S(x)) \right] dx.$$  

We emphasize that function $\pi(q)$ defines the $q_i$, where the cost minimizing contracts change. Now for every $x \in (x_{j-1}, x_j)$, $S(x) \in (q_j, q_{j-1})$, $\pi(S(x)) = r_j + u_j S(x)$ because $\pi(S(x))$ is linear on $(x_{j-1}, x_j]$. Because $\pi'(S(x)) = u_j$, on $(x_{j-1}, x_j]$, the integrand simplifies to $r_j$. Thus,

$$\int_0^\infty \left[ \pi(S(x)) - S(x)\pi'(S(x)) \right] dx = \sum_{j=1}^n \int_{x_{j-1}}^{x_j} [r_j + u_j S(x) - S(x)u_j] dx$$

$$= \sum_{j=1}^n \int_{x_{j-1}}^{x_j} r_j dx$$

$$= \sum_{j=1}^n r_j(x_j - x_{j-1})$$

$$= R(X).$$
Similarly for $U(X, d)$, let $k$ be such that $x_k < d \leq x_{k+1}$. Then we have

$$
\int_0^d \pi'(S(x))dx = \sum_{j=1}^{k} \int_{x_{j-1}}^{x_j} \pi'(S(x))dx + \int_{x_k}^{d} \pi'(S(x))dx
$$

$$
= \sum_{j=1}^{k} \int_{x_{j-1}}^{x_j} u_j dx + \int_{x_k}^{d} u_{k+1} dx
$$

$$
= \sum_{j=1}^{k} u_j (x_j - x_{j-1}) + u_{k+1} (d - x_k)
$$

$$
= \sum_{j=1}^{n} u_j [\min(x_j, d) - \min(x_{j-1}, d)]
$$

$$
= U(X, d).
$$

This completes the calculations. $\square$

**Proposition 2: Incentive Compatibility.** Let $S$ denote the class of possible contract schedules, which will be functionally equivalent to the class of (left continuous) complementary cumulative tail distribution functions. Consider a customer with a demand $D$ with a complementary c.d.f. $Q(x)$. This customer would then minimize his expected fee by picking $S = Q$. In other words, the optimization

$$
\Pi^* \equiv \min_{S \in S} [\Pi(S, D) = R(S) + E[U(S, D)]]
$$

is solved by $S^* = Q$, and its optimal value $\Pi^*$ is

$$
\Pi^* = \int_0^\infty \pi(Q(x))dx.
$$

**Proof:** Note that we can rewrite

$$
\Pi(S, D) = R(S) + E[U(S, D)]
$$

$$
= \int_0^\infty \left[ \pi(S(x)) - \pi'(S(x))S(x) \right] dx + \int_0^\infty \int_0^d \pi'(S(x)) dx \, dP(d).
$$

Interchanging the order of integration in the second integral yields

$$
\Pi(S, D) = \int_0^\infty \left[ \pi(S(x)) - \pi'(S(x))S(x) \right] dx + \int_0^\infty \int_x^\infty dP(d) \pi'(S(x)) dx
$$

$$
= \int_0^\infty \left[ \pi(S(x)) - \pi'(S(x)) [S(x) - Q(x)] \right] dx.
$$

Note that when $S = Q$, this will be simply

$$
\Pi(Q, D) = \int_0^\infty \pi(Q(x))dx.
$$
Now in order to show that \( S^* = Q \) we need to show that \( \Pi(Q) \leq \Pi(S) \) for all \( S \). Making the above substitution yields
\[
\int_0^\infty \pi(Q(x)) \, dx \leq \int_0^\infty \left[ \pi(S(x)) - \pi'(S(x)) [S(x) - Q(x)] \right] \, dx \quad (A1)
\]
or
\[
\int_0^\infty \left[ \pi(Q(x)) - \pi(S(x)) + \pi'(S(x)) [S(x) - Q(x)] \right] \, dx \leq 0.
\]

We now consider the integrand for an \( x \) where \( Q(x) \neq S(x) \); for clarity we drop the explicit dependence on \( x \). The integrand can be rewritten as
\[
\pi(Q) - \left[ \pi(S) + \pi'(S)(Q - S) \right]. \quad (A2)
\]

Note that the term within brackets is the tangent to \( \pi(\cdot) \) at \( S \). Because \( \pi(\cdot) \) is concave, the tangent line will exceed or equal the functional value for all \( Q \neq S \). Thus, Equation (A2) is nonpositive, showing that Equation (A1) is satisfied and that setting \( S^* = Q \) minimizes the customer’s expected costs. If we assume that \( \pi(\cdot) \) is strictly concave, then \( Q \) will be the unique optimal solution in \( S \). To prove this, observe that if \( S \) is different from \( Q \) at some \( x \), then there is a nonempty interval \([a, x]\) for which \( S \) differs from \( Q \) because they are left continuous. If \( \pi(\cdot) \) is strictly concave and thus strictly increasing and \( S \) differs from \( Q \) on an interval \([a, x]\), then Equation (7) will be strictly negative for any value in \([a, x]\), so the inequality in Equation (6) will strictly hold. Thus, the optimal \( S \) is unique in the class \( S \) of (left continuous) complementary c.d.f.’s. \( \square \)

**Proposition 4:** Consider some given family of location-scale probability distributions for nonnegative demand, parameterized by \( \mu \) and \( \sigma \). Then for a fixed set of contracts \( \pi(\cdot) \), there exist constants \( a > 0 \) and \( b > 0 \) such that for all \((\mu, \sigma)\), \( \Pi_{\mu, \sigma} = a\mu + b\sigma \). That is, all other things being equal, the expected fee is a linear function of \( \mu \) and \( \sigma \).

**Proof:** Consider a family of distributions of nonnegative random variables with mean and standard deviation of \( \mu \) and \( \sigma \), respectively. Let \( Q_{\mu, \sigma} \) be the complementary c.d.f. of the distribution with mean \( \mu \) and standard deviation \( \sigma \). This set of distributions is related by
\[
Q_{\mu, \sigma}(\mu + \sigma z) = Q_{0,1}(z)
\]
for all \( z \), for all \((\mu, \sigma) \in R \subset \mathbb{R}_2\), where \( R \) defines the region for which this equivalence holds and demand is nonnegative. For example, consider the set of demand distributions that are normally distributed and truncated at \(-2.5\sigma \). Then in order for demand to be nonnegative, \( R \) will in this case be defined as
\[
R = \{ (\mu, \sigma) | 0 \leq 2.5\sigma \leq \mu \}. 
\]
Define $c$ to be the supremum, over $R$, of the coefficient of variation $\sigma/\mu$; in the above example, $c = 0.4$. Since demand is nonnegative,

\[ Q_{\mu,c\mu}(0) = 1 \Rightarrow Q_{0,1}(-1/c) = 1. \]

Note that although we consider only nonnegative random variables, our analysis is based on the standard normalization with mean 0 and standard deviation 1, for which the associated random variable will take on negative values between $-1/c$ and 0. Now consider any $(\mu, \sigma) \in R$, and define $\Pi^*_{\mu,\sigma}$ to be the optimal expected cost for the customer whose demand follows a probability distribution with mean $\mu$ and standard deviation $\sigma$,

\[ \Pi^*_{\mu,\sigma} = \int_0^\infty \pi(Q_{\mu,\sigma}(x))dx. \]

Redefining the variable of integration to be $z \equiv (x - \mu)/\sigma$ yields

\[ \Pi^*_{\mu,\sigma} = \sigma \int_{-\mu/\sigma}^{\infty} \pi(Q_{0,1}(z))dz. \]

By definition of $c$ and $R$, $-\mu/\sigma \leq -1/c$. For $(\mu, \sigma) \in R$, $Q_{\mu,\sigma}(0) = Q_{0,1}(-\mu/\sigma) \geq Q_{0,1}(-1/c) = 1$. Thus

\[
\begin{align*}
\Pi^*_{\mu,\sigma} &= \sigma \int_{-\mu/\sigma}^{-1/c} \pi(Q_{0,1}(z))dz + \sigma \int_{-1/c}^{\infty} \pi(Q_{0,1}(z))dz \\
&= (\mu - \sigma/c)\pi(1) + \sigma \int_{-1/c}^{\infty} \pi(Q_{0,1}(z))dz \\
&= \mu\pi(1) + \sigma \left[ \int_{-1/c}^{\infty} \pi(Q_{0,1}(z))dz - \pi(1)/c \right].
\end{align*}
\]

Defining $a = \pi(1)$ and $b = \int_{-1/c}^{\infty} \pi(Q_{0,1}(z))dz - \pi(1)/c$ completes the proof. □

**Proposition 5:** The expected cost paid by a customer, $\Pi^*$, is bounded both above and below as

\[ \pi(1)E(D) \leq \Pi^* \leq \pi'(0)E(D). \]

**Proof:** Note that for $q \in [0, 1]$, $\pi(1)q$ represents the chord connecting $0, \pi(0)$ = $(0, 0)$ and $(1, \pi(1))$. Because $\pi(\cdot)$ is concave, this chord will fall below $\pi$ for $q \in [0, 1]$. Thus

\[ \Pi^* = \int_0^\infty \pi(Q(x))dx \geq \int_0^\infty \pi(1)Q(x)dx = \pi(1)E(D). \]

The tangent line to $\pi(\cdot)$ at 0 will just be $\pi'(0)q$, which exceeds $\pi(q)$ because $\pi$ is concave. Thus, $\Pi^* = \int_0^\infty \pi(Q(x))dx \leq \int_0^\infty \pi'(0)Q(x)dx = \pi'(0)E(D)$. □
**Proposition 6:** There exists a choice of \( \pi \) that allows the service provider to learn the entire demand distribution of the customer and obtain a revenue arbitrarily close to the maximum possible, \( pE(D) \).

**Proof:** Define \( \pi(q) = p\left(\frac{(q+1)^{\varepsilon+1}-1}{1-\varepsilon}\right) \), and recall that \( p \) is the willingness of the customer to pay for a unit of service. It is easily verified that \( \pi(0) = 0 \), \( \pi'(0) = p \), \( \pi(1) = p\left(\frac{1}{1-\varepsilon}\right) \leq p \), and \( \pi(1) \to p \) as \( \varepsilon \to 0 \). Thus, from the lower bound in Proposition 3, the revenue generated from the customer can be made arbitrarily close to \( pE(D) \). \( \square \)

**APPENDIX B: TABLE OF NOTATION**

**Customer Information**

\( p \) The customer’s reservation price for a unit of service.
\( D \) The random demand of the customer.
\( d \) A realization of \( D \).
\( P(x) \) The cumulative distribution of \( D \).
\( Q(x) = 1 - P(x) \), the complementary cumulative distribution of \( D \).
\( q \) A quantile of the complementary cumulative distribution.

**Contracts**

\( j \) Index for contracts.
\( n \) The number of contracts.
\( y_j \) The amount of contract \( j \) purchased by the customer.
\( x_j \) The cumulative amount of contracts \( 1, \ldots, j \) purchased by the customer.
\( X = (x_0, \ldots, x_n) \), the vector of cumulative contract choices.
\( X(D) \) The vector of cumulative contract choices based on a demand distribution \( D \).
\( S \) An alternative way to describe the contract choice which is in the form of a complementary cumulative probability distribution.

**Fees for Customer Associated with Contracts**

\( r_j \) Up-front fee for a unit of contract \( j \).
\( u_j \) As-used fee for a unit of contract \( j \).
\( \pi(q) \) An alternative representation of the up-front and as-used fees.
\( R(X) \) (or \( R(S) \)) The up-front fee associated with a contract choice \( X \) (or \( S \)).
\( U(X,d) \) (or \( U(S,d) \)) The as-used fee associated with a contract choice \( X \) (or \( S \)) when the demand is \( d \).
\( \Pi(X, D) = R(X) + E(U(X, D)) \) (or \( \Pi(S) = R(S) + E(U(X, D)) \)) the expected fee for the customer (or revenue for the firm) associated with a contract choice \( X \) (or \( S \)).
\( \Pi^* \) The optimal value of \( \Pi(S) \) over all possible \( S \).

**Capacity Types**

\( i \) Index for types of capacity.
\( m \) The number of types of capacity.
A Service Provider’s Elicitation

\( z_i \)  The number of units of capacity type \( i \) purchased by the service provider.

\( w_i \)  The cumulative number of units of capacity type \( i \) purchased by the service provider.

\( W = (w_0, \ldots, w_n) \), the vector of cumulative capacity choices.

\( W(D) \)  The vector of cumulative capacity choices based on a customer demand \( D \).

**Costs Associated with Capacity Types**

\( h_i \)  The cost of having a unit of capacity of type \( i \).

\( c_i \)  The cost of using a unit of capacity of type \( i \).

\( \gamma(q) \)  An alternative representation of the having and using costs of capacity.

\( H(W) \)  The cost of having the capacity associated with capacities \( W \).

\( C(W,d) \)  The cost of using the capacity associated with capacities \( W \), when demand is \( d \).

\( \Gamma(W, D) = H(W) + E(C(W, D)) \), the expected costs associated with capacities \( W \).

**Gregory Dobson** is an associate professor of operations management at the William E. Simon Graduate School of Business Administration, University of Rochester. He holds a BS in operations research and industrial engineering from Cornell University and PhD in operations research from Stanford University. His current work concentrates on the application of queueing theory and process improvement principles, in particular Six-Sigma, to health care and other service industries. Past research includes work in the areas of product line design, and batching and scheduling in manufacturing.

**Phillip J. Lederer** is associate professor of operations management at the William E. Simon Graduate School of Business Administration, University of Rochester. He holds a BS in physics from SUNY Stony Brook and an MS and PhD in applied mathematics from Northwestern University. His research interests lie in the interface of operations management and economics. His past research has spanned the areas of location and network competition, economic evaluation of operations capital projects, pricing for services, and measurement of operations performance. His papers have appeared in such journals as *Econometrica, Operations Research, Transportation Science, International Journal of Flexible Manufacturing Systems, Decision Sciences*, and many others.

**Lawrence W. Robinson** is a professor of operations management at Cornell University’s Johnson Graduate School of Management. He received his MBA and PhD degrees from the Graduate School of Business at the University of Chicago in 1986. His research focuses on problems of operating in an uncertain service environment; in particular, on developing practical heuristic policies which perform well and can be easily calculated. He is a member of INFORMS and Rotary.
Queries

Q1 Author: Please provide the keywords for this article.

Q2 Author: Please identify the corresponding author.

Q3 Author: Please check the suggested Running head for correctness or provide a suitable one.

Q4 Author: McCarthy (1956) has not been included in the Reference List, please supply full publication details.

Q5 Author: Please check the reference Szekely (2003) has been changed to Szekely (2000) to match the reference list.

Q6 Author: Please provide the last access date for the link provided in reference Buja et al. (2005).

Q7 Author: Please provide the publisher location for reference Cervera & Munoz (1996).

Q8 Author: Please check the reference Thomson (1979) has not been cited in the text, please cite it in appropriate position or delete from the reference list.

Q9 Author: Please check the reference Wilson (1993) has not been cited in the text, please cite it in appropriate position or delete from the reference list.