Risk-Sensitive Sizing of Responsive Facilities

Sergio Chayet,[1] Wallace J. Hopp[2]

[1] *Olin Business School, Washington University in St. Louis, St. Louis, Missouri 63130*

[2] *Stephen M. Ross School of Business, University of Michigan, Ann Arbor, Michigan 48109*

Received 2 July 2004; revised 21 December 2007; accepted 27 December 2007
DOI 10.1002/nav.20278
Published online 19 February 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: We develop a risk-sensitive strategic facility sizing model that makes use of readily obtainable data and addresses both capacity and responsiveness considerations. We focus on facilities whose original size cannot be adjusted over time and limits the total production equipment they can hold, which is added sequentially during a finite planning horizon. The model is parsimonious by design for compatibility with the nature of available data during early planning stages. We model demand via a univariate random variable with arbitrary forecast profiles for equipment expansion, and assume the supporting equipment additions are continuous and decided ex-post. Under constant absolute risk aversion, operating profits are the closed-form solution to a nontrivial linear program, thus characterizing the sizing decision via a single first-order condition. This solution has several desired features, including the optimal facility size being eventually decreasing in forecast uncertainty and decreasing in risk aversion, as well as being generally robust to demand forecast uncertainty and cost errors. We provide structural results and show that ignoring risk considerations can lead to poor facility sizing decisions that deteriorate with increased forecast uncertainty. Existing models ignore risk considerations and assume the facility size can be adjusted over time, effectively shortening the planning horizon. Our main contribution is in addressing the problem that arises when that assumption is relaxed and, as a result, risk sensitivity and the challenges introduced by longer planning horizons and higher uncertainty must be considered. Finally, we derive accurate spreadsheet-implementable approximations to the optimal solution, which make this model a practical capacity planning tool. Naval Research Logistics 55: 218–233, 2008

Keywords: capacity planning; long-term planning; responsiveness; risk

1. INTRODUCTION

Capacity planning decisions are a critical part of a firm's manufacturing strategy, affecting responsiveness, flexibility, exposure to risk, ability to meet customer needs and many other tactical issues. This article considers the facility sizing decision faced by a risk-averse agent. Once deployed, the facility size cannot be adjusted, and limits the quantity of production equipment in the plant. During the useful life of the facility, equipment can be added sequentially to address demand growth and control delivery leadtimes. This problem is most relevant in high-tech industries such as semiconductors and bio-technology, where the aforementioned conditions prevail and demand is highly uncertain. The nonadjustable nature of the sizing decision and prolonged facility lives (typically five to seven years) give rise to long planning horizons, which combined with high demand uncertainty make the agent's decision a challenging problem. In this article we develop a model for finding the optimal facility size by maximizing the expected utility of profits, which include investment in both the initial facility as well as subsequent equipment additions. For consistency with the level of data available at the time of the sizing decision, we model multi-period demand using a univariate random variable and arbitrary demand profiles, assuming ex-post and continuous equipment additions. Under these assumptions and constant absolute risk aversion (CARA), operating profits are the closed-form solution to a nontrivial linear program, which allows to characterize the sizing decision via a single first-order condition.

There is a vast literature on capacity planning under uncertainty. With a few exceptions, we restrict our review to papers studying single-agent, single-location optimal capacity level decisions. For a recent comprehensive review, including timing of capacity adjustments, multiple capacity types, multiple agents, and more, see Van Mieghem [31].

Correspondence to: Sergio Chayet (chayet@wustl.edu)

Most capacity planning studies assume risk-neutrality, including articles analyzing equipment addition. Swaminathan [28] studies multi-period equipment procurement with a scenario-based mixed-integer stochastic program which does not directly model responsiveness. Several papers study single-period equipment configuration using queuing models to estimate cycle times. For instance, Hopp et al. [18] develop an optimization model that minimizes tool-set cost subject to throughput and cycle time constraints, evaluated using a queuing network model and a decomposition solution approach. Other examples in this group include Bard et al. [6], Bitran and Tirupati [10], Suri et al. [27], Rajagopalan and Yu [23], and Benjaafar [8]. Another line of research studies continuous-time models to determine the timing and scale of capacity expansions, as in Ryan [25] (and references therein), who considers fixed installation leadtimes. She incorporates responsiveness by developing a policy that provides a specified service level. Within the semiconductor capacity planning literature, a few authors have proposed modular designs, which spread the facility sizing decision over time to reduce initial capital expenditures and facilitate better alignment between capacity and demand. Benavides et al. [7] calculate facility expansion times and scales using a cash flow model that ignores operational details. Angelus et al. [3] obtain (s, S) type structural results for a finite horizon capacity expansion model with fixed installation lead times. Cakanyildirim et al. [11] provide a bottleneck-based algorithm to determine tool expansion and contraction times, along with modular facility expansions. Although modular expansions may be practical in some industries, they remain to be proven as a practical alternative in high technology capital intensive settings including semiconductor manufacturing. See Wu et al. [30] for a recent review of capacity planning in the high-tech industry. These capacity-planning models incorporate operational responsiveness requirements, but their degree of detail requires a level of data accuracy available only for planning horizons associated with modular facility expansions or equipment addition. In this article we develop a model to support facility sizing decisions made over the longer planning horizons imposed by nonmodular designs. Even though the model addresses responsiveness requirements in subsequent equipment addition decisions, its parsimony and transparency make it compatible with data availability associated with longer planning horizons. As far as we are aware, this is the first model combining these two features. The solution developed below is robust to cost estimate errors and easily implementable, which makes the model compelling as a decision-support tool.

From a corporate finance perspective it can be argued that publicly traded firms should make capacity decisions to maximize their market capitalization while adopting a risk-neutral attitude towards unsystematic or "non-diversifiable" risk. However, this conclusion relies on strong assumptions about capital markets, and if the market has imperfections, firms should be risk-sensitive [31]. Still, the literature on risk-sensitive capacity planning by single agents is quite limited and mostly restricted to inventory settings. Several authors have studied risk-sensitive newsvendor models. Risk preferences are usually incorporated into single-agent models (inventory and otherwise) either using variations of a mean-variance formulation or via an explicit utility function. Under mean-variance preferences, the agent maximizes expected profits minus a fixed parameter times their standard deviation, where the parameter is an measure of risk aversion. Lau [21] and Chen and Federgruen [12] use this approach for several basic inventory models and find the effect of risk aversion on standard policies. The mean-variance method is attractive for its simplicity, but is seriously limited by symmetrically treating positive and negative deviations of operating profits from the mean. This has led to alternative heuristic formulations. Sankarasubramanian and Kumaraswamy [26] maximize the probability of exceeding a target profit for specific demand distributions, and Li et al. [22] extended that model to study two products with uniformly distributed demands. Additional formulations can be found in Anvari [4], Chung [13], and Anvari and Kusy [5]. Another approach is to maximize the expected von Neumann-Morgenstern utility. Although it involves an explicit utility function, this method is more general and does not assume a symmetric distribution of operating profits. Indeed, the expected utility formulation under the assumptions of perfect capital markets, CARA, and normally distributed operating profits, is equivalent to the mean-variance formulation [31]. Lau [21] calculates the optimal order quantity for the risk-averse newsvendor assuming a polynomial approximation to a general utility. Eeckhoudt et al. [15] consider a newsvendor model with general risk-averse utility function and general demand distribution, and examine the sensitivity of the solution to changes in the various price and cost parameters. They show that the optimal value is always decreasing in risk aversion for general concave utility functions. Van Mieghem [32] extends their analysis to a newsvendor network. Agrawal and Seshadri [2] analyze quantity and pricing decisions of a newsvendor with general concave risk-averse utility. They show that in comparison to a risk-neutral newsvendor, the risk-averse newsvendor's order quantity is lower while her price can be higher or lower depending on how pricing affects the demand distribution. We use the expected utility method, and following Howard [19] and Walls and Dyer [29], we assume the existence of a risk-averse corporate utility function, which is independent of any external effects on the firm's equity. For practical purposes, we generally assume CARA corporate preferences and thus ignore wealth effects. This article extends the risk-sensitive newsvendor model to nonmodular facility sizing with subsequent equipment additions, the

two models coinciding under the assumption of a stationary demand over the planning horizon. Along with other structural results, we show that the feature of the risk-sensitive newsvendor's optimal decision being decreasing in risk aversion shown by Eeckhoudt et al. [15], extends to our setting; a desired property for a decision support tool.

We have limited our review of risk-sensitive models to research that is most closely related to ours. But as evidenced by Gan et al. [16] and the references therein, this area has recently received increased attention, with particular emphasis on multiple agents in supply chains, and to a lesser extent on multiperiod inventory models.

An alternative approach for incorporating risk sensitivity to decision-making is via financial or operational hedging, where instead of maximizing the expected utility reflecting risk-adjusted profits, the focus is on mitigating risk without substantially affecting expected profits. Birge [9] analyzes how to integrate financial hedging in a linear capacity investment model, while Gaur and Seshadri [17] show the viability of financial hedging using instruments correlated with the firm's profits. Financial hedging can be effective, but it requires specific tradable instruments, which not always exist. Ding et al. [14] study integrated operational and financial hedging decisions faced by a global firm selling in domestic and foreign markets subject to currency exchange rate risk, and analyze the relationship between financial and operational hedges.

The goal of this article is to develop a model to support facility sizing decisions affecting future equipment additions. The model is sensitive to the risk inherent in the capacity decision and anticipates the operational context it will influence. Its parsimony makes it both capable of yielding structural results and consistent with the level of data available for long-term planning; thus, an effective decision support tool. The rest of the article is organized as follows. In Section 2 we formulate a modeling framework for capacity planning that incorporates the requirements for designing responsive plants. In Section 3 we derive structural results for the optimal solution. In Section 4 we describe the steps necessary for using our models in practice, including two alternative approximations along with numerical tests of their accuracy. The article concludes in Section 5.

2. MODEL FORMULATION

2.1. Problem Description

Strategic capacity planning for a new production facility can be viewed as a single facility sizing (FS) decision followed by a series of equipment addition (EA) decisions. The FS decision establishes the size (floorspace) of the plant, while the EA decisions populate it with equipment. Our focus is primarily on the FS decision. In particular, we concentrate on the variable part of the FS decision (i.e., floorspace proportional to the number of machines to be installed) as opposed to the fixed part (i.e., floorspace for administration, utilities, etc.).

We assume fixed products and production processes, and a finite number of EAs which take place at predefined time periods. We also assume a known utility function reflecting the firm's risk attitude. The problem consists of selecting a FS, and an EA sequence consistent with it, to maximize expected utility of profits with respect to a demand forecast.

An important consequence of capacity decisions is the responsiveness of the production facility. One way to consider this in a model is to assume the existence of a constraint on cycle time. If demand is known with a high degree of certainty (e.g., at the EA decision level), queueing network approximations can be used to evaluate cycle time (see e.g., Bitran and Tirupati [10], Suri et al. [27] or Hopp et al. [18]). Unfortunately, the same approximations cannot be used at the FS decision level because at this early decision phase demand is highly uncertain. So, as a proxy for a cycle time constraint we assume a limit on utilization for all resources. By capping utilization we limit the amount of queueing that can occur. In practice, and depending upon data availability, the utilization limits can be obtained by (1) using simulation or analytic models (e.g., Bitran and Tirupati [10] or Hopp et al. [18]) for a representative equipment configuration, (2) using historical utilizations of similar stations in existing facilities, or (3) setting uniform utilization limits across all stations (our tests indicate that 70% to 85% is appropriate in most cases). It is worth noting that a major semiconductor manufacturer uses precisely this type of utilization constraint in sizing their wafer fabs.

2.2. Model Formulation

To address the above problem we formulate a two-stage model. In the first stage a demand forecast is generated for a set of future expansion times $t_1, t_2, \ldots, t_T$ and the FS decision is made. In the second stage demands for all t_i are revealed simultaneously and an optimal EA schedule, subject to the FS decision, is generated. Although in reality demands are revealed sequentially, we approximate the timing of the EA decisions in this manner for purposes of modeling the FS decision. Note that in this framework, the first EA corresponds to the initial equipment configuration.

Within the plant we assume each product follows a deterministic routing, visiting some of the N different stations in the facility. Processes and technologies are fixed over the planning horizon and we assume that every station is part of at least one product's routing. We define

S_p = number of total steps in the routing of product p $(1 \le p \le P)$,
n_{ps} = station visited by product p, on the sth step of its routing $(1 \le s \le S_p)$,
$\hat{\alpha}_{ps}$ = surviving fraction of incoming parts of product p, due to yield loss at station $n_{ps}(0 < \hat{\alpha}_{ps} \le 1)$,
$\alpha_{ps} = \prod_{j=1}^{s} \hat{\alpha}_{pj}$ = cumulative yield of product p after completing the first s steps in its routing, with $\alpha_{p0} = 1$ $(0 \le s \le S_p)$,
τ_{ps} = effective mean process-time for the sth step of product p,
a_n = floorspace requirement per tool (including aisle, support, etc.) at station n $(1 \le n \le N)$,
u_n = maximum utilization allowed at station n $(1 \le n \le N)$,
k = net cost per unit floorspace for the duration of the planning horizon discounted to the time of the FS decision (includes salvage value),
c_{nt} = marginal installation cost per additional tool for periods t through T at station n discounted to the time of the FS decision (includes salvage value),
r_{pt} = present value of net revenue per unit throughput of product p for period t,
z = total floorspace (primary (FS) decision variable),
x_{nt} = number of tools added at station n for period t (secondary (EA) decision variable),
λ_{pt} = release rate of product p during period t (secondary decision variable).

The demand rate for product p in period t is $f_p q_t D$, for $1 \le p \le P$, $1 \le t \le T$, where the q_t and f_p, which are positive with $\sum_{t=1}^{T} q_t = \sum_{p=1}^{P} f_p = 1$, constitute a *demand profile*, and D is a random variable representing total demand rate across all products and installation periods. A demand forecast consists of a demand profile together with estimates of the median m and coefficient of variation (cv) v of D. For a measure of central tendency, we choose the median instead of the mean because we believe its compatibility with scenarios makes it easier for the forecaster to estimate, and also because it is not subject to a disproportionate influence of rare events with extreme values. In this context, the cv can be interpreted as a measure of confidence in the forecaster's prediction of D. Note that the demand rate expression implies that the proportion of demand for each product remains constant across all periods. We also assume a fixed product mix, and $\lambda_{pt}\alpha_p = f_p\theta_t$, for $1 \le p \le P$, $1 \le t \le T$, where θ_t is the aggregate throughput for period t and $\alpha_p = \alpha_{pS_p}$ is the total cumulative yield of product p. (Note that the λ_{pt} represent *release* rates, while the θ_t represent throughput or *output* rates.)

To formulate the model in terms of aggregate product, a unit of which consists of f_p units of product p for all products, we define

$$r_t = \sum_{p=1}^{P} f_p r_{pt}, \quad 1 \le t \le T$$

$$\tau_n = \sum_{p=1}^{P} \frac{f_p}{\alpha_p} \sum_{s=1}^{S_p} \alpha_{p,s-1}\tau_{ps}[n_{ps} = n], \quad 1 \le n \le N,$$

where the notation $[A]$ stands for 1 if statement A is true and 0 otherwise. Note that τ_n is the average process time at station n adjusted for yield. Since the number of tools at station n in period t is $\sum_{i=1}^{t} x_{ni}$, the utilization constraints are

$$\frac{\theta_t \tau_n}{\sum_{i=1}^{t} x_{ni}} \le u_n, \quad 1 \le n \le N, 1 \le t \le T, \tag{1}$$

where the left-hand side of (1) is station n's utilization in period t. Let

$$J_n = \frac{\tau_n}{u_n}, \quad 1 \le n \le N, \tag{2}$$

When (1) is binding, $\theta_t J_n = \sum_{i=1}^{t} x_{ni}$ is the number of tools required at that station to achieve utilization u_n. Therefore, if station n's utilization is u_n for every n, the total amount of floorspace required is $\theta_t a$, where

$$a = \sum_{n=1}^{N} a_n J_n.$$

Given D, the discounted net revenues generated by a facility of size z (excluding floorspace costs), can be represented as a solution to the following optimization problem (**P**):

$$R(z|D) = \max \quad \sum_{t=1}^{T} r_t\theta_t - \sum_{t=1}^{T}\sum_{n=1}^{N} c_{nt}x_{nt}$$

$$\text{s.t.} \quad \theta_t J_n - \sum_{i=1}^{t} x_{ni} \le 0, \quad 1 \le n \le N, 1 \le t \le T \tag{3}$$

$$\theta_t \le q_t D, \qquad 1 \le t \le T \tag{4}$$

$$\sum_{t=1}^{T}\sum_{n=1}^{N} x_{nt}a_n \le z \tag{5}$$

$$\theta_t, x_{nt} \ge 0, \quad 1 \le n \le N, 1 \le t \le T.$$

Note that (3) are the maximum utilization constraints (1) rewritten using (2). Constraints (4) limit throughput to available demand, and (5) is the floorspace constraint.

Hence, the optimal FS decision z^* is the floorspace that maximizes expected utility of net profit, that is

$$z^* = \max_{z \ge 0} EU[R(z|D) - kz]. \tag{6}$$

Note that the FS and EA decisions are linked via equation (6) and problem (**P**). The latter ignores any integrality conditions for the equipment capacity variables x_{nt}. The main reasons for this choice are tractability and robustness of the $R(\cdot|D)$ function. These are primary concerns in an environment of highly uncertain data like ours. If the optimal solution z^* is not a sum of multiples of the floorspace requirements a_n then, for implementation purposes, a more practical floorspace decision may be obtained by fine-tuning the utilization parameters u_n. But even without this, a solution obtained by the methods we describe here captures the tradeoffs among the main decision factors. The final decision will necessarily have to be adjusted in accordance to additional practical requirements not considered in any long-term strategic model.

3. STRUCTURAL RESULTS

3.1. Flat Demand Profiles

We begin by considering the simplest case where the forecast is *flat*, that is $q_t = q$ for all t. In some instances, a flat demand profile may be consistent with anticipated market conditions at the time of forecasting. In others, it is an additional approximation of reality. However, even in cases where such a profile is not expected, the forecast may be too uncertain to build in any other shape. We treat the flat demand case for its own sake in this section; in Section 4 we use flat profiles to develop two approximations for the general profile case. Without loss of generality, we assume $a = 1$, which is equivalent to selecting convenient floorspace units.

To find the optimal floorspace, we first compute the expected utility in equation (6) as a function of the decision variable z. To do this, we condition on the random component of the demand D in the following lemma (proof in appendix).

LEMMA 1: If $q_t = q$ for all t, given D, the discounted net revenue generated by a facility of size z (excluding floorspace costs) is

$$R(z|D) = r \min(z, qD), \tag{7}$$

where $r = \max\{\sum_{i=t}^{T} r_i - \sum_{n=1}^{N} J_n c_{nt}^* : 1 \le t \le T\}^+$ is the dual price associated with the floorspace variable z, and $c_{nt}^* = \min\{c_{ni} : 1 \le i \le t\}$.

With closed-form expression (7) available, we can use equation (6) to calculate the optimal floorspace; which is our main goal. Let $F(\cdot)$ represent the cdf of the demand variable D, and $U(\cdot)$ the utility function of a risk-averse decision maker. We assume F to be continuous and U to be concave and twice differentiable with $U' > 0$. Note that $R(\cdot|D)$ is concave for every D, and by the assumptions on the utility function, so is $EU[R(z|D) - kz]$. Therefore, the optimal floorspace z^* is well-defined in terms of equation (6). Computing z^* is just a matter of solving the first order condition, which, denoting $dEU[r \min(z, qD) - kz]/dz$ by $\varphi(z)$ can be written as

$$\varphi(z) = -k \int_0^{z/q} U'(rqy - kz) dF(y) + (r-k) U'((r-k)z) \overline{F}(z/q) = 0, \tag{8}$$

where $\overline{F}$ stands for $1 - F$. If $\varphi(z)$ has a zero, the concavity of the expected utility guarantees its uniqueness. On the other hand, if $\varphi(z) < 0$ for all $z > 0$, the solution is $z^* = 0$, and if $\varphi(z) > 0$ for all $z \ge 0$ the solution is $z^* = \infty$. When the decision-maker is risk-neutral (i.e., $U' = 1$), equation (8) yields the well-known newsvendor solution

$$F(z/q) = 1 - k/r. \tag{9}$$

In general, given $F(\cdot)$, $U(\cdot)$, r, k, and q, one can compute the root of equation (8) using standard numerical methods. In Section 4 we present results for specific functional forms of F and U.

The model defined by equations (6)–(8) can be thought of as a risk-sensitive version of the classic newsvendor problem. CARA preferences can be represented by a utility function of the form $U(x) = -e^{-\gamma x}$, where $\gamma = -U''(x)/U'(x)$ is the degree of risk aversion (see e.g., Keeney and Raiffa [20] p.167). Adopting the CARA assumption allows us to consider compactly the effects of some parameter changes on the FS decision. We summarize the key results in the following theorem.

THEOREM 1: If $r - k > 0$ and $q_t = q$ for all t, the optimal FS, z^*, is increasing in q for any risk-averse preferences, and decreasing in k for risk-averse preferences that exhibit CARA.

PROOF: From equation (8) and the concavity of the expected utility, $\varphi'(z) < 0$ for all z. Hence, to show that $\partial z^*/\partial x \ge 0 (\le 0)$ it is sufficient to show that $\partial \varphi(z^*)/\partial x \ge 0 (\le 0)$. The condition $r - k > 0$ implies that $\varphi(z)$ in (8) is increasing in q, which proves the first result. To prove the second result, assume CARA and use $U(x) = -e^{-\gamma x}$, with $\gamma \ge 0$. Substituting into (8) and carrying out the calculations yields the desired result that $\partial \varphi(z^*)/\partial k \le 0$. □

Note that the requirement $r - k > 0$ simply means that it is profitable to build the facility. To understand the second part of Theorem 1 intuitively, observe that when the floorspace cost k decreases, increasing the facility size by the same amount allows the capture of additional revenue without additional risks. From equation (9) it follows that the same result also holds in the risk-neutral case.

The first part of Theorem 1 implicitly assumes r and k remain fixed while q varies. However, when expressed in terms of original parameters, $q = 1/T$, and since r also depends on T, changes in q alter r. In terms of original parameters, the effect of T on the optimal floorspace z^* involves both $\partial z^*/\partial q$ and $\partial z^*/\partial r$. Before considering that effect, we offer an alternative interpretation for the first part of Theorem 1. Specifically, if we ignore the relation between q and T, from equation (7) note that q can represent a scaling parameter of demand, i.e., $D_q = qD$. Thus, increasing q while keeping all other parameters fixed is equivalent to increasing the demand across all sample paths, which leads to building a larger facility. It follows from equation (9) that this result also agrees with the risk-neutral case.

Let us now consider how the revenue parameter r affects z^*. Note that, as a function of T, from its definition in Lemma 1, $r(T+1) = \max\{\sum_{i=t}^{T+1} r_i - \sum_{n=1}^{N} J_n c_{nt}^* : 1 \le t \le T+1\}^+ = \max[\max\{\sum_{i=t}^{T} r_i + r_{T+1} - \sum_{n=1}^{N} J_n c_{nt}^* : 1 \le t \le T\}^+, r_{T+1} - \sum_{n=1}^{N} J_n c_{nT+1}^*] = r(T) + r_{T+1}$. Therefore, r increases in T. But it is also possible to vary r while T, and hence q, is fixed. To analyze that situation, note that the first term of $\varphi(z)$ in condition (8) is the marginal cost in expected utility of an increase in the order quantity z/q, and the second term is the marginal benefit. Although the benefit increases with r through the factor $r-k$, there is also the opposite effect of a decrease in both cost and benefit because of diminishing returns (i.e., lower U') for higher net revenues. Which of the two effects dominates depends upon the particular choice of parameters and preferences. Therefore, z^* is not necessarily increasing in r, as in the risk-neutral case, for which the result follows directly from equation (9). For the risk-averse case, see Eeckhoudt et al. [15] for sufficient conditions that guarantee z^* increasing in r; Neither CARA nor decreasing absolute risk aversion (i.e., $-U''(\cdot)/U'(\cdot)$ decreasing) are. Note that the authors show this for the newsvendor problem, which includes our flat profile problem as a special case.

Observe from the definition of r that any sensitivity result with respect to it can be traced back to the original set of r_t and c_{nt} parameters. However, we caution against singling out the specific parameters that affect r because this constitutes an a posteriori analysis on the optimal EA schedule, the modeling of which is a rough approximation whose only purpose is to support the FS decision. As for the effect of T on z^*, $dz^*/dT = (\partial z^*/\partial r)(\partial r/\partial T) + (\partial z^*/\partial q)(\partial q/\partial T)$, which is negative when z^* is decreasing in r because r and z^* are increasing in T and q respectively, and q is decreasing in T. Otherwise, the sign of dz^*/dT depends on whether $(\partial z^*/\partial r)(\partial r/\partial T)$ or $(\partial z^*/\partial q)(\partial q/\partial T)$ dominates.

In addition to comparative statics for cost, revenue, and planning horizon parameters, since risk-aversion is an important element of the model, it is sensible to determine how changes in risk attitude affect the FS decision z^*. It is straightforward to show that z^* decreases as risk-aversion increases, where an increase in risk aversion is equivalent to a concave transformation of the utility function (in particular, for CARA preferences with $U(x) = -e^{-\gamma x}$, an increase in risk aversion is achieved simply by increasing γ). This result is not surprising. The proof can be found in Eeckhoudt et al. [15] and we therefore omit it.

As we discussed earlier, of all the data required by our model the forecast is by far the most uncertain. For this reason, it is imperative to investigate the effect of forecast uncertainty on the FS decision z^*. Since demand forecasts tend to be highly uncertain for strategic capacity planning, the region of interest lies towards the higher end of cv values. So we begin by analyzing the behavior of z^* in the limit as v increases to infinity, while the median and other parameters remain fixed. This process involves the transformation of the distribution function F, which in general, is not uniquely determined by two parameters. To ensure the desired effect on the distribution in the limit, namely, a complete shift of probability mass towards the two tails, we restrict the class of distribution functions according to the following definition.

DEFINITION 1: A continuous distribution $F(\cdot|m,v)$, with median m and coefficient of variation v, satisfies the median-cv limit conditions if for every $m > 0$

1. $F(0|m,v) = 0$ for every $v > 0$,
2. $\lim_{v\to\infty} F(z|m,v) = \frac{1}{2}$ for all $z \in (0,\infty)$,
3. There exists a function $G(\cdot|m)$, integrable on $[0,\infty)$, and some $v_0 > 0$ such that $F(z|m,v) \le G(z|m)$ for all $z \in [0,\infty)$ and $v \ge v_0$.

The last condition of this definition is a technical requirement added for convenience in proofs. In practice, when time comes to evaluate z^* numerically, an explicit functional form for the distribution F will be required. Given the high degree of uncertainty, and the fact that the forecast information is limited to the median and cv, we opt for simplicity and recommend fitting a two-parameter distribution. Two reasonable choices for modeling demand forecasts are the lognormal and gamma. The lognormal distribution function with median m and cv v is

$$F(z|m,v) = \Phi\left(\frac{\log(z/m)}{\eta(v)}\right), \quad z > 0, \tag{10}$$

where $\eta(v) = \sqrt{\log(v^2+1)}$ and $\Phi(x) = (1/\sqrt{2\pi}) \int_{-\infty}^{x} e^{-u^2/2} du$ is the standard normal distribution function. The gamma distribution function with median m and cv v

is most conveniently defined in terms of the complimentary function

$$\overline{F}(z|m,v) = Q\left(\eta(v), \frac{z}{m}Q^{-1}(\eta(v),1/2)\right), \quad z \geq 0, \quad (11)$$

where $\eta(v) = 1/v^2$ and $Q(\eta,x) = (1/\Gamma(\eta))\int_x^\infty u^{\eta-1}e^{-u}du$ is the regularized incomplete gamma function, with $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$ the Euler gamma function (see e.g., Abramowitz and Stegun [1]). The inverse function $x = Q^{-1}(\eta,1/2)$ is defined as the solution to $Q(\eta,x) = 1/2$. With these definitions, we can formally state the following result (proof in appendix).

LEMMA 2: The lognormal and gamma distributions as defined by equations (10) and (11) satisfy the median-cv limit conditions.

The limiting behavior of the optimal FS decision as the cv of demand forecast approaches infinity is characterized in the following theorem (proof in appendix).

THEOREM 2: If $q_t = q$ for all t, the demand variable D has distribution function $F(\cdot|m,v)$ satisfying the median-cv limit conditions, and the decision-maker is either risk-neutral, or strictly risk-averse with a utility function such that $\lim_{x\to\infty}U'(x) = 0$ and $\lim_{x\to-\infty}U'(x) = \infty$, then

$$\begin{cases} z^*(\infty) = 0 & \text{if } r \leq 2k, \\ z^*(\infty) = \infty & \text{if } r > 2k \text{ and } U \text{ is risk-neutral}, \\ z^*(\infty) < \infty & \text{if } r > 2k \text{ and } U \text{ is strictly risk-averse}, \end{cases} \quad (12)$$

where $z^*(\infty) := \lim_{v\to\infty} z^*(v)$. When the last case of (12) holds, $z^*(\infty) > 0$ is the unique solution of

$$U'(-kz)/U'((r-k)z) = r/k - 1, \quad (13)$$

and for the special case of CARA preferences with $U(x) = -e^{-\gamma x}$,

$$z^*(\infty) = qm\log(r/k-1)/\beta, \quad (14)$$

where $\beta = \gamma rqm$.

The most noteworthy conclusion from Theorem 2 is the qualitative difference, for high cv values, in the behavior of the FS decision between the risk-neutral and strictly risk-averse cases when the systems are sufficiently profitable (i.e., $r > 2k$). For the risk-neutral case, given the divergence of z^* as $v \to \infty$, the solution is necessarily sensitive to perturbations in the data in the region of interest of high v values. On the other hand, for the strictly risk-averse case, the solution is asymptotically stable in v.

To examine this difference more closely, we make the assumptions of CARA preferences and a lognormal demand distribution with median m and cv v. For the risk-neutral case, substituting (10) into equation (9) and solving for z yields

$$z^* = qm\exp(\Phi^{-1}(1-k/r)\sqrt{\log(v^2+1)}), \quad (15)$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal probability distribution. There are two cases: (1) when $r \leq 2k$, $\Phi^{-1}(1-k/r) \leq 0$ so z^* is decreasing in v, and (2) when $r > 2k$, $\Phi^{-1}(1-k/r) > 0$ so z^* is strictly increasing in v. The latter case implies that when using a risk-neutral model, and to the extent that v represents the confidence in the forecaster's prediction, being conservative (i.e., overestimating v) leads to building a larger facility than being overconfident (i.e., underestimating v).

We would like to go beyond the limiting case and compare the above risk-neutral solution to the one for strictly risk-averse preferences. Unfortunately, in general it is impossible to express the risk-averse solution in closed form. However, we can draw some qualitative conclusions, which we present in the following theorem (proof in appendix).

THEOREM 3: If $q_t = q$ for all t, the demand variable D is lognormal with median m and cv v, the decision-maker's preferences exhibit CARA with risk aversion $-U''(x)/U'(x) = \gamma > 0$, and $\log(r/k-1)/\beta < 1$, where $\beta = \gamma rqm$, then there exists $v_m \in [0,\infty)$ such that $z^*(v)$ is decreasing in v for $v \geq v_m$.

Note that if $r/k \leq 2$, the sufficient condition $\log(r/k-1)/\beta < 1$ is satisfied for any $\beta > 0$. Theorem 3 implies that under the CARA assumption the optimal facility size is eventually decreasing in the cv of the demand forecast for systems with profitability region defined by $0 < \log(r/k-1) < \beta$. This is in sharp contrast with the risk-neutral case, in which the optimal facility size is always increasing in v.

3.2. General Demand Profiles

In the previous subsection, the demand forecast was restricted to flat profiles. In this subsection, we consider the more general case where the demand forecast allows levels to vary from one period to the next. As with the flat profile case, to find the optimal floorspace we first obtain an expression for the expected utility as a function of z from equation (6). We do this by conditioning on the random variable D and computing the discounted net revenue function $R(z|D)$, defined as a solution to optimization problem (**P**). Two characteristics of the function $R(z|D)$ are readily available. First, substituting $z = 0$ into (**P**) collapses the feasible region to a single point and yields $R(0|D) = 0$. Second, the function $R(\cdot|D)$ is concave and piecewise linear. This follows directly from

observing that z is an objective function coefficient in the dual of (**P**); a minimization problem which we denote by (**D**):

$$R(z|D) = \min \quad \sum_{t=1}^{T} \sigma_t q_t D + \mu z$$

$$\text{s.t.} \quad \sum_{n=1}^{N} \pi_{nt} J_n + \sigma_t \geq r_t, \qquad 1 \leq t \leq T \quad (16)$$

$$\sum_{i=t}^{T} \pi_{ni} - \mu a_n \leq c_{nt}, \qquad 1 \leq n \leq N, 1 \leq t \leq T \quad (17)$$

$$\pi_{nt}, \sigma_t, \mu \geq 0, \qquad 1 \leq n \leq N, 1 \leq t \leq T,$$

where μ, σ_t, π_{nt} for $1 \leq n \leq N$, $1 \leq t \leq T$, are the dual variables for problem (**P**).

Hence, fully characterizing $R(\cdot|D)$ requires only identifying the number and location of the breakpoints, and the values of the constant slopes between them. We do this in Lemma 3 (proof in appendix), which is a generalization of Lemma 1. Before proceeding, it is convenient to define b_i to be the period of the ith smallest q_t, i.e., for $1 \leq i \leq T$, $b_i = \arg\min\{q_t : 1 \leq t \leq T \text{ and } t \neq b_j \text{ for all } j < i\}$, and also $q_{b_0} = 0$, $q_{b_{T+1}} = \infty$. This implies $0 = q_{b_0} < q_{b_1} \leq \cdots \leq q_{b_T} < q_{b_{T+1}} = \infty$. Again, without loss of generality we assume $a = 1$. In Section 4 we show how to modify these results for arbitrary a, and hence, arbitrary units.

LEMMA 3: For any demand profile, given D, the discounted net revenue generated by a facility of size z (excluding floorspace costs) is

$$R(z|D) = \sum_{t=1}^{T} \Delta_t q_{b_t} D - \sum_{t=1}^{T} \Delta_t (q_{b_t} D - z)[q_{b_t} D > z], \quad (18)$$

where $\Delta_t = \mu_t - \mu_{t+1}$ for $1 \leq t \leq T$, μ_i is the dual price associated with the floorspace variable z when $z \in (q_{b_{i-1}} D, q_{b_i} D)$, with

$$\mu_i = \max \left\{ \sum_{j=t}^{T} r_j [j = b_i, b_{i+1}, \ldots, b_T] - \sum_{n=1}^{N} J_n c^*_{nt} : 1 \leq t \leq T \right\}^+, \quad 1 \leq i \leq T, \quad (19)$$

$\mu_{T+1} = 0$, and $c^*_{nt} = \min\{c_{nj} : 1 \leq j \leq t\}$. In addition, we use r to denote μ_1.

To calculate the optimal floorspace we proceed as in the flat demand case; assuming the probability distribution F is continuous and using expression (18) together with (6). Since $R(\cdot|D)$ is still concave for general demand profiles, assuming the utility U is concave and twice differentiable with $U' > 0$ guarantees the concavity of $EU[R(z|D) - kz]$. It follows that z^* is well-defined in terms of $\varphi(z) = dEU[R(z|D) - kz]/dz$. If $\varphi(z) < 0$ for all $z > 0$, $z^* = 0$, if $\varphi(z) > 0$ for all $z \geq 0$, $z^* = \infty$; otherwise, z^* is the unique solution to $\varphi(z) = 0$. As we show in the following theorem, $r - k$ is a measure of profitability of the system, and uniquely determines whether it is worthwhile to build the facility or not (i.e., $z^* > 0$ or $z^* = 0$). Note that $r = \mu_1$ in Lemma 3 coincides with r in Lemma 1.

THEOREM 4: Under the same assumptions of Lemma 3, $z^* > 0$ if $r - k > 0$ and $z^* = 0$ otherwise.

PROOF: Since $\varphi(z)$ is decreasing, $z^* > 0$ iff $\varphi(0) > 0$. But $R(0|D) = 0$ implies $\varphi(0) = E[U'(R(0|D) - 0)(R'(0|D) - k)] = U'(0)(r - k)$. Hence $\varphi(0) > 0$ iff $r - k > 0$. □

If $r - k > 0$, calculating the optimal FS solution amounts to solving the first order condition $\varphi(z) = 0$. Its explicit form is the following generalization of (8):

$$\varphi(z) = (\mu_1 - k)U'((\mu_1 - k)z)\overline{F}(z/q_{b_1}) + \sum_{i=1}^{T} (\mu_{i+1} - k) \times \int_{z/q_{b_{i+1}}}^{z/q_{b_i}} U' \left(\sum_{t=1}^{i} \Delta_t q_{b_t} y + (\mu_{i+1} - k)z \right) dF(y) = 0. \quad (20)$$

Given $F(\cdot)$ and $U(\cdot)$, the numerical calculation of z^* can be carried out using standard techniques. Equation (20) indicates that allowing non-flat demand profiles introduces functional complexities to the solutions. As a result, only part of the structural results from the previous subsection can be extended. The following is a generalization of Theorem 1.

THEOREM 5: For a general demand profile and risk-averse preferences that exhibit CARA, if $r - k > 0$ the optimal FS z^* is decreasing in k.

PROOF: For risk-neutral preferences, $\varphi(z) = E[R'(z|D)] - k$, where $R'(z|D) = \mu_i$ if $q_{b_{i-1}} D \leq z < q_{b_i} D$ for $0 \leq i \leq T + 1$, is independent of k. Hence $\varphi(z)$ is decreasing in k. The result follows from $\varphi(z)$ being decreasing in z. For strictly risk-averse CARA preferences, let $U(x) = -e^{-\gamma x}$ with $\gamma > 0$, and $\tilde{\varphi}(z) = \varphi(z)e^{-\gamma kz}/\gamma$. Clearly $\tilde{\varphi}(z) = 0$ iff $\varphi(z) = 0$. Since $\varphi'(z) \leq 0$ for all z, $\tilde{\varphi}'(z^*) = (-k\varphi(z^*) + \varphi'(z^*)/\gamma)e^{-\gamma kz^*} = 0 + \varphi'(z^*)e^{-\gamma kz^*}/\gamma \leq 0$. Hence, it is sufficient to show that $\tilde{\varphi}(z)$ is decreasing in k. But

$\tilde{\varphi}(z) = E[e^{-\gamma R(z|D)}(R'(z|D) - k)]$ for this particular utility, and $\partial\tilde{\varphi}(z)/\partial k = -E[e^{-\gamma R(z|D)}] \leq 0$, which concludes the proof. □

We can also extend our observation on the sensitivity of the flat profile solution to the risk aversion coefficient to the general profile case (proof in appendix).

THEOREM 6: For a general demand profile, as the risk aversion of the decision-maker increases, the optimal FS z^* decreases.

Finally, we can show that for general demand profiles, the behavior of z^* in the limit as v increases to infinity is the same as for flat profiles (proof in appendix).

THEOREM 7: If the flat profile assumption is replaced by the general demand profile, the conclusions of Theorem 2 remain unchanged.

The main idea underlying the above theorem is that as the probability mass is shifted towards the tails, the effect of the region between $q_{b_1}D$ and $q_{b_T}D$ diminishes. It follows that the limiting behavior of the solution, as v grows towards infinity, is independent of the demand profile. In general, the details on how the limit is reached do depend on the profile, so there is no straightforward generalization of Theorem 3. However, a flat demand solution is an asymptotically correct approximation for a general profile system. In Section 4, we exploit this to use the flat profile model as a heuristic for the general profile case.

4. IMPLEMENTATION

We now consider the practical issues involved in using the above framework to compute the optimal FS solution from the data described in Section 2. Henceforth, we assume CARA preferences with $U(x) = -e^{-\gamma x}$, and D to be lognormally distributed according to (10).

To illustrate the solution method, we use the following example representing a microprocessor wafer fab. In constructing this example, we used representative industry data (see e.g., Van Zant [33]), and dataset 4 of the Sematech semiconductor testbed datasets.

The planning horizon consists of $T = 5$ one-year periods, and there are $N = 150$ stations. We assume nominal installation costs of \$3,000,000 per tool. The total yield loss is 50%, but for simplicity we assume it all occurs *after* the last operation, so that $\alpha_p = 1$ for every product $p = 1, \ldots, P$. To calculate revenues, we use an \$85 net revenue per aggregate product chip, and assume 350 die per wafer, resulting in a yearly nominal revenue per throughput of $85 \times 350 \times 12 \times 0.5 =$ \$179,000 per wafer per month (wpm). Table 1 contains the cost and revenue parameters for a 15% discount rate. The c_{nt} are in units of \$ per tool, and the r_t in \$ per wpm.

Table 1. Cost and revenue data for sample problem.

t	1	2	3	4	5
c_{nt}	3,000,000	2,550,000	2,167,000	1,842,000	1,566,000
r_t	179,000	152,000	129,000	110,000	93,000

To keep the example uncomplicated, we assume zero annual labor costs, but these can be easily incorporated into the c_{nt} (e.g., for annual labor costs of \$200,000 per tool across all stations, $c_{n3} = (.85)^2[3{,}000{,}000 + 200{,}000 + (.85)200{,}000 + (.85)^2 200{,}000] = 2{,}539{,}000$).

To compute the profitability r defined in Lemma 1, we must first calculate J_n, the expected number of tools per unit throughput at each station for the maximum utilization u_n defined in (2). Dataset 4 contains an average of 1.97 tools per station, and a release rate of 3,400 wpm results in 95% utilization. Setting maximum utilization u_n to 0.75 across all stations implies $J_n = 1.97 \times (0.95/0.75)/3400 = 7.34 \times 10^{-4}$ per wpm for all n. A direct calculation yields $r =$ \$333,000 per wpm, and $a = \sum_{n=1}^{N} a_n J_n = 0.110\bar{a}$ per wpm, with $\bar{a} = \sum_{n=1}^{N} a_n/N$ representing the mean footprint per tool. The floorspace cost can be specified as $k =$ \$1,000,000$/\bar{a}$, where $k\bar{a}$ is the mean floorspace cost per tool.

Recall that in Section 3 we set $a = 1$ by choosing "normalized" floorspace units. Note that z and a_n have dimensions of floorspace, k of money/floorspace, and a of floorspace×time, and that these are the only elements in the model involving the floorspace dimension. Let $(\tilde{z}, \tilde{a}_n, \tilde{k}, \tilde{a})$ denote the corresponding elements in normalized units. There exists $\xi > 0$ such that $a_n = \xi\tilde{a}_n$ for all n, and consequently $z = \xi\tilde{z}$, $k = \xi^{-1}\tilde{k}$, and $a = \sum_{n=1}^{N} \xi\tilde{a}_n J_n = \xi\tilde{a}$. But by design $\tilde{a} = 1$, so $\xi = a$, and $\tilde{z} = z/a$, $\tilde{k} = ka$. Therefore, to generalize the expressions in Section 3 for any a, it is sufficient to replace every occurrence of z by z/a, and every k by ka. For example, equation (15) for the flat demand and risk neutral case becomes $z^*/\bar{a} = (qma/\bar{a})\exp(\Phi^{-1}(1 - ka/r)\sqrt{\log(v^2+1)})$. Using the example's data, $ka/r = 0.330$, and assuming a median demand of $qm = 3{,}000$ wpm for each period, $qma/\bar{a} = 330$, and $\Phi^{-1}(1 - ka/r) = 0.440$. For $v = 2$ this leads to $z^*/\bar{a} \approx 880$. Note that $z^*/\bar{a}$ is the fab size in total number of average sized tools.

4.1. Risk Attitude Assessment

For the risk-averse case, the solution is determined by first order condition (20) which involves the degree of risk aversion γ. Hence, we must assess the decision-maker's utility function to use the above method. For simplicity we assume

CARA preferences, which is equivalent to restricting the utility function to the simple form $U(x) = -e^{-\gamma x}$. Following Howard [19], the usual simple procedure for assessing the corporate risk tolerance (i.e., the inverse of γ) is to find the amount w such that the senior executives are indifferent as a company investment to a 50-50 chance of winning w and losing $w/2$. The result w is a very close approximation to γ^{-1}.

Assuming the value of γ has been assessed, we now show how to calculate the optimal FS solution, which involves solving first order condition (20). We begin with the special case of flat demand profiles, and present the case of general profiles in a separate subsection.

4.2. Flat Demand Profiles

4.2.1. *Solution*

The case of risk-neutral preferences (i.e., $\gamma = 0$) admits the closed-form solution (15) (with z/a and ka replacing z and k, respectively). If preferences are strictly risk-averse (i.e., $\gamma > 0$), the optimal FS is the solution to first order condition (8), which cannot be expressed in closed form. However, after some algebra and a change of variable, (8) leads to the equivalent condition

$$e^{\beta\zeta}\int_0^{\zeta} e^{-\beta u}\Phi'\left(\frac{\log u}{\eta}\right)\frac{du}{\eta u} + (\rho - 1)\left[\Phi\left(\frac{\log \zeta}{\eta}\right) - 1\right] = 0, \tag{21}$$

where $\Phi(\cdot)$ is the standard normal distribution, and $\zeta = z/aqm$, $\beta = \gamma rqm$, $\rho = r/ka$, and $\eta = \sqrt{\log(v^2+1)}$ are all dimensionless. It follows from this equation that the optimal solution ζ^* depends only upon β, ρ, and v. Figure 1 contains plots of $\zeta^*(v, \beta, \rho)$ as a function of v for several values of β and ρ, generated by solving (21) via standard numerical methods.

The ranges for v and β in the plots of Figure 1 were chosen to include the majority of cases one would expect when solving these types of FS problems, and hence define the practical region of interest. These plots contain the necessary information to calculate an approximation of z^* for any set

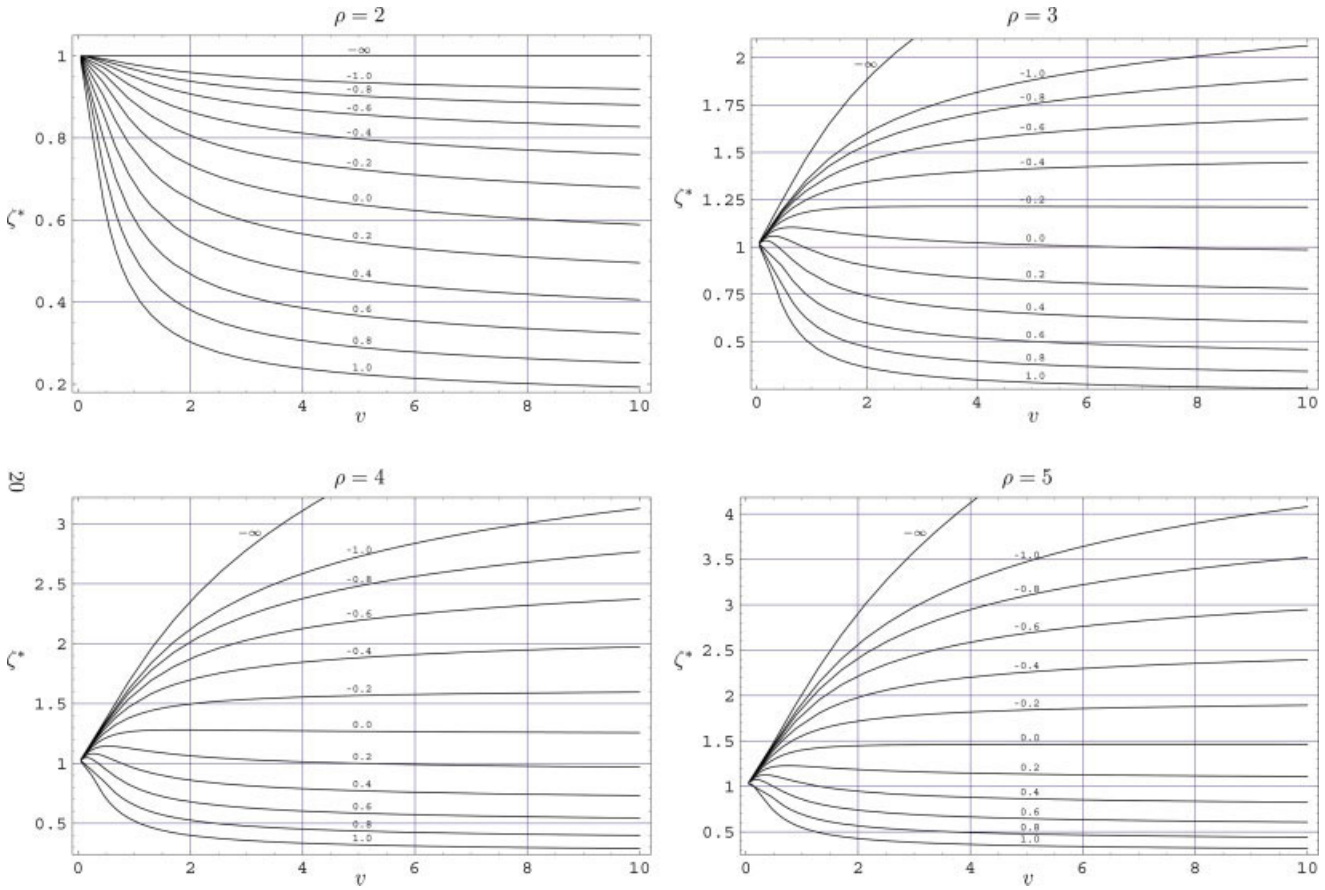

Figure 1. Optimal FS in terms of dimensionless parameters. Curve labels correspond to $\log_{10}\beta$. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

of parameter values in the practical region of interest. We illustrate by calculating the optimal FS for our sample problem. Assuming a risk tolerance of $w = \$500{,}000{,}000$ (see previous section), we get $\gamma = w^{-1} = 2 \times 10^{-9}/\$$. The first step is to calculate $\rho = 333{,}000/110{,}000 = 3.0$ and $\beta = 2 \times 10^{-9} \times 333{,}000 \times 3000 = 2.0$, with $\log_{10} \beta = 0.3$. Using the values for the sample problem yields $qma/\bar{a} = 330$, and $z^*/\bar{a} = 330\zeta^*$. The value of ζ^* for any v in the practical region of interest can be interpolated using the $\log_{10} \beta = 0.2$ and 0.4 curves in the $\rho = 3$ plot from Figure 1. Hence, for $v = 2$, ζ^* has lower and upper bounds of 0.75 and 0.90, and $z^* \approx 270\bar{a}$ (compared with $880\bar{a}$ for the risk-neutral case), where $\bar{a}$ is the average footprint per tool. Evidently, risk aversion level has a substantial impact on the optimal facility size.

4.2.2. Observations

The general nature of the solutions depicted in Fig. 1 allows us to complement the structural analysis of Section 3 with the following numerically-based observations, valid over the practical region of interest: (1) the sufficient condition $\log(r/k - 1) < \beta$ is not necessary for z^* to be eventually decreasing in v, (2) the condition $\log(r/k - 1) < \beta$ is violated only in cases of extremely high profitability and/or extremely low risk-averseness, (3) if the sufficient condition is satisfied (and even if it is slightly violated, i.e., $\log(r/k - 1)/\beta < 2$)), z^* is quite insensitive to changes in v, for $v \geq 2$ ($v \leq 10$), (4) if the sufficient condition is strongly violated, z^* is always increasing in v; representing a continuous transition from the risk-neutral case, and (5) even for moderate risk-averseness and relatively low v, the quantitative discrepancy with the risk-neutral solution is substantial.

4.3. General Demand Profiles

4.3.1. Solution

The first order condition for general demand profiles is (20). For the risk neutral case (i.e., $U'(\cdot) = 1$), and assuming $F(0) = 0$ it reduces to

$$\sum_{t=1}^{T} \Delta_t F(z/q_{b_t})/r = 1 - k/r, \tag{22}$$

which is the generalization of newsvendor first order condition (9). However, in this case, there is no general closed-form solution for $T > 1$. If F is lognormal, after substituting the functional form (10) one must solve (22) numerically.

For the strictly risk-averse case (with CARA preferences and lognormal F), the first order condition (20) is equivalent to

$$\sum_{i=1}^{T} (\rho\mu_{i+1}/r - 1) e^{\beta\zeta(1-\mu_{i+1}/r)} \int_{\zeta q_{b_1}/q_{b_{i+1}}}^{\zeta q_{b_1}/q_{b_i}} e^{-\beta\delta_i u} \Phi'\left(\frac{\log u}{\eta}\right) \times \frac{du}{\eta u} + (\rho - 1)\left[1 - \Phi\left(\frac{\log \zeta}{\eta}\right)\right] = 0, \tag{23}$$

where $\zeta = z/aq_{b_1}m$, $\beta = \gamma r q_{b_1} m$, $\delta_i = \sum_{t=1}^{i} \Delta_t q_{b_t}/r q_{b_1}$, and $\eta = \eta(v)$ and ρ are defined above. The solution ζ^* is uniquely determined by ρ, β, v, μ_i/r, and q_{b_1}/q_{b_i} $(2 \leq i \leq T)$; altogether $2T + 1$ dimensionless parameters. This large number makes it impracticable to generate plots or tables similar to those for flat demand profiles, so equation (23) must be solved numerically on a case-by-case basis.

The floorspace dual prices μ_i can be calculated directly using equation (19). Alternatively, one can solve T versions of the linear program (**D**) with $D = 1$ and fixed z (using floorspace units such that $a = 1$). To calculate μ_i, where $1 \leq i \leq T$, set z to any value in $(q_{b_{i-1}}, q_{b_i})$, e.g., $z = (q_{b_{i-1}} + q_{b_i})/2$, and solve; the resulting μ^* corresponds to μ_i. Table 2 shows a typical demand profile, and the corresponding dual prices (in \$ per wpm) for the sample problem.

4.3.2. Two Approximations

To avoid solving equation (23) numerically, we propose two approximations involving flat demand profiles, with solutions readily available from the plots in Fig. 1. As we pointed out in Section 3, these approximations are asymptotically correct.

In both approximations, all parameters in the flat-demand model retain their original values, and the value of q is set to either

$$q = 1/T, \tag{24}$$

or

$$q = \sum_{t=1}^{T} q_{b_t} \Delta_t / r. \tag{25}$$

Table 2. Demand profile and floorspace dual prices for sample problem.

t	1	2	3	4	5
q_t	2/14	4/14	4/14	3/14	1/14
μ_t	333,000	240,000	114,000	3,550	0

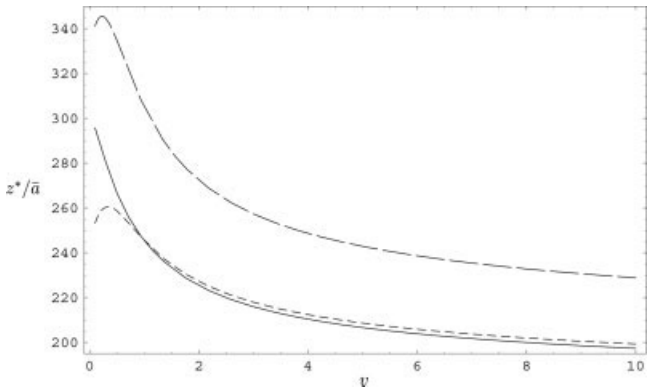

Figure 2. Comparison of approximations (24) (long dashes) and (25) (short dashes) with the exact solution for the sample problem with non-flat profile.

Approximation (24) avoids the calculation of the dual prices μ_t, but proved to be less accurate than approximation (25) in almost all numerical tests. For the sample problem with the profile from Table 2, $q = 0.200$ for (24) and $q = 0.148$ for (25). Figure 2 depicts plots of the exact solution and the two approximations as functions of v (recall that $z^*/\bar{a}$ is the floorspace in total number of average sized tools). For $v = 2$, $z^*/\bar{a} = 225$, approximations (24) and (25) yield respectively 272 ($< 21\%$ relative error) and 227 ($< 1\%$ relative error).

4.3.3. Example Extension

We extend the example from the beginning of this section to illustrate how the demand profile, cost structure and risk coefficient γ affect the FS decision. Table 3 shows a set of six sample demand profiles. Profiles A–C are permutations of each other and thus exhibit different trends and a common dispersion, which is the highest in the set. Profiles D and E exhibit similar dispersions and trends. We include flat profile F for comparison purposes.

In addition to the demand profile, the cost structure is a strong driver of EA decisions. To gauge the cost structure effect on EA and FS decisions, we include the results of a no discounting cost structure with $c_{nt} = 3,000,000$, $r_t = 170,000$ for all n and t. With no penalty for early equipment additions, our model suggests identical EA schedules for demand profiles A–C, with all the equipment added in the first period, which leads to identical FS decisions as illustrated in the lower portion of Fig. 3.

Table 3. Sample demand profiles.

Profile	q_1	q_2	q_3	q_4	q_5
A	4/30	5/30	6/30	7/30	8/30
B	4/30	6/30	8/30	7/30	5/30
C	8/30	7/30	6/30	5/30	4/30
D	14/72	14/72	15/72	15/72	15/72
E	14/72.5	14.25/72.5	14.5/72.5	14.75/72.5	15/72.5
F	1/5	1/5	1/5	1/5	1/5

Under the cost structure from Table 1 in which both equipment cost and revenue parameters depreciate at a typical rate of 15%, our model suggests different EA schedules for profiles A–C. Given the optimal FS z^*, the increasing demand profile A calls for EAs in periods 1–4, provided the realized demand D is moderate (i.e., $D < z^*/q_{b_4}$). The EA for the first period (x_{n1}) increases linearly with D and is of course flat beyond $D > z^*/q_{b_1}$. The EAs for periods 2 to 4 (x_{nt}, $t = 2, 3, 4$) increase linearly with D (all with the same rate, which is lower than that of x_{n1}), peak (at z^*/q_{b_t}) and decrease linearly thereafter since $D > z^*/q_{b_t}$ implies the FS constraint is binding for $b_t, b_{t+1}, \ldots b_T$. Similarly, the increasing-decreasing demand profile B calls for EAs in periods 1–3, x_{n1} increasing in D (with the highest rate) and flat beyond $D > z^*/q_{b_1}$. Both x_{n2} and x_{n3} increase linearly with D (with the intermediate and lowest rates respectively), peak (at z^*/q_{b_3} and z^*/q_{b_4} respectively) and decrease thereafter. The decreasing demand profile C calls for an EA in the first period only. The EA decisions for demand profiles D and E are qualitatively similar to those for profile A, but as a result of the small demand dispersion, any EA beyond the first period is very small relative to the EA for the first period. Using different cost structures produced qualitatively anticipated EA schedules. For instance, when revenues decrease at a faster rate than equipment costs, EAs decrease in general and are omitted in late periods relative to when revenues and costs decrease at the same rate, since it takes more revenue periods to amortize the EA investment. Also, if equipment costs for any period are higher than in an earlier period, the model may call for buying equipment earlier than needed to take advantage of lower early costs.

The upper portion of Fig. 3 illustrates the differences in the optimal FS decision for all profiles and two risk coefficients. Note that the FS is increasing from A to B and from B to C, which reflects the fact that under the given conditions, the effect of the higher net present value in the decreasing demand profile overcomes the value of EA postponement in the increasing demand profile. Figure 3 suggests that the risk coefficient γ can be expected to have the most significant effect on the FS decision, followed by cost structure, demand profile dispersion and demand profile trend. Observe that in this model the benefits from a low dispersion demand profile come from a more efficient use of equipment across periods rather than from reduced demand uncertainty. Given a risk factor and cost structure, Fig. 3 also illustrates the value of using non-flat demand profiles if available. Moreover, the reduction in FS resulting from the use of non-flat demand profiles should be considered conservative because

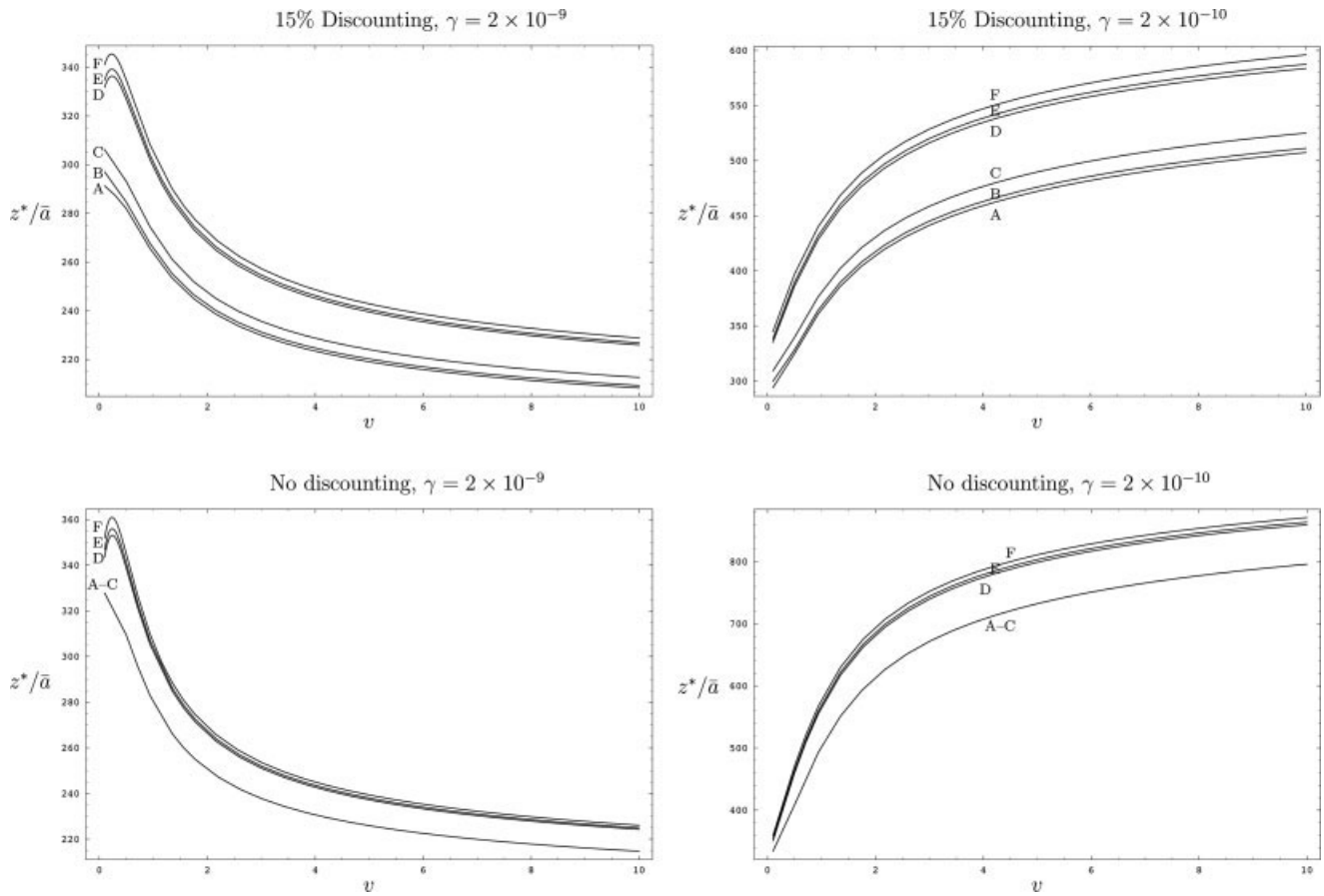

Figure 3. Facility sizing decision under demand profiles in Table 3, two cost structures and two risk coefficients.

in this model the EA decisions represent a best possible case since they follow demand realization. This fact has managerial relevance for decision makers considering adoption of the model.

5. CONCLUSIONS

In this article, we have developed a practical facility sizing tool that can be implemented with obtainable data. This approach is superior to existing capacity planning frameworks because it incorporates risk attitude considerations and explicitly considers the need for designing responsive plants. The solution has several desirable features, including the optimal facility size being eventually decreasing in forecast uncertainty and decreasing in risk aversion, as well as being generally robust to cost errors, forecast uncertainty, and changes in the demand profile—except when the uncertainty is unusually low. The approximations, which can be easily implemented in a spreadsheet, proved to be quite accurate in extensive numerical tests.

Analysis of the optimal solution shows that neglecting risk attitude considerations can result in poor facility sizing decisions that deteriorate with increased forecast uncertainty, which constitutes an important managerial insight.

APPENDIX: PROOFS

PROOF OF LEMMA 1: We construct an explicit solution for problem (**P**) and prove optimality. There are two cases.

Case 1.

Suppose $r > 0$. We want to show that an optimal solution involves capturing demands only during periods t_* through T, where $t_* = \arg\max\{\sum_{i=t}^{T} r_i - \sum_{n=1}^{N} J_n c_{nt}^* : 1 \le t \le T\}$, so that sufficient equipment capacity must be available on or before period t_*. The best installation time for station n is $t(n) = \arg\min\{c_{ni} : 1 \le i \le t_*\}$. Consider the proposed solution: $x_{nt} = \min(z, qD)J_n[t = t(n)]$, and $\theta_t = \min(z, qD)[t \ge t_*]$, for all t and n, which is feasible and has objective value (7), as can be verified by direct substitution into the constraints (3)–(5) and the objective function of problem (**P**). To prove optimality, we show the existence of a feasible dual solution with the same objective value.

The dual of problem (**P**), for the special case of a flat demand profile ($q_t = q$ for all t) is: $\min\{\sum_{t=1}^{T} \sigma_t q D + \mu z : \sum_{n=1}^{N} \pi_{nt} J_n + \sigma_t \geq r_t; \sum_{i=t}^{T} \pi_{ni} - \mu a_n \leq c_{nt}; \pi_{nt}, \sigma_t, \mu \geq 0$; for all $1 \leq t \leq T$ and $0 \leq n \leq N\}$. We can characterize the dual solution in terms of a linear system of equations. Let n_0 be a station index such that $t(n_0) = t_*$ (its existence is a direct consequence of the definition of t_*), and define the index set $X = \{(n,t) : t = t(n)$ for all n, or $1 \leq t < t(n)$ for all $n \neq n_0$, or $t_* < t \leq T$ for all $n \neq n_0\}$. Let the linear system $\Omega = \{\sum_{n=1}^{N} \pi_{nt} J_n + \sigma_t = r_t$ for all $1 \leq t \leq T; \sum_{i=t}^{T} \pi_{ni} - \mu a_n = c_{nt}$ for all $(n,t) \in X; \pi_{nt} = 0$ for all $n \neq n_0$ and $t(n) \leq t < t_*; \sigma_t = 0$ for all $1 \leq t < t_*\}$. For the case $z < qD$, the proposed dual solution is the solution to the system $\Omega \cup \Omega_1$, and for the case $z > qD$, the solution to the system $\Omega \cup \Omega_2$; where $\Omega_1 = \{\sigma_t = 0$ for all $t_* \leq t \leq T\}$, and $\Omega_2 = \{\mu = 0; \sum_{i=t}^{T} \pi_{ni} - \mu a_n = c_{nt}$ for all $t_* < t \leq T$ and $n \neq n_0\}$. The proposed dual solution is well-defined because the rank of both systems is equal to the number of dual variables ($NT + T + 1$). Moreover, manipulating the equations of the systems $\Omega \cup \Omega_1$ and $\Omega \cup \Omega_2$, we show that any solution of either system satisfies the dual constraints, hence, the proposed dual solution is feasible. Finally, combining some of the equations of system Ω yields: $\sum_{t=t_*}^{T} \sigma_t + \mu = \sum_{t=t_*}^{T} r_t + \sum_{n=1}^{N} J_n c_{nt(n)}$. Comparing this expression with the dual objective function $\sum_{t=1}^{T} \sigma_t q D + \mu z$ for the cases $z < qD$ and $z > qD$, and using the equations from Ω_1 and $\mu = 0$ from Ω_2 respectively, we show that the objective value of the proposed dual solution is (7).

Case 2.

Suppose $r = 0$. In this case, the system is not profitable; no stream of revenues can exceed the equipment costs. Therefore, the optimal solution is $x_{nt} = \theta_t = 0$ for all n and t, with objective value zero. □

PROOF OF LEMMA 2: For the lognormal, condition (i) is verified by direct substitution. If $0 < z/m < \infty$ implies $\lim_{v\to\infty} F(z|m,v) = \lim_{\eta\to\infty} \Phi(\log(z/m)/\eta) = \Phi(0) = 1/2$, from which condition (ii) follows. Note that conditions (i) and (ii) imply that in the limit, half of the probability mass shifts to 0^+ and half to infinity. For (iii), observe that $\Phi(\log(z/m)/\eta)$ is decreasing in η when $z > m$. Then, for any v_0, if we define $G(z|m) = F(z|m,v_0)$ for $z > m$, and $G(z|m) = 1/2$ for $z \leq m$, G is integrable and clearly satisfies $F(z|m,v) \leq G(z|m)$ for all $z \in [0,\infty)$ and $v \geq v_0$. The proof for the gamma distribution is analogous and is omitted. □

PROOF OF THEOREM 2: Integrating by parts the first term of equation (8) and simplifying yields

$$\begin{aligned}\varphi(z|m,v) &= k\int_0^{z/q} F(y|m,v)dU'(rqy - kz) \\ &\quad + [r - k - rF(z/q|m,v)]U'((r-k)z) \\ &\quad - kF(0|m,v)U'(-kz) \\ &= k\int_0^{z/q} F(y|m,v)dU'(rqy - kz) \\ &\quad + [r - k - rF(z/q|m,v)]U'((r-k)z),\end{aligned}$$

where the second equality follows from condition (i) in Definition 1. Taking the limit we get

$$\lim_{v\to\infty} \varphi(z|m,v) = U'((r-k)z)(r-k)/2 - U'(-kz)k/2. \qquad (26)$$

When computing the limit of the integral, we exchanged limit and integration; which is valid by virtue of condition (iii) in Definition 1, and Lebesgue's dominated convergence theorem (see e.g., Rudin [24] p. 26).

For the risk-neutral case, substituting $U' = 1$ into the right hand side of (26) yields $r/2 - k$. It follows that the optimal FS solution is $z^*(\infty) = 0$ when $r/2 - k \leq 0$, and $z^*(\infty) = \infty$ otherwise. For the strictly risk-averse case, if $r - 2k \leq 0$, from (26) we get $\lim_{v\to\infty} \varphi(z|m,v) \leq [U'((r-k)z) - U'(-kz)]k/2 < 0$ for all $z > 0$ (the second inequality due to U' being strict decreasing), and hence, $z^*(\infty) = 0$. If $r - 2k > 0$, expression (26) as a function of z has a zero iff equation (13) has a solution. But since U' is strictly decreasing in its argument, for all $z \geq 0$, the numerator of the left hand side of (13) is increasing in z, and the denominator is decreasing in z, and hence, the quotient is strictly increasing in z. In addition, evaluating the quotient at zero yields $U'(0)/U'(0) = 1$, and using the limit values of U', $\lim_{z\to\infty} U'(-kz)/U'((r-k)z) = \infty$. Therefore, as $1 < r/k - 1 < \infty$, equation (13) has a unique solution $z^*(\infty) \in (0,\infty)$. Finally, substituting $U(x) = -e^{-\gamma x}$ into equation (13) yields (14). □

PROOF OF THEOREM 3: If $r - k \leq 0$, $z^* = 0$ for all $v \geq 0$. Hence, assume $r - k > 0$. The proof has three steps: (1) we derive an equivalent first order condition $\tilde{\varphi}(z|v) = 0$, and show that $\tilde{\varphi}(z|v)$ is decreasing in z, (2) we show that if there exists a v_m such that $z^*(v_m) \leq qm$, then $z^*(v)$ is decreasing in v for $v \geq v_m$, and (3) we prove the existence of v_m.

(1) Let $\tilde{\varphi}(z|v) = \varphi(z|v)e^{-\gamma kz}/\gamma$. Then $\tilde{\varphi}(z|v) = 0$ iff $\varphi(z|v) = 0$. Using this definition and substituting $U(x) = -e^{-\gamma x}$ into (8) yields $\tilde{\varphi}(z|v) = -k\int_0^{z/q} e^{-\gamma rqy} dF(y|m,v) + (r-k)e^{-\gamma rz}\overline{F}(z/q|m,v)$, which is decreasing in z. (2) Integrating by parts and simplifying we can write $\tilde{\varphi}(z|v) = -\gamma rqk\int_0^{z/q} e^{-\gamma rqy}F(y|m,v)dy - e^{-\gamma rz}[rF(z/q|m,v) - r + k]$. For any $z \leq qm$, the argument of F in both terms is $\leq m$, and hence, from definition (10), F is increasing in v. It follows that $\tilde{\varphi}(z|v)$ is decreasing in v for $z \leq qm$. Therefore, if $z^*(v_m) \leq qm$ then $\tilde{\varphi}(z^*(v_m)|v) \leq \tilde{\varphi}(z^*(v_m)|v_m) = 0$ for any $v \geq v_m$, which together with step (1) and the definition of $z^*(v)$ implies $z^*(v) \leq z^*(v_m)$. As $z^*(v) \leq qm$, given any $v' \geq v$, the result $z^*(v') \leq z^*(v)$ follows directly from the same reasoning. (3) Combining limiting result (14) with the condition $\log(r/k - 1)/\beta < 1$ implies that $\lim_{v\to\infty} z^*(v) < qm$, and hence, there exists a sufficiently large v_m such that $z^*(v_m) \leq qm$. This concludes the proof. □

PROOF OF LEMMA 3: The main part of the proof is computing the dual price of z. Before proceeding, note that any optimal solution to (**P**), denoted by $\{\theta_t^*, x_{nt}^*\}$, satisfies

$$\max_{1\leq t\leq T} \theta_t^* = \sum_{i=1}^{T} x_{ni}^*/J_n, \quad 1 \leq n \leq N. \qquad (27)$$

To show this observe that because of the utilization constraints (3), if (27) is false, there exists some station n for which the inequality in (3) is strict. In this case, a new solution can be constructed by modifying the x_{nt}^* variables according to $\sum_{i=1}^{T} x_{ni}^* \leftarrow \min(\max_t \theta_t^* J_n, \sum_{i=1}^{T} x_{ni}^*)$. But by direct substitution, it follows that this modified solution is feasible and has a higher objective value, which contradicts the optimality of the original solution, and thus proves (27). Notice that (27) implies $\sum_{t=1}^{T} x_{nt}^*/J_n = \sum_{t=1}^{T} x_{\hat{n}t}^*/J_{\hat{n}}$, for any stations n and $\hat{n}$. Substituting into (5) yields the tighter floorspace constraints

$$\sum_{t=1}^{T} x_{nt} \leq zJ_n, \quad 1 \leq n \leq N. \qquad (28)$$

Having shown these two results, we are now ready to compute the dual price of z. We begin with the case $z > q_{b_T} D$. Combining (27) with the demand constraints (4) leads to $\sum_{i=1}^{T}\sum_{n=1}^{N} x_{ni}^* \leq q_{b_T} D$. This means that the floorspace constraint (5) is not tight, and hence, $\mu_{T+1} = 0$ by complementary slackness. Observe that in this case, the optimal net revenue is not affected by floorspace and is only limited by profitability. Note also that even without the floorspace constraint (5), given D, the finite forecast always guarantees a bounded solution.

Turning to the more general case, let $z \in (q_{b_{i-1}} D, q_{b_i} D)$, with $1 \le i \le T$. According to (28), a unit increase in z allows an increase of J_n units in the total equipment capacity at station n. To evaluate the net benefit of such an increase in capacity, note that because of the utilization constraint (3), to produce an additional unit of throughput during the periods t through T, J_n units of equipment must be added at each station n, on or before period t. For station n, this represents a cost increase of $c^*_{nt} J_n$, where $c^*_{nt} = \min\{c_{nj} : 1 \le j \le t\}$ is the incremental cost per unit of capacity at station n for use during period t and beyond. This reflects the possibility of obtaining lower costs by installing the equipment before it is required. The total incremental cost is obtained by adding costs for all stations, i.e., $\sum_{n=1}^{N} c^*_{nt} J_n$. Although this additional capacity allows to increase production during periods t through T, additional revenues are only accrued for the periods where more demand can be captured by increasing z; namely $b_i, b_{i+1}, \ldots, b_T$. The additional revenue for this increase in equipment capacity available during period t is then $\sum_{j=t}^{T} r_j [j = b_i, b_{i+1}, \ldots, b_T]$, and the net revenue is obtained by selecting the most convenient period of availability, i.e., $\max\{\sum_{j=t}^{T} r_j [j = b_i, b_{i+1}, \ldots, b_T] - \sum_{n=1}^{N} J_n c^*_{nt} : 1 \le t \le T\}$. If this quantity is positive, it corresponds to μ_i; the additional profits resulting from a unit increase in z. On the other hand, if the quantity is negative, the original optimal net revenue was limited by profitability, and an increase in floorspace is not useful, hence $\mu_i = 0$. Equation (19) follows from combining the two cases, and since the reasoning only used the assumption $z \in (q_{b_{i-1}} D, q_{b_i} D)$, the expression for μ_i is valid for the whole interval. This completes the determination of all the breakpoints and slopes.

Finally, using $R(0|D) = 0$ and the continuity at the breakpoints yields $R(z|D) = \mu_i z + \sum_{j=1}^{i-1} (\mu_j - \mu_{j+1}) q_{b_j} D$ for $z \in (q_{b_{i-1}} D, q_{b_i} D)$, where $1 \le i \le T + 1$. This expression can be cast into (18) after some algebra. □

PROOF OF THEOREM 6: If $r - k \le 0$, $z^* = 0$ for any risk-averse utility, hence, we assume $r - k > 0$. An increase in risk is equivalent to a concave transformation $V(\cdot)$ of the risk-averse utility $U(\cdot)$, where $V' > 0$ and $V'' \le 0$. Let $\hat{\varphi}(z) = dE[V \circ U(R(z|D) - kz)]/dz$. Since $\hat{\varphi}(\cdot)$ is decreasing, it is sufficient to show that $\hat{\varphi}(z^*) \le 0$, where z^* solves $\varphi(z) = 0$. Towards this end, we first show that $R(z|\cdot)$ is increasing for every z. Let $D' > D$, then for every $z \ge 0$ there exist $j \le i$ such that $z \in [q_{b_{i-1}} D, q_{b_i} D) \cap [q_{b_{j-1}} D', q_{b_j} D')$. From (18), $R(z|D) = \mu_i z + \sum_{t=1}^{i-1} (\mu_t - \mu_{t+1}) q_{b_t} D$, but $q_{b_{i-1}} D < z$ implies $\sum_{t=j}^{i-1} (\mu_t - \mu_{t+1}) q_{b_t} D \le (\mu_j - \mu_i) z$, which leads to $R(z|D) \le R(z|D')$. Recall that the μ_t are decreasing in t, with $\mu_1 - k = r - k > 0$ and $\mu_{T+1} - k = -k < 0$, therefore, there exists some $1 \le \ell \le T$ such that $\mu_\ell - k > 0$ and $\mu_{\ell+1} - k \le 0$. Partitioning into the different intervals where $R'(z|D)$ is constant,

$$\hat{\varphi}(z) = \sum_{i=1}^{T+1} (\mu_i - k) E[V'(U(R(z|D) - kz)) \times U'(R(z|D) - kz); q_{b_{i-1}} D \le z < q_{b_i} D].$$

For $1 \le i \le \ell$, $\mu_i - k > 0$, and we use $E[V'(U(R(z|D) - kz)) U'(R(z|D) - kz); q_{b_{i-1}} D \le z < q_{b_i} D] \le V'(U(R(z|z/q_{b_\ell}) - kz)) E[U'(R(z|D) - kz); q_{b_{i-1}} D \le z < q_{b_i} D]$, where the inequality follows from $R(z|\cdot)$ and $U(\cdot)$ being increasing and $V'(\cdot)$ decreasing. On the other hand, for $\ell + 1 \le i \le T$, $\mu_i - k \le 0$, and we use $E[V'(U(R(z|D) - kz)) U'(R(z|D) - kz); q_{b_{i-1}} D \le z < q_{b_i} D] \ge V'(U(R(z|z/q_{b_\ell}) - kz)) E[U'(R(z|D) - kz); q_{b_{i-1}} D \le z < q_{b_i} D]$. Combining the two sets of inequalities yields $\hat{\varphi}(z) \le V'(U(R(z|z/q_{b_\ell})))\varphi(z)$. It follows that $\hat{\varphi}(z^*) \le 0$, and hence $\hat{z}^* \le z^*$; where $\hat{z}^*$ solves $\hat{\varphi}(z) = 0$, so the proof is complete. □

PROOF OF THEOREM 7: Integrating by parts the last T terms of (20) and simplifying leads to

$$\begin{aligned}\varphi(z|m,v) &= (\mu_1 - k) U'((\mu_1 - k)z) \\ &\quad - \sum_{i=1}^{T} \Delta_i F(z/q_{b_i}|m,v) U' \left(\sum_{t=1}^{i-1} \Delta_t q_{b_t} z/q_{b_i} + (\mu_i - k)z \right) \\ &\quad + kF(0)U'(-kz) - \sum_{i=1}^{T} (\mu_{i+1} - k) \\ &\quad \times \int_{z/q_{b_{i+1}}}^{z/q_{b_i}} F(y|m,v) dU' \left(\sum_{t=1}^{i} \Delta_t q_{b_t} y + (\mu_{i+1} - k)z \right).\end{aligned}$$

To calculate the limit as $v \to \infty$, we exchange the limit and integration in the last T terms of this expression, and after some algebra, we get equation (26). The result follows from the remainder of the proof of Theorem 2. □

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the University of Chicago Graduate School of Business, The Olin Business School, and the National Science Foundation under grants DMI-9322830, DMI-9732868, and DMI-0114598.

REFERENCES

[1] M. Abramowitz and I.A. Stegun, Handbook of mathematical functions, Dover, New York, 1972.
[2] V. Agrawal and S. Seshadri, Impact of uncertainty and risk aversion on price and order quantity in the newsvendor problem, Manufact Serv Oper Management 2 (2000), 410–423.
[3] A. Angelus, E.L. Porteus, and S.C. Wood, Optimal sizing and timing of capacity expansions with implications for modular semiconductor wafer fabs. Working Paper (1997), Graduate School of Business, Stanford University.
[4] M. Anvari, Optimality criteria and risk in inventory models: The case of the newsboy problem, J Oper Res Soc 38 (1987), 625–632.
[5] M. Anvari and M. Kusy, Risk in inventory models: Review and implementation, Eng Costs Product Econ 19 (1990), 267–272.
[6] J. Bard, J.K. Srinivasan, and D. Tirupati, An optimization approach to capacity expansion in semiconductor manufacturing facilities, Int J Product Res 15 (1999), 3359–3382.
[7] D.L. Benavides, J.R. Duley, and B.E. Johnson, As good as it gets: Optimal fab design and deployment, IEEE Trans Semiconduct Manufact 12 (1999), 281–287.
[8] S. Benjaafar, Modeling and analysis of congestion in the design of facility layouts, Management Sci 48 (2002), 679–704.
[9] J.R. Birge, Option methods for incorporating risk into linear capacity planning models, Manufact Serv Oper Management 21 (2000), 19–31.
[10] G.R. Bitran and D. Tirupati, Capacity planning in manufacturing networks with discrete options, Ann Oper Res 17 (1989), 119–135.
[11] M. Cakanyildirim, R.O. Roundy, and S.C. Wood, Optimal machine capacity expansions with nested limitations under stochastic demand, Naval Res Logist 51 (2004), 217–241.

[12] F. Chen and A. Federgruen, Mean-variance analysis of basic inventory models, Working Paper (2000) Graduate School of Business, Columbia University, New York.

[13] K.H. Chung, Risk in inventory models: The case of the newsboy problem—Optimality conditions, J Oper Res Soc 41 (1990), 173–176.

[14] Q. Ding, L. Dong, and P. Kouvelis, On the integration of production and financial hedging decisions in global markets, Oper Res 55 (2007), 470–489.

[15] L. Eeckhoudt, C. Gollier, and H. Schlesinger, The risk-averse (and prudent) newsboy, Management Sci 41 (1995), 786–794.

[16] X. Gan, S.P. Sethi, and H. Yan, Coordination of supply chains with risk-averse agents, Product Oper Management 132 (2004), 135–149.

[17] V. Gaur and S. Seshadri, Hedging inventory risk through market instruments, Manufact Serv Oper Management 7 (2005), 103–120.

[18] W.J. Hopp, M.L. Spearman, S. Chayet, K.L. Donohue, and E.S. Gel, Using and optimized queueing network model to support wafer fab design, IIE Trans 34 (2002), 119–130.

[19] R.A. Howard, Decision analysis: Practice and promise, Management Sci 34 (1988), 679–695.

[20] R.L. Keeney and H. Raiffa, Decisions with multiple objectives: Preferences and value tradeoffs, Wiley, New York, 1976.

[21] H.S. Lau, The newsboy problem under alternative optimization objectives, J Oper Res Soc 31 (1980), 525–535.

[22] J. Li, H. Lau, and A. Lau, Some analytical results for a two-product newsboy problem, Decision Sci 21 (1990), 710–726.

[23] S. Rajagopalan and S. Yu, A capacity planning model with congestion costs, Eur J Oper Res 134 (2001), 137–149.

[24] W. Rudin, Real and complex analysis, McGraw-Hill, New York, 1987.

[25] S.M. Ryan, Capacity expansion for random exponential demand growth with lead times, Management Sci 50 (2004), 740–748.

[26] E. Sankarasubramanian and S. Kumaraswamy, Note on "optimal ordering quantity to realize a pre-determined level of profit," Management Sci 29 (1983), 512–514.

[27] R. Suri, G.W. Diehl, S. de Treville, and M. Tomsicek, From CAN-Q to MPX: Evolution of queueing software for manufacturing, Interfaces 25 (1995), 128–150.

[28] J.M. Swaminathan, Tool procurement planning for wafer fabrication facilities: A scenario-based approach, IIE Trans 34 (2002), 145–155.

[29] M.R. Walls and J.S. Dyer, Risk propensity and firm performance: A study of the petroleum exploration industry, Management Sci 42 (1996), 1004–1021.

[30] S.D. Wu, M. Erkoc, and S. Karabuk, Managing capacity in the high-tech industry: A review of literature, Eng Econ 50 (2005), 125–158.

[31] J.A. Van Mieghem, Capacity management, investment, and hedging: Review and recent developments, Manufact Serv Oper Management 5 (2003), 269–302.

[32] J.A. Van Mieghem, Risk mitigation in newsvendor networks: Resource diversification, flexibility, sharing and hedging, Working paper (2004), Kellogg School of Management, Northwestern University, Evanston, IL.

[33] P. Van Zant, Microchip fabrication: A practical guide to semiconductor processing, 3rd ed., McGraw-Hill, New York, 1997.