ELSEVIER

Journal of Operations Management 12 (1995) 273–296

JOURNAL OF OPERATIONS MANAGEMENT

Plant objectives as revealed by shop floor activities

Robert A. Miller *

Graduate School of Industrial Administration, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA

Prasad Ramnath

Parsifal Systems, Pittsburgh, PA, USA

Abstract

One problem management scientists face in adapting heuristics to actual applications on the factory floor is eliciting preferences from plant managers about the importance of different jobs. Even when the goal of the firm is apparently as straightforward as profit maximization, reputation and goodwill can play such critical roles in affecting the revenue flow, that exclusively focusing on a few key variables (such as the price of individual items, their production times, and the raw material cost) may give a distorted picture about which jobs the firm values most highly. This paper offers a new way of eliciting preferences, utilizing high-frequency data on plant operations that are routinely collected by many firms, in order to infer the direct and indirect cost of scheduling jobs, from actual job schedules that managers reveal by their choices. We then apply our method to a scheduling problem in a steel tube manufacturing plant. After estimating the preferences of the plant manager, we demonstrate how our estimates can be used to evaluate heuristics for hard scheduling problems, and to forecast the effects of structural change, such as expansion in plant capacity, or shifts in job order flow.

1. Introduction

Which new orders fail to materialize because current jobs are tardy or defective, the willingness of the plant to hold finished-goods inventory rather than insist on immediate delivery, even the reputation effect from continuing a product line versus withdrawing it, are three examples of how shop floor activities help determine the plant's performance. It is important to know exactly how the goals and objectives of a company are implemented on the shop floor through its scheduling and capital acquisition decisions. If actions taken at the plant level reflect priorities that are not aligned with each other or, alternatively, are at variance with the espoused goals of management, this could reveal conflict or inefficiency within the organization. Also, unless management consultants know the plant's objectives, useful advice about which decision rule to adopt is hard to provide. Similarly, anticipating the effects of proposals to change the shop floor typically requires the forecasting agency to know what priorities will determine how the new work environment will function. Let us briefly consider these factors in turn.

Suppose the priorities revealed by decisions made at the plant level do not mirror stated managerial

* Corresponding author. Tel.: 412-268-3701.

SSDI 0272-6963(95)00003-8

objectives. This could signal a communication problem between upper level management and operatives, or incentives that do not align rewards and penalties for shop-floor employees with company goals. Similarly, when actions in one part of the plant cannot be reconciled, through the same set of objectives, with those taken elsewhere, there is scope to improving coordination and resolving conflicts thus enhancing overall firm performance. Identifying these sources of inefficiency is a catalyst for improving the overall performance of the plant.

Another reason why decisions taken on the plant floor might sometimes appear at odds with other ways the managerial goals are revealed is the sheer difficulty of determining how to best implement commonly shared objectives. In this case an outside consultant might be called to advise the plant on which decision rule to adopt. One way or another, the consultant must elicit the preferences managers exhibit over the trade-offs that limited resources impose on the plant, before he can provide useful input to his client.

If, on the other hand, managers are confident that the company is efficiently organized, maximizing performance subject to the constraints imposed by the technology and markets (both of which may be fashioned by the firm), then detailed information about how plant floor activities affect company goals is, potentially, a very useful resource for forecasting the effects of innovations (such as the purchase of new equipment, or changing the product mix) on the shop floor.

Yet, precisely how such detailed knowledge of these very tangible alternatives translate to even a simply stated company goal like value maximization, is difficult to elicit from management. Unlike academics, their goals are chiefly concerned with the specifics of the shop floor under their responsibility (rather than searching for unifying principles common to all shop floors), and have only a tangential interest in articulating, or even understanding, their own plant in broader terms. Therefore, compared to academia, the culture of the workplace is less introspective, placing more emphasis on direct measures of plant performance.

Decision theorists have confronted the task of eliciting preferences with mixed success; in interviews, managers are questioned about hypothetical situations in order to determine their preferences. [1] This approach is an expensive form of information retrieval, because the manager must learn decision theory well enough to articulate useful responses. Indeed, it might be argued that the degree of introspection required to answer such questions makes this approach valuable as a teaching tool for imparting powerful concepts, but not as a practical guide for designing superior heuristics. To put this argument another way, shop floor operatives say that much learning takes place through experience on the job; it is hard to fathom why such a costly form of education would be frequently resorted to, if an analytic approach, such as that taught in classrooms, provided an effective low-cost alternative.

Protocol analysis circumvents the problem by encoding the manager's actions directly. [2] Consequently the resulting heuristic implicitly incorporates the manager's preferences and the constraints imposed by the production technology, as reflected in his responses to the management scientist who is conducting the analysis. However, knowledge of the coded decision rule by itself cannot be used to diagnose communication, incentive and complexity problems; nor are such coded rules useful for forecasting the effects of structural changes the plant managers might contemplate. To undertake such diagnosis and forecasting, we need to know how the coded rules would change in response to new incentive plans, better computational algorithms and structural shifts on the shop floor. For these reasons we believe that protocol analysis is more suited for replacing (expensive) human resources with (inexpensive) software that makes almost identical decisions, rather than improving the quality of those decisions.

This paper develops and implements a new approach to elicit managerial preferences based on their past decisions concerning the scheduling of jobs on a day-to-day basis, as found in plant records. One advantage of our procedure is that it expends

[1] See (Keeney and Raiffa, 1976) for a definitive treatment of this approach or (Zeleny, 1976; Hogarth, 1987) for additional work in this area.

[2] See, for example, (Newell and Simon, 1972) or (Ericsson and Simon, 1984).

less managerial time than the other two approaches: rather than require the manager to articulate himself, demanding that he recast his decisions from the perspective of the management scientist, our approach gleans information directly from what the manager is paid to do, namely, from production decisions that he makes on the plant floor.

We formulate the preferences of the manager as an objective function that is known up to a vector of parameters to be estimated. This objective function could encompass any set of characteristics of the jobs, customers, or the condition of the shop floor. We assume the plant manager makes scheduling decisions that are determined by these preferences, that some of those decisions are easy to make optimally, and that we have access to high frequency data which records the outcomes of those optimally made (scheduling) decisions. The vector of unknown parameters is estimated by matching features of the decisions that the manager makes with analogues of those generated by the optimal rule.

This approach has been widely used in economics to estimate structural dynamic models of sequential decision making under uncertainty with sample populations of panel data. The first applications by Miller (1984), Wolpin (1984), Pakes (1986) and Rust (1987) have spawned a large and still growing literature on estimating models of dynamic discrete choice. [3] Following their precedent, formulating a dynamic programming model and numerically solving the optimal decision rule for many different parameter values has become the standard approach to estimating models of dynamic discrete choice, despite the onerous computational burdens entailed. Recently Hotz and Miller (1993) have developed a semiparametric estimator for dynamic programming problems that avoid the substantial computational costs of solving them. [4]

Here we take a slightly different tack. Although the optimal decision rule for our scheduling application is NP hard, we demonstrate that part of it, scheduling tardy jobs, is quite straightforward, using a forward index rule that is easy to calculate. This index rule makes pairwise comparisons between jobs, to determine the schedule for tardy jobs. Rather than assuming the plant manager can solve the whole scheduling problem, we merely assume that, since tardy jobs are easy to schedule, he solves this part of the scheduling problem correctly. This far weaker, identifying assumption enables us to infer some features of the plant's objectives from his decisions about the set of tardy jobs alone.

The next section models a generic factory floor and the class of preferences considered here. Section 3 lays out the procedure used to identify and estimate managerial preferences, and establishes the properties of those estimates. We modify the simulation techniques developed by McFadden (1989) and Pakes and Pollard (1989) for independent processes to take account of finite time dependence that one typically encounters in manufacturing environments, appealing to work by Andrews (1991) to handle serial correlation in our nonlinear framework. In the fourth section we describe the factory floor of our empirical application, and then report our estimates of the plant's objectives. The fifth section puts our empirical results to work. First we conduct counterfactual experiments using different heuristics previously proposed in the scheduling literature (after some adaptation), comparing them with each other on the basis of the estimated objective function obtained in Section 4. Then we demonstrate how to predict the effects, on output measures and the firm's objectives, of changing the shop floor capacity if a given rule is used. The last section discusses the managerial implications more broadly, and mentions some ideas for future research. [5]

2. A framework

The framework for our analysis can be described as follows. A plant manufactures products of differ-

[3] See (Eckstein and Wolpin, 1989) and (Rust, 1994) for surveys of this fast growing field.

[4] Applications of the Conditional Choice Probability Estimator, as it is known, appear in (Hotz and Miller, 1993; Hotz et al., 1994; Sanders, 1994; Aguirregabiria, 1994; Slade, 1994).

[5] A companion paper (Miller and Ramnath, 1994a), also analyzes situations where the computational capabilities of the researcher may be dominated by the specific knowledge of the practitioner, and explains how, even in this situation, structural econometrics might help managers who confront hard scheduling problems.

ent types, on a made to order basis. Let w_t denote the inflow of new orders at the beginning of day t. Foreshadowing our empirical study, w_t represents the jobs welded upstream on the previous day that arrive at the current production stage, but more generally this could denote work done on orders at various upstream stages or at other plants before arriving at the current production stage as raw material input. Denote by x_{tj} a generic order waiting to be processed at the beginning of date t, and by $x_t = \sum_{j=1}^{J_t} x_{tj}$ the vector sum of outstanding orders, where there are J_t outstanding orders on date t. Let $d_{tj} \in [0, 1]$ denote the fraction of x_{tj} processed on date t. For the purpose of this analysis, we assume that w_t and x_t are elements of a vector space X. The vector sum of orders processed on date t is given by

$$\sum_{j=1}^{J_t} d_{tj} x_{tj}. \tag{2.1}$$

Each day x_t is updated by the incoming new orders w_t which add to the stock and job completions $\{d_{tj}\}$ which deplete it. Unprocessed orders change their characteristics only by becoming one day later; none of the other characteristics change. Denoting $x_{tj} = (x_{tj0}, x_{tj1}, \ldots, x_{tjK}, x_{tj,K+1})$, where each argument represents one of the characteristics of x_{tj}, let x_{tj0} denote its lateness in days, that is, the current date t less its actual due date. Thus, each day a job is left unprocessed, it becomes late by one more day, denoted by $f: X \to X$, where

$$f(x_{tj}) = (x_{tj0} + 1,\ x_{tj1}, \ldots, x_{tjK}, x_{tj,K+1}) \tag{2.2}$$

formally denotes how each job progresses from day to the next (if unprocessed). Based on the above:

$$x_{t+1} = w_t + \sum_{j=1}^{J_t} f(x_{tj} - d_{tj} x_{tj}). \tag{2.3}$$

The production stage can in general be a job shop consisting of multiple machines and multiple routes. When jobs are processed at this stage, they consume a certain percentage of the total available capacity. We abstract this production stage as a single aggregate machine, where the processing time of each job on this machine, is equivalent to the fraction of the daily available capacity at the production stage. Besides parsimony, this assumption reflects some practical limitations we encountered; the data we analyze subsequently in Section 4 does not reveal exactly what happens to orders within the production stage, just information on when they enter and when they leave the shop floor.

The asymptotic properties of the econometric procedures used in the empirical portion of this paper also rely on the assumption that w_t is an exogenous, stationary markov process. Note that the markov assumption is not a substantive restriction, providing one includes lagged variables within the state space, appropriately models seasonal effects, and so forth. Stationarity is harder to defend, especially given the possibility that the 3 year period of daily data we exploit in our application may be too short to properly accommodate anything but high frequency events. In practice, model builders are faced with a trade off between shorter data sets like ours, chosen partly because there was limited structural change occurring on the plant floor during this period, versus data sets covering a longer time period. Although the latter deal with low frequency stationary processes, they may be more susceptible to nonstationary structural shifts such as technological change.

Based on the assumptions above about the daily effective capacity at the production stage, the total amount of time devoted to processing jobs each day, is a random variable that depends on machine breakdowns and other unforseen stoppages. We normalize the mean amount of time available each day to 1 unit, and denote by ξ_t the amount of time lost on day t due to stoppages, as a proportion of the mean time available. Thus $E[\xi_t] = 0$ and negative values indicate days when effective capacity was greater than usual. The mapping $g: X \to \mathbb{R}^+$ determines how much time a job x_{tj} takes to process. We further impose the restriction that the processing time of a job is independent of its tardiness:

$$g(f(x_{tj})) = g(x_{tj}). \tag{2.4}$$

Processing time is assumed additive in jobs, which implies that the daily capacity constraint can be stated as

$$\sum_{j=1}^{J_t} d_{tj} g(x_{tj}) \leqslant 1 - \xi_t. \tag{2.5}$$

The firm is assumed to have preferences over processing outstanding orders x_{tj}, which we formally express as a cost function associated with the respective job completion times. Let $c(x_{tj})$ denote the cost of processing x_{tj} on date t. To make headway on partially characterizing the optimal rule, we impose three further assumptions on preferences (which for the most part were viewed by the managers of the plant we studied as being reasonable approximations). First, it is assumed that the cost is additive in the jobs that comprise x_t. Thus, the reputational loss and financial penalties incurred from delivering one job late are independent of the plant's performance on other jobs. Accordingly let

$$\lim_{T\to\infty}\left[\frac{1}{T}\sum_{t=0}^{T}\sum_{j=1}^{J_t} d_{tj}c(x_{tj})\right] \tag{2.6}$$

denote the average cost of pursuing a job-processing policy of $\{\{d_{tj}\}_{j=1}^{J_t}\}_{t=0}^{\infty}$ when the order flow is $\{w_t\}_{t=0}^{\infty}$, and it is understood that jobs left unprocessed on the shop floor are treated the same way as those completed at date infinity. [6] The manager's goal is to choose a policy $\{\{d_{tj}\}_{j=1}^{J_t}\}_{t=0}^{\infty}$ to minimize (2.6) subject to (2.5).

Since costs are additively separable across jobs, the tardiness of one job does not affect the costs associated with making another job late, or the probability that another order might arrive. To investigate the importance of the first factor, one would like to know whether the timing of various orders is coordinated by their respective buyers, so that making one job late would reduce the value of another order. It is hard to believe that orders placed by different firms affect each other directly, since the completed jobs are to be used in products of different firms. More surprisingly, management informed us that even in the case of two orders by the same firm, there seemed to be little coordination on the part of the buyers from different parts of the same client firm. The assumption that firm performance does not affect future orders (but only prices of future sales) is both a convenient simplification and an important limitation of this analysis. The limitation, that can be leveled at almost all the academic literature on scheduling, stems from the observation that if the plant manager changes his scheduling policy, client firms will ultimately adopt new expectations about how new orders will be treated, and this might affect their future demand. [7] This important point is taken up again in the conclusion: we regard policy recommendations coming from analyses like this one as useful but, nevertheless, partial.

The second assumption is that the manager's objective function (2.6) minimizes average costs per period over an infinite horizon. Our formulation differs from standard economic frameworks, which typically discount costs to reflect the importance of the present over the future. The optimal decision rule for this problem is, in fact, the limit of the optimal rule for the discounted case as the interest rate facing the firm converges to zero. The average loss criterion can be defended along empirical lines: since long delays are rare in the data we study, the effects of incorporating discounting into the empirical analysis are probably negligible. Similarly, assuming an infinite horizon is innocuous. This specification circumvents the end point, or terminal period, phenomena of a finite horizon problem, which would affect both the optimal scheduling rule (which would no longer be stationary), and hence the assumptions exploited in estimation. Lacking detailed information about what might happen to the plant in the long run, investigating the infinite horizon case is a reasonable benchmark.

The third assumption is about the functional form of $c(x_{tj})$. We assume that if a job is processed more than ρ days before its due date, the current cost of processing the job for a unit of time is just τ, a constant. Such jobs are said to be processed early. Likewise, jobs processed less than ρ days before their due dates are considered tardy. [8] We also place

[6] This condition is necessary to remove the otherwise trivial policy of never scheduling any jobs from being optimal.

[7] In other words, the scheduling literature implicitly assumes that the order flow does not change when an optimal rule replaces an inferior heuristic.

[8] The traditional assumption in the scheduling literature has been that jobs are tardy when completed after their due dates. By controlling for a parameter ρ we see whether there occurs a marked difference in the way jobs are treated by managers when tardy based on the classical definition or possibly at a different date in the neighborhood of the jobs' recorded due date.

some regularity conditions on the curvature of how tardiness affects cost, both directly and through its interactions with other job characteristics. Our theoretical framework analyzes two classes of preferences. The first assumes that costs increase geometrically with tardiness at a rate which is independent of the other characteristics of the job. The second assumes that the cost of scheduling a job is quadratic in tardiness, and that its quadratic coefficient is common across all jobs, but permits the linear coefficient to depend on the job's characteristics. Our discussion is summarized by the following formalization. Let $x'_{tj} \equiv (x_{tj1}, \ldots, x_{tj,K+1})$. In the geometric class

$$c(x_{tj}) = \begin{cases} \tau g(x_{tj}) & \text{if } \rho + x_{tj0} < 0, \\ \exp[c_0 x_{tj0} + c_1 x'_{tj}] g(x_{tj}) & \\ \quad \text{otherwise}, & \end{cases} \tag{2.7}$$

where c_0 is a positive real number, and $c_1 x'_{tj}$ is real valued, only depending on the fixed characteristics of x_{tj}. For the quadratic preference class

$$c(x_{tj}) = \begin{cases} \tau g(x_{tj}) & \text{if } \rho + x_{tj0} < 0, \\ (c_0 x'_{tj} & \\ \quad + x_{tj0} c_1 x'_{tj} & \\ \quad + x_{tj0}^2 c_2) g(x_{tj}) & \text{otherwise}, \end{cases} \tag{2.8}$$

where $c_0 x'_{tj}$ and $c_1 x'_{tj}$ are real valued functions of the fixed characteristics of x_{tj}, while c_2 is a positive real number.

Under these assumptions, scheduling tardy jobs is a straightforward matter. The optimal policy is to assign each job an index and prioritize them according to their index values. The index for job x_{tj} is defined as any monotonic transformation of

$$c(f(x_{tj}))/g(x_{tj}) - c(x_{tj})/g(x_{tj}). \tag{2.9}$$

In particular, the index for preference class (2.7) specializes to

$$i(x_{tj}) = \exp(c_0 x_{tj0} + c_1 x'_{tj}) \tag{2.10}$$

and to

$$i(x_{tj}) = c_2 2 x_{tj0} + c_1 x'_{tj} \tag{2.11}$$

for preference class (2.8).

Thus pairwise comparisons between tardy jobs are sufficient to determine their relative importance; tardy jobs with higher marginal penalties (index values) get scheduled first. Proposition 1 below formally states this partial characterization of the optimal policy. [9] Note though, that without further restrictions on τ (the cost of scheduling a unit of an early job), Proposition 1 does not fully characterize an optimal forward rule; it only does so for tardy jobs.

Proposition 1. *Suppose that* $\{x_{ti}, x_{tj}\} \subseteq X$, $\rho + x_{ti0} \geqslant 0$ *and* $\rho + x_{tj0} \geqslant 0$ (*both jobs are tardy*). *If* $i(x_{tj}) > i(x_{ti})$, *then* x_{tj} *must precede* x_{ti} *in the optimal schedule.*

Appendix A contains the proofs to all the propositions. [10]

3. Estimating preferences

The introduction explained the basis for making inferences about the manager's preferences: if some jobs are relatively straightforward to schedule optimally, then his preferences should be reflected in scheduling such jobs. Indeed, if this was false, it would be instructive to compare discrepancies between the enunciated objectives of the firm with the ones the plant manager's actions reveal. As we mentioned in the introduction, such discrepancies might arise from not aligning the incentive structure to organizational goals, or from poor channels for communicating those goals through the hierarchy. The previous section showed that scheduling tardy jobs is much easier than scheduling early jobs in our framework, and so we estimate preferences from decisions made for tardy jobs. Our approach estimates the parameters underlying the cost function $c(x_{tj})$ by simulating schedules for tardy jobs, which are then matched with the managers' decisions. This section accordingly analyzes how to estimate these preferences.

[9] This proposition is very similar to the WSPT rule in the deterministic scheduling literature. See (Baker, 1979; Morton, 1993) for examples.

[10] The proof for Proposition 1 relies on an interchange argument. Since the costs increase from one day to next, it depends on the fractional assumption for d_{tj}, so as to allow partial jobs to be completed towards the end of the day, instead of trying to find a job that will fit exactly in the remaining time available.

We index the class of parameterizations by θ, now writing $c(x_{tj0}, \theta^o)$ for $c(x_{tj0})$, where $c(x_{tj}, \theta)$ is a known function and $\theta^o \in \Theta$ is the unknown parameter vector to be estimated from data on job flow. If all the information about jobs at the manager's disposal was also recorded without errors in the data set, and the model was correctly specified, then conditional on x_t, the predicted rankings from the estimated framework would mimic those actually observed in the sample. Since a perfect fit is never attainable in practice, we further assume that there is information relevant to processing the jobs not contained in the data.

We investigate two cases: unobserved heterogeneity in the characteristics of the job, or measurement error in the due dates of jobs (due, perhaps, to revisions by customers about when they would like their orders fulfilled). Let x_{tjK} denote the recorded due date of a job. Accordingly suppose the characteristics $(x_{tj1}, \ldots, x_{tjK})$ are observed, whereas $x_{tj,K+1}$ is a random variable that is distributed independently and identically with probability distribution function $G(x_{tj,K+1}; \theta^o)$.

In the first case we assume that the punctuality of job shipments is measured without error:

$$x_{tj0} = t - x_{tjK}. \tag{3.1}$$

Since Proposition 1 allows us to characterize an optimal rule for jobs that are tardy, we can form an optimal rule for those jobs, x_{tj}, satisfying the inequality $\rho + x_{tj0} \geq 0$. Under this assumption, τ, the cost of scheduling early jobs, is unidentified, because we do not know what principles guide the scheduling of early jobs and therefore cannot exploit data on them.

In the second case we assume $\tau = \infty$ and that the true lateness x_{tj0} is given by [11]

$$x_{tj0} = t - x_{tjK} - x_{tj,K+1}. \tag{3.2}$$

Under Eq. (3.2), lateness or equivalently tardiness is measured with error. This discrepancy in the recorded due date, and the one the plant actually observes, could be due to changes in the due date of the order that were made after data entry because of further negotiations between the firm and the customer. [12] Now it is impossible for us to distinguish tardy jobs from early jobs, so additional restrictions must be imposed on τ, to apply Proposition 1 in estimation. We impose unbounded costs on early delivery, and thus completely characterize the optimal rule from Proposition 1 (for it is never optimal to process early jobs if $\tau = \infty$). The two preference classes illustrate how subtle differences in assumptions with regard to unobservables can change the characterization of optimal rules, as well as the estimation procedure.

3.1. Estimation procedure

Suppose the conditions for Proposition 1 are met, and the plant follows the simple optimal decision rule for completing orders under either preference class (2.7) or (2.8). Then it is possible to identify and estimate preferences up to a parameterization for $c(x_{tj}, \theta^o)$.

We first partition the vector space X into X_ρ, and its complement

$$X_\rho = \begin{cases} \{x_{tj} : x_{tj} \in X \text{ and } \rho + x_{tj0} \geq 0\} & \text{if } \tau < \infty, \\ X & \text{if } \tau = \infty. \end{cases} \tag{3.3}$$

This partition identifies which jobs were tardy under assumption (3.1) and therefore can be used in estimation. Under Eq. (3.2), X_ρ consists of all jobs since a full characterization of the optimal policy is possible. In principle, a maximum likelihood (ML) estimator could be formed from the probability that the observed schedule occurs conditional on x_t at t. However, because of the computational difficulties associated with implementing ML, we adopted a Method of Simulated Moments (MSM) strategy instead. [13] Intuitively the MSM estimator chooses θ to match

[11] In both assumptions the noise term $x_{tj,K+1}$ is time invariant. An order is assigned a unique noise term that is propagated under $f(x_{tj})$ from one day to the other. Notice that Proposition 1, which assumes that only tardiness changes with time, will not hold if $x_{tj,K+1}$ changes over time.

[12] Our analysis can be easily extended to account for unobserved characteristics, as well as measurement error in the other characteristics.

[13] The ML estimator is based on the likelihood that the processed jobs had an index higher than the other outstanding orders in X_ρ. Since the number of outstanding belonging to X_ρ is on average 500, numerous multifold integrals must be computed to evaluate the likelihood of any candidate parameter vector.

sample moments of the schedule with those of a schedule for a simulated set of jobs that have the same distributional characteristics as the actual jobs. Under some regularity conditions described in Proposition 2, the MSM estimator, denoted by $\hat{\theta}$, converges to θ^o, and $T^{1/2}(\hat{\theta}-\theta^o)$ is an asymptotic normal random variable centered at 0.

More specifically, the MSM estimation procedure generates N sample paths as follows. First, a normal random number generator assigns (simulates) a value $x^n_{tj,K+1}$ for the unobserved component of each job x_{tj}. The random variable $x^n_{tj,K+1}$ is uniquely assigned once for each job in the data at the beginning of the simulation, and is therefore invariant in time. This value is propagated under $f(x^n_{tj})$. We define

$$x^n_{tj} \equiv \left(x^n_{tj0}, \ldots, x_{tjK}, x^n_{tj,K+1}\right) \tag{3.4}$$

as the simulated analogue of the jth job in x_t. Under the assumptions in (3.1) $x^n_{tj0} = x_{tj0}$. Under the assumptions in Eq. (3.2) tardiness is now measured as $x^n_{tj0} = t - x_{tjK} - x^n_{tj,K+1}$ for each simulated job. Each day the vector sum of the J_t simulated outstanding orders is

$$x^n_t = \sum_{j=1}^{J_t} x^n_{tj}. \tag{3.5}$$

Second, the simulated jobs are ranked by their respective index values. Under the first preference class tardy jobs are assigned a positive index value $i(\cdot)$, and under the second preference class, tardy jobs are assigned a positive index value while early jobs are assigned a value of $-\infty$. Accordingly let

$$i\left(x^n_{tj};\, \theta\right) = \begin{cases} c\left(f\left(x^n_{tj}\right),\, \theta\right)/g\left(x^n_{tj}\right) - c\left(x^n_{tj},\, \theta\right) \\ \quad /g\left(x^n_{tj}\right) \quad \text{if } \rho + x^n_{tj0} \geqslant 0, \\ -\infty \quad \text{otherwise,} \end{cases} \tag{3.6}$$

denote the index of x^n_{tj} as a function of θ, and let $r(x^n_{tj};\, \theta)$ denote its rank amongst the J_t simulations for that day. That is,

$$r\left(x^n_{tj};\, \theta\right) = \#\left\{x^n_{tk} : i\left(x^n_{tk};\, \theta\right) \geqslant i\left(x^n_{tj};\, \theta\right)\right\}, \tag{3.7}$$

where $\#\{\cdot\}$ denotes the cardinality of set $\{\cdot\}$.

Jobs are scheduled in order of their rankings, provided that there is enough time (capacity) available and that they do not have an index value of $-\infty$ (that is, provided they are not early). In this fashion the simulations generate an index for some particular parameter value θ, in the process also determining the schedule for that day t. So the third step is to simulate scheduling for the hypothetical tardy jobs. Denote by

$$T(t) \equiv \sum_{j=1}^{J_t} d_{tj} 1\left\{x_{tj} \in X_\rho\right\} g\left(x_{tj}\right) \tag{3.8}$$

the available processing time (or actual capacity) for the subset of outstanding jobs that are in X_ρ at t, where $1\{\cdot\}$ is 1 if its argument is true and 0 otherwise. We define $h(x^n_{tj};\, \theta)$ as the difference between the actual schedule and the simulated outcomes that would have occurred if θ had been the true parameter value, and $\{x^n_{tj,K+1}\}$, not $\{x_{tj,K+1}\}$, were the shocks for simulation n. Then for each observation, we form

$$h\left(x^n_{tj};\, \theta\right) = \begin{cases} d_{tj} - 1 \\ \quad \text{if } \sum_{k=1}^{J_t} 1\left\{x^n_{tk} \in X_\rho\right\} 1\left\{r\left(x^n_{tk};\, \theta\right)\right. \\ \quad \left. \leqslant r\left(x^n_{tj};\, \theta\right)\right\} g\left(x^n_{tk}\right) \leqslant T(t) \\ \quad \text{and } i\left(x^n_{tj};\, \theta\right) \geqslant 0, \\ d_{tj} \quad \text{otherwise,} \end{cases} \tag{3.9}$$

that is, the difference between the indicator variable for the actual schedule and that in the simulation. To interpret (3.9) note, for example, that jobs that are not scheduled have a simulation indicator of 0, which means that the difference between the actual schedule and the simulation is d_{tj}. By construction

$$E\left[h\left(x^n_{tj};\, \theta^o\right)\right] = 0. \tag{3.10}$$

In expectation the actual and the simulated schedule are the same at the true value of the unknown parameter vector. This condition provides the basis for establishing that $T^{1/2}(\hat{\theta}-\theta^o)$ converges to a normally distributed random variable with mean 0. For each date t, we construct an r-dimensional vector $h_r(x^n_t;\, \theta)$, whose elements are differences, $h(x^n_{tj};\, \theta)$, obtained by picking r jobs at random from the outstanding orders in X_ρ. These r jobs

along with N simulations are used to construct a vector of r sample moments, $h_t^{(T)}(\theta)$,

$$h_t^{(T)}(\theta) = \frac{1}{N}\sum_{n=1}^{N} h_r(x_t^n;\theta). \tag{3.11}$$

The fourth step obtains a first round estimator, $\tilde{\theta}$, by minimizing the criterion function: [14]

$$\left\{\frac{1}{T}\sum_{t=1}^{T} h_t^{(T)}(\theta)\right\}' A \left\{\frac{1}{T}\sum_{t=1}^{T} h_t^{(T)}(\theta)\right\}, \tag{3.12}$$

with respect to θ, where A is an $r \times r$ dimensional identity matrix.

Then we calculate the estimated covariance matrix:

$$\tilde{\Sigma} = \frac{T}{T-q}\sum_{j=-T+1}^{T-1} \mathscr{K}\left(\frac{j}{b}\right)\tilde{\Gamma}(j), \tag{3.13}$$

where q is the number of parameters to be estimated and $\mathscr{K}(\cdot)$ is the Tukey–Hanning Kernel (Andrews, 1991):

$$\mathscr{K}(x) = \begin{cases} (1+\cos(\pi x))/2 & \text{for } |x| \leqslant 1, \\ 0 & \text{otherwise,} \end{cases} \tag{3.14}$$

b is a bandwidth parameter which determines how much correlation between periods is accounted for in a sample of finite size and [15]

$$\tilde{\Gamma}(j) \equiv \begin{cases} \dfrac{1}{T}\displaystyle\sum_{t=j+1}^{T} h_t^{(T)}(\tilde{\theta}) h_{t-j}^{(T)}(\tilde{\theta})' & \text{for } j \geqslant 0, \\ \dfrac{1}{T}\displaystyle\sum_{t=-j+1}^{T} h_{t+j}^{(T)}(\tilde{\theta}) h_t^{(T)}(\tilde{\theta})' & \text{for } j < 0. \end{cases} \tag{3.15}$$

The last step begins by constructing an optimal weighting matrix, setting A to $[\tilde{\Sigma}]^{-1}$ which we use when minimizing (3.12) with respect to θ a second time to obtain our (final) estimator $\hat{\theta}$. As we show below, the asymptotic covariance matrix for $\hat{\theta}$ is

$$\left(Q'[\Sigma]^{-1}Q\right)^{-1}, \tag{3.16}$$

where

$$Q = E_0\frac{\partial h_t(\theta^o)}{\partial \theta} \tag{3.17}$$

and

$$\Sigma = E_0\left[\sum_{s=-M}^{M} h_t(\theta^o) h_{t+s}(\theta^o)'\right], \tag{3.18}$$

where M is an (unknown) finite integer which denotes the degree of temporal independence. (See Assumption 2 below.) Consistent estimates of Σ and Q are respectively $\tilde{\Sigma}$ and $\tilde{Q}$ defined as

$$\tilde{Q} = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial h_t^{(T)}(\hat{\theta})}{\partial \theta}. \tag{3.19}$$

(The partial derivatives for the sample moments are obtained numerically by finite differencing.)

3.2. Large sample properties

The basis for this estimation procedure rests on its asymptotic properties, which can be derived from the following assumptions.

Assumption 1. The class of functions $h(x_{tj}^n, \theta)$ indexed by $\theta \in \Theta$, is Euclidean. Also θ^o belongs to the interior of Θ and is the unique root for $E[h(x_{tj}^n; \theta)]$.

This regularity assumption is typically satisfied when Θ is a closed convex space in $\mathbb{R}^k$ (implying that θ^o is a k-dimensional parameter to be estimated). [16]

Assumption 2. $\{x_t\}$ is a stationary finite dimensional markov process.

[14] The dimension of $h_r(x_t^n; \theta)$, namely, r, can at most be equal to the number of outstanding orders on date t that also are in X_ρ. However, this number changes with t, since different days have different collections of outstanding orders. This means that when we construct sample moments using $h_t^{(T)}(\cdot)$ we cannot average them evenly. There are several ways of dealing with this problem. In our application we restrict r to be the smallest number of jobs that on any date t also are in X_ρ.

[15] The larger the value of b the greater the correlation that is accounted for in estimation. Thus correlation between observations greater than b periods apart is ignored.

[16] Pakes and Pollard (1989) provide a useful discussion on how to check whether a class of functions is Euclidean or not.

Loosely speaking, this assumption says something about how well the sample represents the process generating the data. It is weaker than independence, permitting correlation between x_s and x_t providing $|s-t|$ is less than some finite integer M. Even if the order flow process from the customers is independent over time, the flow from w_t into x_t will typically be time dependent. This occurs, for example, if orders are processed upstream (at various upstream plants) in batches of jobs that have similar characteristics. [17]

Assumption 3. $g(x_{tj})$ is measured without error.

This assumption, made to facilitate the exposition, imposes the restriction that the processing times of the jobs are known and can be easily relaxed, see (Miller and Ramnath, 1994b).

These assumptions are used to establish the consistency and asymptotic normality of $\hat{\theta}$.

Proposition 2. *Let $\tilde{\theta}$ be a $k \times 1$ vector which minimizes* (3.12), *when A is set to the $r \times r$ identity matrix, and define $\tilde{\Sigma}$ using expressions* (3.13) *through* (3.15). *Let $\hat{\theta}$ be the $k \times 1$ vector which minimizes* (3.12) *when A is set to $\tilde{\Sigma}^{-1}$. Then $\hat{\theta}$ converges to θ^o in probability and $T^{1/2}(\hat{\theta} - \theta^o)$ is asymptotically distributed as a normal random variable with mean* 0 *and variance* $(Q_0 \Sigma_0^{-1} Q_0)^{-1}$. *A consistent estimator for* $(Q_0 \Sigma_0^{-1} Q_0)^{-1}$ *is* $(\tilde{Q}' \tilde{\Sigma}^{-1} \tilde{Q})^{-1}$.

4. An application

The latter half of this article illustrates how the techniques described in the previous sections can be applied to a plant that produces steel tubes. We applied the methods developed in Section 3 to estimate the manager's preferences. Then the heuristics described below in Section 5 are run on the plant and evaluated relative to the current procedures, as well as to each other. We conclude the paper with a discussion of some managerial implications.

4.1. The plant

The plant in question is a steel-tube mill which buys steel billets and coils, and produces customized tubing tailored to the client's specifications. A detailed description of the plant floor and the data is given in (Miller and Ramnath, 1994b). Briefly, there are two parts to the production process: milling (which turns the raw materials into tubes of approximately desired dimensions), and finishing (where the tubes are made to conform exactly to customer specifications). Tubes are welded from steel coil, which is first unrolled and twisted with the use of massive dies to form a tube. Thereafter the twisted ends of the coil are welded together. Several stages comprise the subsequent finishing process. After annealing (heating to eliminate the stress points), the tube is pickled (cleaned in an acid bath). Then to bring the tube closer to the customer's specifications, it is push-pointed to create a rough edge to allow a vice to grip the tube, and cold drawn (clamped at both ends in a vice and stretched), annealed and pickled again. This process is repeated until the outer diameter and thickness of the tubes are within the tolerance levels of the customer's specifications, at which point the tube is cut to size, packaged and shipped. The milling process is relatively capital intensive, and the finishing labor intensive. Each sequence of annealing, pickling, push-pointing, cold-drawing comprises a pass in the finishing process. Different tubes undergo different numbers of passes before being shipped out to the customer. [18]

The orders from customers which are welded upstream generate w_t. This flow of orders generates work-in-process in front of the finishing stage, which we treat here as the set of outstanding orders x_t, to be processed. The scheduling at the finishing stage is the process being estimated. The assumptions regarding the processing-time characteristics allow us to

[17] In our application consecutive observations are correlated with each other because job orders are not processed immediately. The lead time between the appearance of a job order and its actual shipment date is, however, finite. Therefore, assuming the process of job order arrivals exhibits only finite dependence, observations sufficiently apart from each other are also uncorrelated.

[18] Bertrand et al. (1990) discuss a similar problem faced by a steel tube factory in The Netherlands.

abstract the finishing stage as a single aggregate machine, where the processing time of any job is proxy for the daily capacity utilized at this stage by it.

The data used to analyze this problem are routinely collected by the firm to facilitate scheduling, and retained to identify jobs coming from the same batch in case they suffer from common defects. They consist of 17,206 orders processed through the plant over a period of 1,055 days starting from 1 January 1989. Over this period the plant operated for 843 days. After preliminary cleansing, the data used in our study is on 13,647 jobs for a period of 790 working days, a working day being defined as a day when the finishing stage was operational. [19] Of these 790 days, only 600 are used for the estimation, while the entire set is used for reporting summary statistics and conducting the counterfactual experiments. The data are organized on an order-by-order basis. Each entry includes the customer number, the requested outer diameter of the tube, the requested wall thickness, the requested weight, the requested footage, the dollar value of the order, the actual outer diameter of the welded tube, the actual wall thickness of the welded tube, the actual weight of the order when it was welded, the number of passes that the order went through at the finishing stage, the date that the order was entered in the records, the actual date that the order was welded, the week that the order was promised to be shipped to the customer, the date that the customer had requested shipment (which we treat in this paper as the due date), and the actual date when the order was shipped. The 8 largest customers placed 49 percent of the orders, each one placing more than 500 orders in total over the three-year period. Nineteen percent of the orders are placed by about ten medium-sized customers who placed between 200 and 500 orders in total. The remaining 32 percent of the orders were placed by the remaining small customers. There are 183 customers in total, with many of them having placed between one and twenty orders over the three-year period.

[19] 2328 orders with either invalid welding dates, or shipping dates were first deleted. Then orders with invalid prices, or passes numbering a total of 393, were removed. Since orders are entered by the firm in order of entry dates and not on the actual dates they were shipped or welded, the information about what happened at the plant on the first and the last few days is incomplete. Therefore we delete all observations about orders welded before the first 40 days and those shipped after the 1026th day. As a result another 838 observations were dropped.

Table 1
Summary statistics for job characteristics (standard deviations in parenthesis)

Characteristic		Average
Tardiness (days)	$\max(0, x_{tj0})$	23.66 (43.52)
Urgency (days)	x_{tj1}	74.22 (46.2)
Customer size (number of orders)	x_{tj2}	758.97 (672.87)
Order size (feet)	x_{tj3}	1071.88 (1922.32)

Table 1 reports some of the average statistics for the characteristics of the orders. The term x_{tj0} denotes, as before, the job's lateness in days on the current date t, and let x_{tj1} represent the urgency of the order in days, that is, the difference between its recorded due date and the date the order was placed. The size of the customer who placed the order which is evaluated as the total number of orders placed over a period of 3 years is denoted by x_{tj2}. Denote by x_{tj3} the size of the order in feet. The figures in Table 1 are based on averages shipped per day. For example, since x_{tj1} denotes urgency, the urgency per day is defined as

$$\frac{1}{\sum_{j=1}^{J_t} d_{tj}} \sum_{j=1}^{J_t} d_{tj} x_{tj1}, \tag{4.1}$$

and Table 1 reports the mean and standard deviation of the above random variable for $t = 30 \ldots 820$ days. In other words, the table reports the mean $\bar{x}_{tji}$, which is the mean of the characteristic x_{tji}:

$$\bar{x}_{tji} = \frac{1}{790} \sum_{t=31}^{820} \frac{1}{\sum_{j=1}^{J_t} d_{tj}} \sum_{t=1}^{J_t} d_{tj} x_{tji}. \tag{4.2}$$

In the table tardiness is evaluated as 0 if the job was shipped ahead of its due date, and is equal to its lateness otherwise.

4.2. Estimating preferences from the finishing stage

The methods described in Section 3 were applied in order to estimate 2 parameterizations of the plant manager's objective function. Not surprisingly, the manager was not able to provide us with a paramet-

ric specification of his preferences, so we felt that estimating two might help us gauge the sensitivity of our empirical results to alternative assumptions about structure. In addition, important differences between the findings generated by the respective specifications might illuminate aspects of the data that could otherwise remain hidden in the process of imposing a particular model structure on the data.

In addition to the characteristics defined before, let x_{tj4} denote the recorded due date of the order, and x_{tj5} the measurement error which is a random variable generated (for each sample path) from a normal distribution with mean 0 and standard deviation θ_4 (another parameter to be estimated). In the parameter specifications below we also assume that θ_0, θ_1, θ_2 and θ_3 are associated with x_{tj0}, x_{tj1}, x_{tj2} and x_{tj3}, respectively, for some $\theta \in \Theta$. The estimator of the true unknown $\theta^o \in \Theta$, denoted $\hat{\theta}$, is based on how the data shows the work in process inventory is finished for shipping. As described in Section 3 the disturbance x_{tj5} is invariant over time.

Both parameterizations fall within the exponential parameter class (2.7).

(i) Eq. (3.1) holds, and $\rho = 0$. For all jobs for which $\rho + x_{tj0} \geq 0$,

$$c_0(x_{tj};\theta) = \exp(\theta_0 x_{tj0}) \exp(\theta_1 x_{tj1} + \theta_2 x_{tj2} + \theta_3 x_{tj3} + x_{tj5}) g(x_{tj}). \quad (4.3)$$

In the scheduling literature the first term $\exp(\theta_0 x_{tj0})$ would denote the penalty for tardiness, whereas the remaining part of the cost function that is invariant in time would represent the weight of the job, making the scheduling cost an exponentially increasing weighted tardiness function. The cost of tardiness for this specification is increasing geometrically in tardiness. This weight of the job, namely

$$\exp(\theta_1 x_{tj1} + \theta_2 x_{tj2} + \theta_3 x_{tj3} + x_{tj5}) g(x_{tj}), \quad (4.4)$$

depends on the urgency of the order, the size of the customer who placed the order, its size in footage and the unobservable characteristics. Because of the assumptions in Eq. (3.1), the measurement error is only over this time-invariant part of the cost of scheduling the job. From (2.7) and (4.3),

$$c_0 \equiv \theta_0,$$

$$c_1(x_{tj1}, \ldots, x_{tj5}) \equiv \theta_1 x_{tj1} + \theta_2 x_{tj2} + \theta_3 x_{tj3} + x_{tj5}.$$

From Eqs. (2.10) and (3.6) the index is

$$i(x_{tj}^n;\theta) = \exp(\theta_0 x_{tj0}) \exp(\theta_1 x_{tj1} + \theta_2 x_{tj2} + \theta_3 x_{tj3} + x_{tj5}^n). \quad (4.5)$$

Since for this specification early jobs are never included in X_ρ, no jobs take on an index value of $-\infty$.

(ii) Eq. (3.2) holds, $\rho = 7$ and $x_{tj0} = t - x_{tj4} - x_{tj5}$. For all jobs $x_{tj} \in x_t$ such that $\rho + x_{tj0} \geq 0$,

$$c_0(x_{tj};\theta) = \exp(\theta_0 x_{tj0}) \exp(\theta_1 x_{tj1} + \theta_2 x_{tj2} + \theta_3 x_{tj}) g(x_{tj}). \quad (4.6)$$

In this case, since the assumption in Eq. (3.2) is assumed to hold, the measurement error occurs only in the lateness term (or equivalently in tardiness) and therefore in the tardiness cost, $\exp(\theta_0 x_{tj0})$. The cost of tardiness is geometric after accounting for the measurement error. Otherwise, this parameterization is similar to specification (i). Since $\rho = 7$, jobs earlier than 7 days will never be scheduled using an optimal decision rule because under Eq. (3.2) the cost of scheduling early jobs is ∞. The index for specification (ii) is very similar to the index for (i). For those jobs that are tardy, that is jobs for which $x_{tj0}^n \equiv \rho + t - x_{tj4} - x_{tj5}^n \geq 0$,

$$i(x_{tj}^n;\theta) = \exp(\theta_0 x_{tj0}^n) \exp(\theta_1 x_{tj1} + \theta_2 x_{tj2} + \theta_3 x_{tj3}). \quad (4.7)$$

In addition, early jobs (which will incur an infinite cost if scheduled on date t) are assigned an index, $i(x_{tj}^n;\theta) = -\infty$, so that they will never be scheduled.

We set the number of sample moments r to 55, the number of sample paths to N to 3 and the bandwidth b for computing the covariance matrix (using the Tukey–Hanning Kernel defined in (3.14)) to 10. The size of the data T is 600 (days). The criterion function in (3.12) is constructed, by finding r sample moments for $t = 1$ through 600. The simulations are repeated for each θ as the minimization routine repeatedly evaluates the criterion function. At the end of the second round the covariance matrix

Table 2
Estimation of preferences (asymptotic standard errors in parentheses)

Characteristic	Variable	Coefficient	(i) $\exp(\theta_0 x_{tj0})\exp(\theta_1 x_{tj1} + \theta_2 x_{tj2} + \theta_3 x_{tj3} + x_{tj5})g(x_{tj})$	(ii) $\exp(\theta_0, x_{tj0})\exp(\theta_1 x_{tj1} + \theta_2 x_{tj2} + \theta_3 x_{tj3})g(x_{tj})$
Tardiness	x_{tj0}	$\hat{\theta}_0$	2.1×10^{-4} (1.75×10^{-5})	1.15×10^{-4} (1.84×10^{-5})
Urgency	x_{tj1}	$\hat{\theta}_1$	-9.1×10^{-4} (8.57×10^{-6})	4.27×10^{-3} (6.78×10^{-5})
Customer	x_{tj2}	$\hat{\theta}_2$	1.07×10^{-3} (3.24×10^{-7})	2.81×10^{-5} (3.98×10^{-5})
Feet	x_{tj3}	$\hat{\theta}_3$	2.99×10^{-4} (2.65×10^{-6})	1.58×10^{-4} (2.87×10^{-5})
Error	x_{tj5}	$\hat{\theta}_4$	8.49×10^{-3} (1.3×10^{-4})	1.055 (4.57×10^{-1})
J_T			37.70	43.69
p			0.90	0.77

for $\hat{\theta}$ in (3.16) is obtained by varying $\hat{\theta}$ in each dimension by 1 percent while computing the partials in (3.19). [20]

The results of the estimation procedure for specification (i) and (ii) are shown in Table 2. In parameterization (i), $\hat{\theta}_1$ the coefficient governing the urgency or orders is negative, the more urgent an order is, the more it adds to cost. [21] Since both $\hat{\theta}_2$ and $\hat{\theta}_3$ are positive, orders have increased priority when placed by big customers, or when they are large. All the coefficients $\hat{\theta}_0, \ldots, \hat{\theta}_4$ are significantly different from zero. Using Tables 1 and 2, we can compare the sensitivity of the cost function with respect to any particular characteristic. [22] Letting $\bar{x}_{tj} = (\bar{x}_{tj0}, \ldots, \bar{x}_{tj5})$ be a nominal job, with average characteristics using up one unit of processing time, $c(\bar{x}_{tj}, \hat{\theta})$ denotes the cost of scheduling it. Our measure of sensitivity is obtained by increasing each characteristic of $\bar{x}_{tj}$ by one standard deviation and noting the percentage increase with respect to $c(\bar{x}_{tj}, \hat{\theta})$. Evaluated at $\bar{x}_{tj}$, the cost $c(\bar{x}_{tj}, \hat{\theta})$ is 2.916, and if we increase the average tardiness by one standard deviation, that is, from 23.66 days to 67.18, then the cost increases by $\exp(43.52 \times 2.1 \times 10^{-4})$, by 0.92 percent. [23] Similarly increases in urgency x_{tj1}, customer size x_{tj2}, order size x_{tj3}, or the noise term x_{tj5} (with a mean of 0) by one standard deviation result in a percentage increase of $c(\bar{x}_{tj}, \hat{\theta})$ by -4, 105, 77 and 0.86, respectively. Therefore, the most important variables as far as the cost function goes, in determining the job priorities are customer size and order size. In other words, customer size is the variable most likely to determine why a particular job gets scheduled ahead of another. The increase in costs on account of tardiness alone is about 2.1×10^{-2} percent a day, as a result of which tardiness becomes an important criterion (everything else being equal) only for very tardy jobs. Although the coefficient for the unobservable x_{tj5}, namely its variance $\hat{\theta}_4$, is significantly different from 0, its sensitivity is small relative to the other characteristics. One can test the overidentifying restrictions implied by the parameterization and the sample moments. [24] In the case of the sample moments for parameterization (i), the test shows that the model has a p value of 0.90, or that the restrictions cannot be rejected at the 90 percent significance level. [25]

[20] A simplex method using the IMSL math library was called to minimize the criterion function.

[21] Since urgency is defined as the difference between the request date and the date the order was placed, a smaller value for urgency means that a shorter response time is available for the plant, and therefore the more urgent the jobs are.

[22] In other words, we compare how the index (or cost) changes across a cross-section of jobs available on the shop floor.

[23] The term $\bar{x}_{tj5}$, that is, the mean of x_{tj5}, is zero.

[24] See (Hansen, 1982). The test is based on the fact that under the null hypothesis, the function J_T defined by T times the optimally weighted criterion function is asymptotically χ^2 distributed with degrees of freedom equal to the number of overidentifying restrictions.

[25] We did try a larger value for r (105) and N (5), however, the asymptotic gains from these larger values were minimal. In addition a larger value for N implies that there are as many simulations to be conducted while constructing the orthogonality conditions. This gets to be computationally expensive.

All the coefficients are significant in parameterization (ii) except for customer size. The coefficient for urgency is now positive, implying that less urgent orders have higher priorities than more urgent ones. An intuitive explanation for this sign reversal exists. The second parameterization sets in, a week earlier than the first. Consequently, simulated jobs which are regarded as early in (i), and are therefore not included in the analysis, are being defined as late in (ii), and hence matched up to their real counterparts in estimation. Although including these additional jobs is not the only feature which differentiates the parameterizations, we are tempted to conclude that their inclusion might be the source of the sign reversal. The interpretation is that greater urgency increases the cost index if a job is past its due date but receives a lower priority if its due date has not yet arrived. As we explain below, this plausible interpretation should be viewed as evidence against specification (ii). However, the test of the overidentifying restrictions shows that they cannot be rejected at the 77 percent significance level. Comparing sensitivities again, we note that a nominal job has a cost $c(\bar{x}_{tj}, \hat{\theta}) = 1.667$. (To obtain $\bar{x}_{tj0}$ for this parameterization we add 7 days to the average in Table 1 since ρ is assumed to be 7 for this parameterization.) Increases in tardiness, urgency, customer size, order size, and the unobservable by one standard deviation results in the increase of this nominal cost by 0.50, 37.29, 1.9, 35.4 and 0.012 percent, respectively. So, for parameterization (ii), the most relevant terms in determining priorities are urgency and order size.

Several variations on these two specifications were estimated. Adding other variables, such as the price of the order and the total dollar value from the customer, did not yield statistically significant differences. [26] More specifically, the additional variables were not significant, and the estimated values of the existing coefficients did not change significantly. [27] Unfortunately there is no straightforward way of statistically evaluating the relative performance of the two specifications models, short of nesting them within a common statistical framework. This is a computationally cumbersome exercise to undertake in a nonlinear model like ours, and was not attempted. One could use the Akaike (1973) information criterion to assess them, although we do not find the justification for this procedure convincing.

One further recourse for choosing between competing specifications is to critically study the values of the estimated parameters. As discussed above, we uncovered indirect evidence that the cost of processing an order is not monotonic in urgency. Before their due date, processing less urgent orders seem to be preferred, but after the due date more urgent orders have higher index values. If this is true, the index used in (ii), which assumes there is no interaction between tardiness and the fixed characteristics (specifically the due date), cannot be optimal. Moreover, we cannot appeal to the same argument in the case of the first parameterization, because its indices only apply to jobs processed after the due date. Thus we can reject the second parameterization by a contradiction argument, but not the first. Also the plant manager may have views regarding the relative merits of competing specifications. In this particular application, our findings that the coefficient on urgency takes on a different sign depending on which parameterization is adopted, and that customer size is not significant in (ii) helps him to differentiate between them. Taken overall, these considerations led us to favor the first parameterization.

5. Counterfactual experiments

As we have demonstrated above, the estimated parameters can be used to interpret the nature of the

[26] In addition we tried the linearly weighted case as found in the scheduling literature, but the estimated coefficients were statistically insignificant. Furthermore, a linearly weighted assumption would permit an insignificant job to stay indefinitely on the shop floor. However, even in our application where the shop is always heavily loaded, we never observed such a scenario.

[27] One could also include interaction terms between the fixed characteristics. However, interaction terms between tardiness and the other variables destroy the optimality of our simple index rule. In this case, one could take a different approach to estimation, and we are currently working on this problem. See (Miller and Ramnath, 1994b).

objective functions of the firm when the managers make scheduling decisions. Another use for them is to suggest alternative decision rules for the firm's scheduling decisions. A third use is in forecasting the effects of structural changes on the shop floor and the probability distribution generating job orders. We now examine these latter two uses.

5.1. Evaluating alternative heuristics

Under parameterization (i), only a partial characterization of the decisions at the finishing stage is possible using a forward rule; it can only be characterized for jobs that are tardy. With a tardiness assumption, it is well known that a complete characterization of the problem is NP hard, even for the simpler linear cost found in the literature. This is because once all jobs are considered to be scheduled, the tradeoff between early and tardy jobs is hard to determine. Under such circumstances it is not realistic to think that all the scheduling decisions of managers are optimal, but only decisions comparing tardy jobs. Consequently, the scheduling of early jobs, can potentially be improved upon by constructing different heuristics for the finishing stage.

Our approach cannot directly compare the plant's performance with those of alternative scheduling rules, unless the managerial decisions are coded, because the unobservables known to the managers when they make their decisions affect the plant's objective function. However, we can compare the performance of the plant with our decision rules on the basis of other characteristics, such as average or maximum tardiness. We can also compare different decision rules, on the basis of the evaluated objective function, by simulating the unobservables.

To construct counterfactual histories, we infer effective daily capacity the available capacity $1-\xi_t$ using (2.5),

$$1-\xi_t = \sum_{j=1}^{J_t} d_{tj} g(x_{tj}). \tag{5.1}$$

We next simulate the unobservable x_{tj5} to obtain an equivalent job x_{tj}^* that has the same observable characteristics as the actual job but with simulated characteristics replacing the unknown ones. Given this sequence of capacities $\{1-\xi_t\}$, the simulated jobs $\{x_{tj}^*\}$, the processing times $\{g(x_{tj}^*)\}$, and the estimates for the objective function $c(x_{tj}^*, \hat{\theta})$, we generate scheduling outcomes for the finishing stage denoted $\{\{d_{tj}^*\}_{j=1}^{J_t^*}\}_{t=1}^{T}$, and the sum of outstanding orders from the decision rule, $\{x_t^*\}_{t=1}^{T}$, where J_t^* is the number of orders in x_t^* which emerges from sequentially applying the decision rule. [28] Statistics such as the evaluation of the objective function under the simulation, the average or maximum tardiness can then be easily obtained, for the same arrival process that generated the original data. Noting that the cost of scheduling early jobs is not identified in estimation, we assume that τ, the cost of scheduling a job more than ρ days early is 0. All the decision rules applied to the finishing stage are forward heuristics, using modifications of heuristics already available in the literature. For any date t, these heuristics generate an index $i^*(x_{tj}; \hat{\theta})$, and process orders in a decreasing order of their index values, to fully exploit effective daily capacity (if there are enough jobs to process). That is,

$$\sum_{j=1}^{J_t^*} d_{tj}^* x_{tj}^* \leqslant 1-\xi_t. \tag{5.2}$$

The decision rules are as follows:

Weighted Shortest Processing Time (WSPT):

$$i^*(x_{tj}; \theta) = i(x_{tj}; \theta). \tag{5.3}$$

This heuristic performs well when most jobs are tardy, or when the objective is to minimize weighted flowtime in contrast to the weighted tardiness objective implied in (2.7) or (2.8). In fact, from Proposition 1 if all the jobs were tardy it would be the optimal rule.

Rachamadugu and Morton (R&M):

$$i^*(x_{tj}; \theta) = \begin{cases} i(x_{tj}; \theta) & \text{if } \rho + x_{tj0} \geqslant 0, \\ i(x_{tj}; \theta)\, e^{k_{RM}(\rho + x_{tj0})} & \text{otherwise.} \end{cases} \tag{5.4}$$

[28] The estimation procedure in Section 3 and the previous one use only part of the data ($r = 55$) to estimate the parameters. In contrast, the simulation procedure here is over all the orders observed over the relevant time period. This approach allows us to compare how all the orders that arrive at the plant are scheduled when using different different decision rules.

This heuristic is very similar to the Rachamadugu and Morton heuristic in the scheduling literature (Morton, 1993), with certain modifications to fit the estimate objective function of the plant. The parameter k_{RM} is determined by finding the best schedule. Jobs that are less than ρ days early have the same priority as given by Proposition 1. However, jobs that are more than ρ days early have an exponentially decreasing priority (in earliness), the rate of decrease governed by k_{RM}.

Weighted Covert (WTCOVERT):

$$i^*(x_{tj};\theta) = \begin{cases} i(x_{tj};\theta) & \text{if } \rho + x_{tj0} \geqslant 0, \\ i(x_{tj};\theta)\dfrac{k_{\mathrm{COVERT}} + \rho + x_{tj0}}{k_{\mathrm{COVERT}}} & \\ \quad \text{if } \rho + x_{tj0} + k_{\mathrm{COVERT}} \geqslant 0, & \\ 0 \quad \text{otherwise.} & \end{cases} \tag{5.5}$$

The parameter k_{COVERT} is also determined to minimize the resulting average costs. As before, jobs that are less than ρ days early have a priority given by the optimal rule. However, jobs that are more than ρ days early, but less than $\rho + k_{\mathrm{COVERT}}$ days early have a linearly decreasing priority in earliness, which goes to 0 for jobs earlier than $\rho + k_{\mathrm{COVERT}}$.

The heuristics were run on the data from the 31st working day to the 820th for a total of 790 days. The top half of Table 3 compares the performance of the heuristics on a number of different dimensions, and also the plant (except for average cost per day on account of the aforementioned difficulty in knowing the unobservables). Since such counterfactual experiments apply only under the assumptions in (3.1), when managers cannot make completely optimal decisions, c_0 is obtained using parameterization (i), and therefore $\rho = 0$. The index $i(\cdot)$ is the same as that for the parameterization as in (4.5),

$$i\left(x_{tj}^*;\,\hat{\theta}\right) = \exp\left(\hat{\theta}_0 x_{tj0}^*\right)\exp\left(\hat{\theta}_1 x_{tj1}^* + \hat{\theta}_2 x_{tj2}^* + \hat{\theta}_3 x_{tj3}^* + x_{tj5}^*\right), \tag{5.6}$$

where the simulated unobservable x_{tj5}^* is generated from a normal distribution with mean zero and standard deviation $\hat{\theta}_4$. As in Table 1, tardiness is measured in terms of averages per day to smooth out outliers (particularly in the actual history) and therefore the maximum tardiness is

$$\max_{t\in\{31,\ldots,820\}}\left\{\frac{1}{\sum_{j=1}^{J_t} d_{tj}}\sum_{j=1}^{J_t} d_{tj}\max(0,x_{tj0})\right\}. \tag{5.7}$$

Similarly average tardiness is now given by

$$\frac{1}{790}\sum_{t=31}^{820}\frac{1}{\sum_{j=1}^{J_t} d_{tj}}\sum_{j=1}^{J_t} d_{tj}\max(0,x_{tj0}). \tag{5.8}$$

The average cost is similarly computed as

$$\frac{1}{790}\sum_{t=31}^{820}\frac{1}{\sum_{j=1}^{J_t^*} d_{tj}^*}\sum_{j=1}^{J_t^*} d_{tj}^* c\left(x_{tj}^*\right). \tag{5.9}$$

In the above equations d_{tj}^* replaces d_{tj} and J_t^* replaces J_t, when the statistics are obtained for the heuristics. Using averages per day as summary statis-

Table 3
Counterfactual experiments (asymptotic standard errors or standard deviations in parentheses)

Nominal settings	R&M $k_{\mathrm{RM}} = 1\times10^{-6}$	WTCOVERT $k_{\mathrm{COVERT}} = 79$	WSPT	Plant
Average cost	0.064 (0.027)	0.065 (0.020)	0.064 (0.027)	
Average tardiness (days)	27.15 (8.27)	25.74 (3.27)	27.15 (8.08)	23.66 (43.5)
Maximum tardiness (days)	401.67	353.29	401.67	151.33
5 percent increase in capacity				
Average cost	0.059 (0.019)	0.055 (0.028)	0.059 (0.019)	
Average tardiness (days)	15.08 (2.53)	13.46 (2.92)	15.08 (2.52)	
Maximum tardiness (days)	285.8	279.52	285.8	

tics also enables the construction of the asymptotic properties for these heuristics as shown in (Miller and Ramnath, 1994b). [29] Since all the heuristics use all the available capacity at the finishing stage without enforcing idleness, the cost of the residual work in process at the end of the 820th day, $\{x_t^*\}_{t>820}$ is treated as zero. A similar treatment is made of the residual work in process in the data.

From the table it can be seen that R&M and WSPT had lower average costs. The constant $k_{RM} = 1 \times 10^{-6}$ is very small, which means that the R&M indices are almost the same as the index for WSPT. Similarly, the constant for WTCOVERT, that is k_{COVERT}, is 79, which means its index is also very close to the WSPT index, especially for jobs that are not extremely early. Contrary to the conventional WSPT rule for the linearly weighted tardiness objective in the scheduling literature, this particular version increases in value with increasing tardiness, as seen by the construction of the index $i(x_{tj}; \theta)$ in Eq. (4.5) and $c(x_{tj}; \theta)$ in (4.3). Therefore, our version of the WSPT rule performs as well as the other rules by implicitly assigning lower priorities to early jobs and vice versa.

The parameterizations in Section 4 were generated by conducting a preliminary analysis of the data, to determine which characteristics of the orders seemed to drive the historical scheduling decisions. Direct methods for evaluating the estimates of these parameterizations are the standard errors of the coefficients (which measure the precision of the estimates), and the J_T statistic (that checks how well the equations used in estimation can be reconciled to each other). In the parameterizations we studied, the coefficients are significant and the overidentifying restrictions of the model are not rejected. Now we have just shown in this section, how alternative scheduling rules for integrating early with late jobs orders can be valued, using the estimated preferences. Apart from these internal modes of validation, one can also take recourse to external ones, namely the managers themselves. Indeed, the operations analyst or consultant must ultimately convince the end user of his or her conclusions. But as mentioned in the introduction, there are few grounds to believe that managers can articulate their preferences in an abstract way, so presenting test statistics, while useful to the analyst, is unlikely to help in external validation.

We therefore propose augmenting the formal test statistics with informal measures, when presenting our findings to managers for direct attestation or refutation. Taken overall, comparisons based on these figures illustrate how one can externally verify the estimates and the performance of heuristics. If the managers prefer the distributions generated by heuristics associated with higher costs over those with lower costs, this might indicate that our cost function was misspecified, or that managerial priorities were not coordinated with scheduling on the shop floor. In particular, this could imply that the characteristics (urgency, customer size or order size) which our parameterizations use are either not relevant, or that the geometric tardiness function is misspecified.

Fig. 1 compares the distribution of lateness per day for the plant with that of WTCOVERT. (Recall that lateness is the difference between the delivery date of the order and its due date, and lateness per day is the average lateness of orders shipped on a day.) Over the sample period, the plant incurred an average lateness per day of 21.75 days (with a standard deviation of 20.66 days), whereas the heuristic generates an average of 13 days (with a standard deviation of 58 days). From these statistics it is clear that the heuristic reduces the overall average but increases the variance of lateness. The maximum lateness at the plant is 158 days, whereas the heuristic generates a maximum of 352 days. However, the minimum lateness per day at the plant was -69 days, compared to only -35 days for the heuristic.

Our findings reported in Section 4 showed that tardiness, urgency, customer size and footage significantly affect the revealed cost of tardiness to the plant. Accordingly we now present summary statistics on these dimensions, comparing the plant manager's (historical) decisions with the counterfactual generated by WTCOVERT. Fig. 2 juxtaposes the distribution of lateness per day generated historically against that generated by WTCOVERT, for orders

[29] The standard errors show in parentheses in Table 3 are obtained from these asymptotic results.

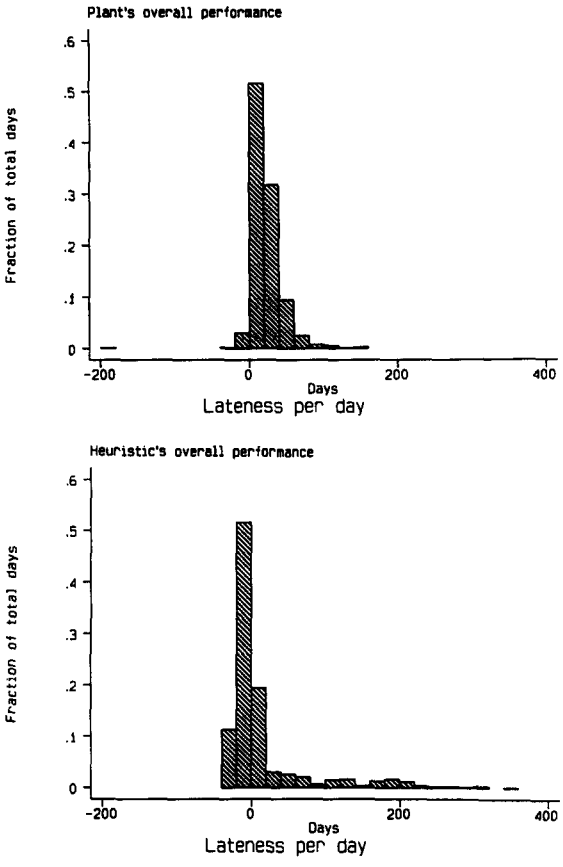

Fig. 1.

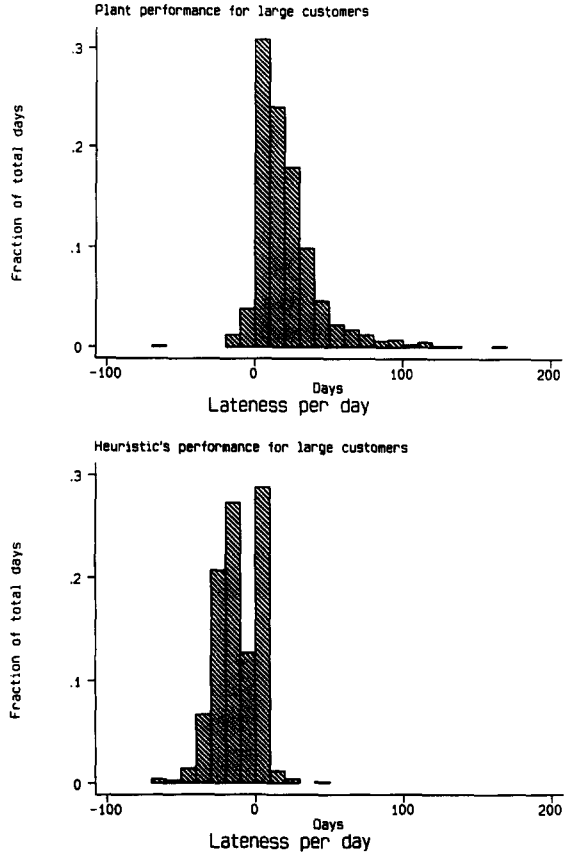

Fig. 2.

from customers within the upper 33rd percentile by total number of orders placed. [30] The figures show the variance in lateness is greater in the plant histories (with a standard deviation of 22 days) than those from the heuristic (14 days). In addition, the historical average for large customers is about 20 days late, whereas WTCOVERT generates an average of -13 days. The difference between the plant's performance and the heuristic's arises because the heuristic makes better tradeoffs between early and tardy jobs. Since the estimates place importance on customer size, the heuristic gives large customers a high priority (resulting in a low lateness performance).

Fig. 3 depicts the distribution of the lateness per day for less urgent orders (more precisely, the upper third of orders based on the urgency measure) for the plant and WTCOVERT. It shows that the heuristic schedules a larger number of these non-urgent orders (average lateness of 21 days and a standard deviation of 91 days) later than the plant (average lateness of 23 days and a standard deviation of 34 days). A similar comparison of urgent orders reveals that heuristic delivers an average lateness performance of 11 days (standard deviation of 55 days) compared to the historical average of 20 days (standard deviation of 21), thus demonstrating that WTCOVERT places a greater priority on urgent orders than the plant managers.

[30] In other words, the top 33 percent are the smallest set of customers who in total placed 33 percent of the total orders at the plant over the time period studied.

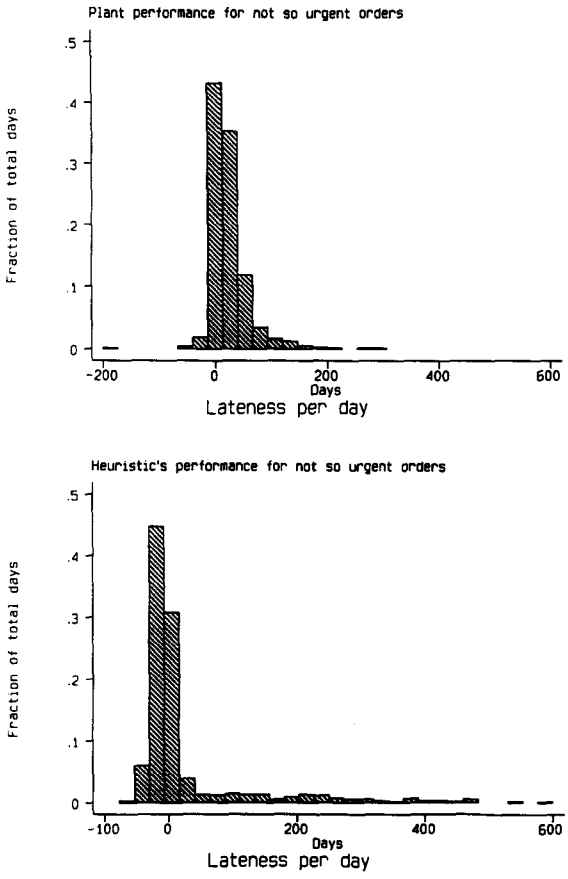

Fig. 3.

5.2. *Forecasting the effects of structural change*

The discussion so far has focused on analyzing alternative decision rules for a given shop floor and estimated set of preferences. We can equally as well investigate counterfactuals that change the underlying structural environment and hold the decision rule fixed. Essentially the same procedure as that described in Section 5.1 is used to forecast the effects of, say, changes in the probability distribution determining job order flow, or expansion in plant capacity. When analyzing capacity expansion, which is what we do below, there are two exercises one might undertake. First, suppose the plant knows the optimal rule, or will continue to follow the same priorities before and after the structural change occurs. In this case, historical data taken from the existing capacity can be used to estimate the priorities that guide decision making, and by relaxing the current constraint we can then forecast how the decisions and hence the overall performance would change (along any dimensions of interest). Second, suppose that in addition to asking for advice about a better rule to manage say, an operation, the plant is considering a capacity expansion there. Then our approach would be to estimate preferences as before, and then show how the candidate heuristic would have performed with the same historical order flow but different capacity constraints. Since we have not encoded the plant's decision rule with respect to early jobs, our application takes the second approach.

To illustrate how the effects of structural change can be forcast, we ran counterfactual histories to predict the effects of running the three rules, R&M, WTCOVERT and WSPT, on average costs and two measures of tardiness after undertaking a hypothetical increase in plant capacity of 5 percent. Here we simply scaled up daily effective capacity $1 - \xi_t$ by 5 percent for each day, although it would be just as straightforward to study an expansion that is not uniform across days. Then we recalculated how each of the decision rules would have processed the historical job order flow and compute the summary performance statistics, using exactly the same approach as above.

Our findings are displayed in the bottom half of Table 3. There are 2 main findings. Regardless of which rule is adopted, costs and tardiness decline quite substantially. This is an interesting result, because it shows that an equivalently modest increase in plant utilization would improve performance quite significantly. The other result is that, based on their respective performances, the ordering of the heuristics switches. Whereas WTCOVERT is more costly than the other two heuristics under the nominal settings, and is not significantly less tardy, if an expansion was undertaken, we predict that WTCOVERT's performance would dominate the others. This should not be too surprising, since none of the heuristics touted here are optimal, they share several common features (for example in the way tardy jobs are treated), and each of them has been justified in the literature. That such heuristics cannot be unambiguously ranked, but depend on the parameters characterizing the shop floor (and also managerial preferences), add further weight to our general argu-

ment for exploiting data on shop floor outcomes to draw a precise picture.

6. Conclusions

There are several implications that emerge from this study, some of which are applicable to industrial consulting, and others to operations management and production economics. It is convenient to discuss them under separate subheadings.

6.1. Managerial implications

The basic ingredient to our estimation strategy is using high-frequency data on inputs, outputs and production times (on flow through the various machines and their setups) for an operation that the plant manager knows how to solve. The manager should be confident that he is using the optimal rule, so that his preferences are reflected in the outcomes the data record. In order to execute this strategy, we ourselves also require the solution to this class of scheduling problems (perhaps by exploiting results in the operations research literature), stated in terms of parameters that at some unknown value (to be estimated) characterize the plant manager's preferences (which presumably reflect demand for the plant's products and its operating costs). With these ingredients we showed how to estimate the objective function the plant manager is optimizing.

The estimated objective function has 3 uses. First, the estimates can be used as a validation device. Confronted with the preferences revealed by his own actions, and various implications of those preferences, does the plant manager (and his boss) still feel comfortable with his decision making abilities? Second, the estimated preferences are helpful to the plant manager if he seeks guidance about a hard scheduling problem in another part of the plant. Using a second source of data on throughput in that part, we can evaluate the performance of competing heuristics, and thus advise him on which heuristic to adopt for the hard scheduling problem. Third, our approach provides a way of responding to requests by the plant manager for forecasts about how plant performance would change if either plant capacity or the job order flow changed.

Along the way, we explained why all the parameters of interest are not necessarily identified, depending on how well the estimated preferences relate to their envisaged uses. Thus preferences estimated about customers served by one product line might not help managers solve a scheduling problem about an entirely different product line say, at another firm facility. When some but not all the parameters are identified (as was the case in our empirical application), comparisons undertaken between competing heuristics (such as the difference in average costs) depend on the value(s) of the unknown parameter(s) assigned by the researcher. Finally, we provided statistical measures of the confidence and accuracy of the evaluations, and discussed how the results might be presented to managers in a digestible form.

Plant records used for this type of exercise are already routinely collected for other reasons, such as quality control and accounting requirements. Therefore, the practices we recommend do not require management to upgrade the plant's information systems. However, as our analysis has demonstrated, these records are not yet being utilized to their fullest extent. Perhaps the most important reason for this state of affairs is a lack of sophistication in data manipulation, and insufficient working familiarity with applied statistical inference at management levels. Given the widespread adoption of computer-based information systems, and the recent development of user-friendly statistical packages, we predict that this knowledge will be soon acquired by successful firms.

6.2. Future research

Our analysis suggests a new use for research in structural econometrics, as a tool for management science. Many heuristics and optimal rules for problems confronting management are characterized by management scientists and operations researchers in terms of features that describe not only the production set, but also preferences over multidimensional objectives. To make such rules operational, the parameters characterizing these preferences must be assigned values. We argued in the introduction that neither decision analysis nor protocol analysis are terribly satisfactory ways of eliciting these values.

Therefore, the heuristics and optimal rules derived in the operations management and operations research literatures are currently missing an important link necessary for their adoption in real world environments. The thrust of our argument is that the findings of structural estimation which exploits high frequency time series data on actual decisions are informative, and can sometimes provide that link. Undertaking the empirical exercise we advocate, exploits an important source of information about managerial preferences, in a systematic fashion. Not only do we recover cost functions that can be directly incorporated into algorithms based on managerial preferences; we also simultaneously estimate the precision of those estimates, and supply evidence about how well the framework is specified.

This paper concludes with some remarks about where future research might lead. An assumption maintained throughout this analysis is that the stochastic process characterizing the order flow is invariant to changes in decision rules and plant capacities. This simplification might be a reasonable approximation in the short term. Over the longer horizon, surely customers would greet more warmly improvements in service due to systematic change, than exactly the same service that arose by chance. In the former instance, but not the latter, one would predict a change in the pattern of future orders. As a first pass, one would expect those customers who were better off under the new regime to place more job orders and those who are worse off to place less (although even this apparently innocuous prediction is not invariably correct). How much the demand for each product lines shifts, and what price response this evokes from the firm, all contribute to determining the new equilibrium.

Beyond acknowledging the partial nature of our analysis, two directions are worth pursuing. Down one road is the challenge of determining when the qualitative features of our predictions continue to hold validity; in particular, under what circumstances do the signs of our cost comparisons remain unaffected by recognizing equilibrating demand shifts that accompany changes in scheduling rules? In the other direction is a more ambitious econometric analysis that exploits price data to estimate product demand directly, and hence incorporate scheduling changes within an equilibrium analysis of dynamic supply and demand. Both avenues are left for future research.

Appendix A

Proof of Proposition 1. We first note that to determine the optimal policy for jobs, it suffices to know the ranking of all possible pairs of jobs, for all possible dates. We now show that the optimal order for processing two tardy jobs is time invariant. In other words, if it is less costly to process job j at time t and job i at time $(t+1)$, than the other way around, then for all $s \in \{1, 2, 3\ldots\}$, it is also less costly to process j at time t and i at $(t+s)$, when compared with inverting the order. This implies that the ranking of all tardy jobs does not change over time. Thus any index that preserves this ranking forms the basis of an optimal policy for tardy jobs.

We establish the time invariance property separately for (2.7) and (2.8). Under (2.7) the difference between the cost of processing one unit of j on date t, one unit of i on $(t+s)$, and the reverse order (when both are tardy on date t) is

$$
\begin{aligned}
&c(x_{tj})/g(x_{tj}) + c(f^s(x_{ti}))/g(x_{ti}) \\
&\quad - c(f^s(x_{tj}))/g(x_{tj}) - c(x_{ti})/g(x_{ti}) \\
&\quad = \exp(c_0 x_{tj0} + c_1 x'_{tj}) \\
&\qquad + \exp(c_0(x_{ti0} + s) + c_1 x'_{ti}) \\
&\qquad - \exp(c_0(x_{tj0} + s) + c_1 x'_{tj}) \\
&\qquad - \exp(c_0 x_{ti0} + c_1 x'_{ti}) \\
&\quad = \exp(c_0 x_{tj0} + c_1 x'_{tj})(1 - \exp(c_0 s)) \\
&\qquad - \exp(c_0 x_{ti0} + c_1 x'_{ti})(1 - \exp(c_0 s)) \\
&\quad = \left[i(x_{tj}) - i(x_{ti})\right](1 - \exp(c_0 s)), \qquad \text{(A.1)}
\end{aligned}
$$

where the last line follows from (2.10). Noting that the sign of (A.1) does not depend on s, time invariance follows immediately.

Now if $i(x_{tj}) \geq i(x_{ti})$, it immediately follows from (A.1) that processing j before i is cheaper, that is, (A.1) is negative, since $\exp(c_0)$ the geometric penalty for tardiness per period is assumed greater than 1.

Now turning to the quadratic cost specification in (2.8), the cost difference between processing one unit of j on date t and one unit of i on $(t+s)$ and

vice versa, providing that both of them are tardy on date t is

$$
\begin{aligned}
&c(x_{tj})/g(x_{tj}) + c(f^s(x_{ti}))/g(x_{ti}) \\
&\quad - c(f^s(x_{tj}))/g(x_{tj}) - c(x_{ti})/g(x_{ti}) \\
&\quad = c_2 x_{tj0}^2 + x_{tj0} c_1 x'_{tj} \\
&\qquad + c_0 x'_{tj} \\
&\qquad + c_2 (x_{ti0} + s)^2 + (x_{ti0} + s) c_1 x'_{ti} + c_0 x'_{ti} \\
&\qquad - c_2 (x_{tj0} + s)^2 - (x_{tj0} + s) c_1 x'_{tj} - c_0 x'_{tj} \\
&\qquad - c_2 x_{ti0}^2 - x_{ti0} c_1 x'_{ti} \\
&\qquad - c_0 x'_{ti} \\
&\quad = s\{[x_{ti0}(2c_2) + c_1 x'_{ti} \\
&\qquad - [x_{tj0}(2c_2) - c_1 x'_{tj}\} \\
&\quad = s[i(x_{ti}) - i(x_{tj})], \qquad \text{(A.2)}
\end{aligned}
$$

where the last line of (A.2) follows from the definition of the index in (2.11). Following the same argument as above, time invariance follows in this case too. Furthermore, if $i(x_{tj}) \geqslant i(x_{ti})$, then it is cheaper to process j ahead of i, since the difference between doing j first, and vice versa shown in (A.2) will now be negative. □

Proof of Proposition 2. The proof draws extensively from Pakes and Pollard (1989), extending their results on inference about independent processes to those with finite dependence. To accomplish this, we introduce notation which is analogous to theirs and strengthens Lemmas (2.8) and (2.17) (on pages 1033 and 1037 of (Pakes and Pollard, 1989), respectively).

Accordingly, let P denote the probability associated with observing an observation y; let P_t denote the empirical measure on y, and write $\nu_t = T^{1/2}(P_t - P)$ for the standardized empirical process. Following Assumption 2 suppose y_t is independent of y_s for all $s < t - M$. For each $m \in \{1, \ldots, M\}$, define P_{tm} as the empirical measure associated with the subsample created by every Mth observation starting at m.

Lemma. *Let $\{h(y, \theta): \theta \in \Theta\}$ be an Euclidean class with envelope H for which $\int H^2 \, dP$ is finite. Then*

$$
\begin{aligned}
&\sup_{\theta \in \Theta} \left| \int h(y, \theta) \, dP_t(y) - \int h(y, \theta) \, dP(y) \right| \\
&\quad = o_{as}(1). \qquad \text{(A.3)}
\end{aligned}
$$

In addition, suppose that the parameterization is L^2 continuous at θ^0. Then for each sequence of positive numbers, $\{\delta_t\}$ converging to zero

$$
\sum_{\|\theta - \theta^o\| < \delta} |\nu_t h(y, \theta) - \nu_t h(y, \theta^o)| = o_p(1). \qquad \text{(A.4)}
$$

Proof. From the definition of P_{tm}:

$$
\begin{aligned}
&\sup_{\theta \in \Theta} \left| \int h(y, \theta) \, dP_t(y) - \int h(y, \theta) \, dP(y) \right| \\
&\quad = \sup_{\theta \in \Theta} \left| \sum_{m=1}^{M} \int h(y, \theta) \, dP_{tm}(y) \right. \\
&\qquad \left. - \int h(y, \theta) \, dP(y) \right| \\
&\quad \leqslant \sum_{m=1}^{M} \sup_{\theta \in \Theta} \left| \int h(y, \theta) \, dP_{tm}(y) \right. \\
&\qquad \left. - \int h(y, \theta) \, dP(y) \right|. \qquad \text{(A.5)}
\end{aligned}
$$

By Lemma (2.8) of Pakes and Pollard (1989, p. 1033)

$$
\begin{aligned}
&\sup_{\theta \in \Theta} \left| \int h(y, \theta) \, dP_{tm}(y) - \int h(y, \theta) \, dP(y) \right| \\
&\quad = o_{as}(1) \qquad \text{(A.6)}
\end{aligned}
$$

for all $m \in \{1, \ldots, M\}$. Since $Mo_{as}(1) = o_{as}(1)$, Eq. (A.3) follows immediately. Similarly, from the definition of ν_t,

$$
\begin{aligned}
&\sup_{\|\theta - \theta^o\| < \delta} |\nu_t h(y, \theta) - \nu_t h(y, \theta^o)| \\
&\quad = \sup_{\|\theta - \theta^o\| < \delta} \left| \sum_{m=1}^{M} \nu_{tm} h(y, \theta) - \nu_{tm} h(y, \theta^o) \right| \\
&\quad \leqslant \sum_{m=1}^{M} |\nu_{tm} h(y, \theta) - \nu_{tm} h(y, \theta^o)| \\
&\quad = Mo_p(1) \\
&\quad = o_p(1). \qquad \text{(A.7)}
\end{aligned}
$$

(The second from the last line in (A.7) follows from Lemma (2.17) from Pakes and Pollard (1989) on p. 1037 for independent processes.) This establishes (A.4). □

The consistency of $\tilde{\theta}$ is established by verifying that the conditions of Corollary (3.2) of Pakes and Pollard (1989, p. 1039) are met. Since $\tilde{\theta}$ minimizes (3.12) the first condition is satisfied by definition. Their second condition is an assumption about identification, which we also make. Finally, their third condition is directly implied by our Eq. (A.3). Using the fact that $\tilde{\theta}$ converges in probability to θ^o, and applying the ergodic theorem, it now follows that $\tilde{\Sigma}$, defined in Eq. (3.13) converges in probability to Σ, and $\tilde{Q}$ converges to Q. Hence, the conditions in Lemma (3.4) are satisfied, allowing us to conclude that $\hat{\theta}$ converges to θ^o in probability as claimed.

The asymptotic distributional properties of $\hat{\theta}$ are established in a similar way. In this case we verify that the five conditions in Theorem (3.3) of Pakes and Pollard (1989, p. 1040), and the conditions in their Lemma (3.4) (p. 1044), is met. Again the first condition of Theorem (3.3) follows directly from our definition of $\hat{\theta}$. Likewise, their second condition is implied by Assumption 1 and that $E[h(y, \theta)]$ is differentiable. Eq. (A.4) is stronger than their third condition, while the martingale central limit theorem meets their condition (iv). Our Assumption 1 implies their condition (v). Finally, the condition in their Lemma (3.5) is satisfied because $\tilde{Q}$ converges to Q in probability and $\tilde{\Sigma}$ converges to Σ in probability. □

References

Aguirregabiria, V., 1994. Dynamic models of price and inventory decisions with fixed adjustment costs, PhD thesis, Universidad Complutense de Madrid.

Akaike, H., 1973. "Information theory and an extension of the maximum likelihood principle", In: B. Petrov and F. Csaki (Eds.), Second International Symposium on Information Theory, Akademia Kiado, Budapest, pp. 267–281.

Andrews, D., 1991. "Heteroskedasticity and autocorrelation consistent covariance matrix estimation", Econometrica, vol. 59, pp. 817–858.

Baker, K., 1979. Introduction to Sequencing and Scheduling. Wiley, New York.

Bertrand, J.J., Wortmann and J. Wijngaard, 1990. Production Control: A Structural and Design Oriented Approach, Elsevier, Amsterdam.

Eckstein, Z. and K. Wolpin, 1989. "The specification and estimation of dynamic stochastic discrete choice models", J. Human Resources, vol. 24, pp. 561–597.

Ericsson, K.A. and H.A. Simon, 1984. Protocol Analysis: Verbal Reports as Data, MIT Press, Cambridge, MA.

Hansen, L., 1982. "Large sample properties of generalized method of moments estimators", Econometrica, vol. 50, pp. 1029–1054.

Hogarth, R.M., 1987. Judgement and Choice: The Psychology of Decision, Wiley, New York.

Hotz, V. and R. Miller, 1993. "Conditional choice probabilities and the estimation of dynamic discrete choice models", Rev. Economic Studies, vol. 60, pp. 497–529.

Hotz, V., R. Miller, S. Sanders and J. Smith, 1994. "A simulation estimator for dynamic discrete choice models", Rev. Economic Studies, vol. 61, pp. 265–289.

Keeney, R.L. and H. Raiffa, 1976. Decisions with Multiple Objectives: Preferences and Value Tradeoffs, Wiley, New York.

McFadden, D., 1989. "A method of simulated moments for estimation of discrete response models without numerical integration", Econometrica, vol. 57, pp. 995–1026.

Miller, R., 1984, "Job matching and occupational choice", J. Political Economy, vol. 92, pp. 1086–1120.

Miller, R. and P. Ramnath, 1994a. Splicing histories to form conditional choice probability estimators, Working paper, Graduate School of Industrial Administration, Carnegie Mellon University.

Miller, R. and P. Ramnath, 1994b. Structural econometrics for management consultants, Working paper, Graduate School of Industrial Administration, Carnegie Mellon University.

Morton, T., 1993. Heuristic Scheduling Systems: With Application to Production Systems and Project Management, Wiley, New York.

Newell, A. and H.A. Simon, 1972, Human Problem Solving, Prentice Hall, Englewood Cliffs, NJ.

Pakes, A., 1986. "Patents as options: Some estimates of the value of holding European patent stocks", Econometrica, vol. 54, pp. 755–784.

Pakes, A. and D. Pollard, 1989. "Simulation and the asymptotics of optimizations estimators", Econometrica, vol. 5, pp. 1027–1057.

Rust, J., 1987. "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher", Econometrica, vol. 55, pp. 999–1033.

Rust, J., 1994. "Structural estimation of markov decision processes", In: D. McFadden and R. Engle (Eds.), Handbook of Econometrics (forthcoming), volume IV, North-Holland, Amsterdam.

Sanders, S., 1994. "A dynamic model of welfare and work, Working paper, Carnegie Mellon University.

Slade, M., 1994. Optimal pricing with costly adjustment and persistence effects: Empirical evidence, Working paper, University of British Columbia.

Wolpin, K., 1994. "An estimable dynamic stochastic model of fertility and child mortality", J. of Political Economy, vol. 92, pp. 852–874.

Zeleny, M., 1976. Multiple criteria decision making. Springer, New York.