MANAGEMENT SCIENCE
Vol. 33, No. 8, August 1987
Printed in U.S.A.

PRIORITY RULES FOR JOB SHOPS WITH WEIGHTED TARDINESS COSTS*

ARI P. J. VEPSALAINEN AND THOMAS E. MORTON
Department of Decision Sciences, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104
Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213

Mainstream research in priority dispatching has considered jobs with equal delay penalties, thereby ruling out strategic differentiation of customer orders. We develop and test efficient dispatching rules for the weighted tardiness problem with job-specific due dates and delay penalties. Our approach builds on previous greedy heuristics which assign the priority on the basis of the expected tardiness cost per immediate processing requirements. In multi-machine applications, estimates of the remaining leadtime are needed to determine local operation due dates and to evaluate the adequacy of the job's slack. Two slightly different "look-ahead" features are identified, and the corresponding priority rules are tested in job shop experiments with a variety of load conditions. The results indicate that the new rules are not only superior to competing rules for minimizing weighted tardiness penalties but are also robust for several other criteria, such as the number of tardy jobs and the costs of in-process inventories.
(PRODUCTION/SCHEDULING; JOB SHOP; DETERMINISTIC)

1. Introduction

In many applications, meeting due dates and avoiding delay penalties are the most important goals of scheduling. The costs of tardy deliveries, such as customer badwill, lost future sales, and rush shipping costs, vary significantly over customers and orders, and the implied "strategic weight" should be reflected in job priority. But instead, traditional research in priority scheduling has assumed all jobs to carry equal weights (as in average tardiness and number of tardy jobs), or studied other objectives (most often average flow time or makespan), see the RAND studies (Conway et al. 1967), and Muth and Thompson (1963), Panwalkar and Iskander (1977), Graves (1981), Baker (1974, 1984). Moreover, most studies have tested simple rules designed for some extreme shop conditions and known to be deficient with certain load levels. For example, Earliest Due Date (EDD), Minimum Slack (MSLACK), and Slack per Remaining Processing Time (S/RPT) rules perform reasonably with light load levels but deteriorate in congested shops, whereas the mainstay of priority scheduling, the Shortest Processing Time (SPT) rule, fails with light load levels and generous due date allowances (Elvers and Taube 1983). There have been attempts to combine the effects of slack (MSLACK, S/RPT) and machine utilization (SPT) into a single rule. Despite demonstrated performance in test problems, however, such rules as the Dynamic Composite Rule (Conway et al. 1967) and COVERT (Carroll 1965) have been shunned by scholars because of reluctance to consider parameterized rules. Just recently has a nonparameterized priority rule, the Modified Operation Due date (MOD (Baker and Bertrand, 1982)), emerged as satisfactory for many average (nonweighted) tardiness problems.

In approaching the more complex weighted-tardiness problem, we analyze dispatching rules that are greedy heuristics: priority is given to the job with highest expected delay cost per imminent machine processing time. For a job with some slack, the expected tardiness cost is reduced according to a "look-ahead" feature that can be

* Accepted by Leroy B. Schwarz; received December 1985. This paper has been with the authors 4 months for 2 revisions.

0025-1909/87/3308/1035$01.25

formalized in two different ways: (a) the remaining time until due date can be compared with the expected total waiting time, as in the COVERT rule (Carroll 1965) or (b) the global slack can be allocated for the remaining operations by establishing local operation due dates (Baker 1984). We apply both of these principles, the first by making a "weighted" version of the COVERT rule, and the second by extending a myopic single-machine rule for weighted-tardiness problems, due to Rachamadugu and Morton (1981), to multi-stage problems. The operation due dates of this rule, here called the *Apparent Tardiness Cost* (ATC) rule, are determined by the expected subsequent leadtimes.

In the following, we discuss first the information used in the priority rules, especially the "weighted" COVERT and our Apparent Tardiness Cost rules. In §3, the weighted-tardiness performance of the different priority rules is reported, thereby complementing the previous studies (Baker 1974, 1984, Conway et al. 1967, Panwalkar and Iskander 1977) with this important criterion. The results of a large-scale computational experiment in general dynamic job shops indicate that the new ATC rule dominates the competing rules in all shop load conditions studied. This is true not only with respect to the average weighted-tardiness criterion but often also for the average portion of tardy jobs. Weighted COVERT is overall second but is not as consistent in terms of number of tardy jobs. Both rules also exhibit good performance in terms of work-in-process and inventory holding costs. These results, achieved with predetermined parameter values for look-ahead and leadtime estimation, can often be improved by adjusting the parameters on the basis of shop load.

Conclusions and extensions are presented in §4.

2. Priority Rules for Weighted-Tardiness Scheduling

2.1. *The Information Used in Local Rules*

Consider scheduling a machine-constrained job shop[1] with m machines and n jobs. Job i has m_i operations in a predetermined sequence on the machines with deterministic processing times, $p_{ij}, j = 1, \ldots, m_i$. There is a delay penalty, or weight, of v_i per unit time, charged if job i is completed after its due date d_i. This penalty, assumed to be constant over time, includes customer badwill, cost of lost sales or changed orders, and rush shipping cost. The objective is to minimize the weighted tardiness of the jobs:[2]

$$WT = \sum_{i=1}^{n} v_i[C_i - d_i]^+, \tag{1}$$

where C_i stands for the completion time of job i:

$$C_i = r_i + \sum_{j=1}^{m_i} (W_{ij} + p_{ij}). \tag{2}$$

Here r_i is the arrival (or release) time, and W_{ij} is the waiting time at operation j.

Most dispatching rules are local, i.e., their priority index is a function of some subset of the job's attributes given above and the current time, t. Representative rules and their priority indexes are shown in Table 1. First Come-First Served (FCFS) is a commonly used benchmark rule which uses the realized completion time of the previous operation, C_{ij-1}, where $C_{i,0} = r_i$, as the priority index of job i waiting for its jth operation.

[1] The standard assumptions apply (see Conway et al. 1967; Baker 1974): no pre-emption, overlapping of operations, or alternative routings allowed, set-up times are independent of sequence and included in the processing times, instantaneous transfers, no rework or simultaneous operations, and machines are available continuously.

[2] Here we use notation $[x]^+ = \max\{0, x\}$.

FCFS is easy to implement, but such a "backward-looking" approach is detrimental for most performance criteria. The Weighted Shortest Processing Time first (WSPT) rule, using the "natural" priority of job i, v_i/p_{ij}, or the penalty avoided per unit of machine time, works analogously to the SPT rule: overall tardiness is reduced in congested shops by giving priority to short jobs, and the weights v_i help in coordinating job priorities across machines. By delaying some long jobs, WSPT can also achieve a remarkably low total number of tardy jobs without using explicit due date information, especially when job earliness is limited by dynamic release dates.

The Earliest Due Date first (EDD) rule emphasizes job urgency by using the global due date d_i, or equivalently the allowance $A_i(t) = d_i - t$, as priority index. The Minimum Slack (MSLACK) index, $S_{ij}(t) = d_i - \Sigma_{q=j}^{m_i} p_{iq} - t$, applies additional processing time information but in a way that actually counteracts the shortest first (SPT) economies: of several jobs with equal allowance, priority is always given to the longest one! No wonder schedulers have experimented instead with the Critical Ratio rule, or allowance per expected leadtime (Berry and Rao 1975), and other similar rules such as Slack per Remaining Processing Time (S/RPT, Baker 1974, Buffa and Miller 1979). These ratio rules compensate for the "anti-SPT" tendency of a plain slack rule; when the allowance turns negative, shorter jobs will get higher priority. This switch from the "slack-mode" to the "SPT-mode", viewed as an anomaly by some (Adam and Surkis 1980), in fact realigns priorities eventually even if the shorter job was due later. In the same spirit, Baker and Bertrand (1982) introduced a dynamic "Modified Due Date" rule (MDD, a single-machine version of the MOD rule) which adjusts the due date to the earliest possible completion date once the job becomes critical:

$$\text{MDD} = \begin{cases} d_i & \text{if} \quad S_i > 0, \\ t + p_i & \text{otherwise.} \end{cases} \tag{3}$$

The job with the earliest modified due date goes first. The MDD and MOD rules are closely related to our new rule (see §2.3 below), and very efficient for nonweighted tardiness scheduling in comparison to many traditional rules (Baker 1984).

Conway et al. (1967) experimented with more explicit combinations of SPT and

TABLE 1

Alternative Dispatching Rules and Priority Indexes for Job i on Operation j of m_i Operations, at Time t.

Rule		Rank	Priority Index
FCFS	First Come-First Served	min	$C_{i,j-1} = r_i + \Sigma_{q=1}^{j-1} (W_{iq} + p_{iq})$
EDD	Earliest Due Date	min	d_i
S/RPT	Slack per Remaining Processing Time	min	$\dfrac{(d_i - t - \Sigma_{q=j}^{m_i} p_{iq})}{\Sigma_{q=j}^{m_i} p_{iq}}$
WSPT	Weighted Shortest Processing Time	max	$\dfrac{v_i}{p_{ij}}$
COVERT	Weighted COVERT	max	$\dfrac{v_i}{p_{ij}}\left[1 - \dfrac{(d_i - t - \Sigma_{q=j}^{m_i} p_{iq})^+}{k\, \Sigma_{q=1}^{m_i} W_{iq}}\right]^+$
ATC	Apparent Tardiness Cost*	max	$\dfrac{v_i}{p_{ij}} \exp\left(-\left[\dfrac{d_i - t - p_{ij} - \Sigma_{q=j+1}^{m_i} (W_{iq} + p_{iq})}{kp}\right]^+\right)$

* ATC rule complete for multiple operations.

Slack per Operation (S/OPN) indexes, with equal weights, and with the "Dynamic Composite Rule" (DCR), a weighted sum of the slack from an operation due date, the immediate processing time, and the work in the current queue and job's next queue. These rules reduced the average tardiness and the portion of tardy jobs over the component rules, but their application in varying shop conditions were not recommended because of the sensitivity to misspecification of the terms and weights of the priority index.

2.2. *The Weighted COVERT Rule*

Carroll (1965) designed a dynamic rule for average tardiness scheduling to be used to incorporate job weights into a slack-based approach. The COVERT priority index represents the expected tardiness cost per unit of imminent processing time, or Cost OVER Time. Job i queuing for operation j with zero or negative slack, $S_{ij}(t) < 0$, is projected to be tardy by completion with an expected tardiness cost of v_i (originally normalized to one) and priority index v_i/p_{ij}. If, on the other hand, the slack exceeds some generous "worst case" estimate of the waiting time, the expected tardiness cost is set to zero. The worst case waiting time serves as a reference for a piecewise-linear "look-ahead", or a mapping of the expected tardiness cost over the slack S_{ij}:

$$\mathrm{COVERT}_{ij}(t) = \frac{v_i}{p_{ij}} \frac{[k \sum_{q=j}^{m_i} \mathrm{W}_{iq} - (S_{ij})^+]^+}{k \sum_{q=j}^{m_i} W_{iq}}. \tag{4}$$

Here W_{iq} is the expected waiting time for a remaining operation q, and k is a multiplier adjusting the expected waiting time to the worst case, say the 99% limit of the cumulative probability distribution. The remaining waiting time is often a sum of several operation waiting times, and the symmetry of that distribution suggests worst case about twice the mean, or length $k = 2$ of look-ahead in equation (4) above. The COVERT priority function is shown in Figure 1.

The original results proved COVERT superior to the competing rules, including a truncated SPT, in the mean tardiness performance (Carroll 1965).

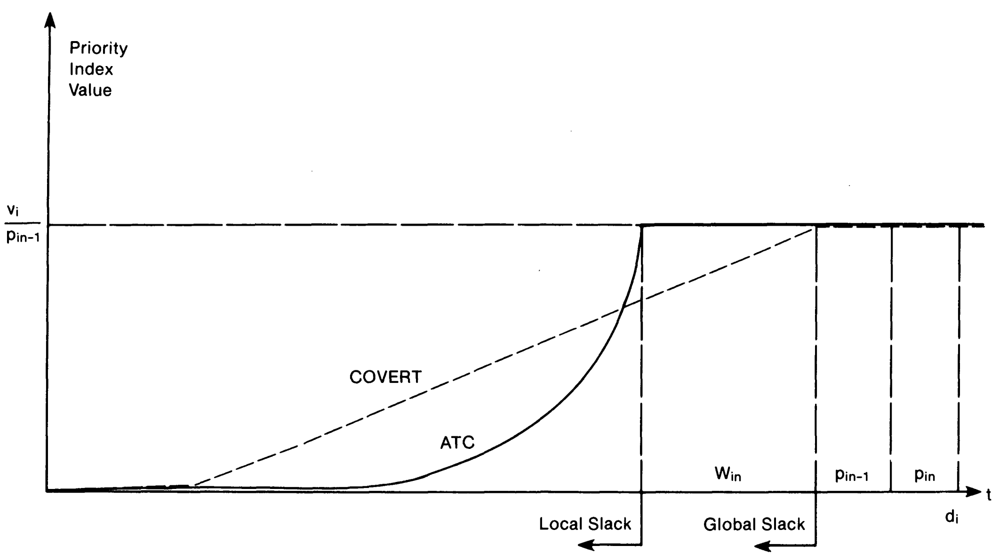

FIGURE 1. An Illustration of the Priority Indexes for Job i with Two Operations Remaining: COVERT (Dashed Line) and Apparent Tardiness Cost (Solid Line).

2.3. *The Apparent Tardiness Cost Rule*

Rachamadugu and Morton (1981) developed a look-ahead rule for the weighted-tardiness problem. Assuming an optimal schedule, initially for a single machine, they used a pairwise interchange to show that for job i to go before job j at time t it is necessary that the marginal costs satisfy the following condition:

$$v_i[t + p_i - d_i]^+ + v_j[t + p_i + p_j - d_j]^+ < v_i[t + p_j + p_i - d_i]^+ + v_j[t + p_j - d_j]^+. \quad (5)$$

To develop a priority index, it is convenient to rewrite (5) in an equivalent form:[3]

$$\frac{v_i}{p_i}\frac{[p_j - (d_i - p_i - t)^+]^+}{p_j} > \frac{v_j}{p_j}\frac{[p_i - (d_j - p_j - t)^+]^+}{p_i}. \quad (6)$$

Now the index has a familiar form: a maximum of v_i/p_i, or the WSPT index, for a critical job and less if there is slack. But instead of trading off the slack of job i against the processing time of job j, and vice versa, a standard reference should be used. A piecewise linear look-ahead was first suggested by replacing the unknown p_j in job i's index (implied by the left-hand side of equation (6)) by a factor kp, or the average processing time of the waiting jobs, p, multiplied by a "look-ahead parameter" k related to the number of competing critical and near-critical jobs. But instead of a linear function, an inverse of allowance is actually closer to the "apparent cost" of tardiness implied by the break-even priority of tardy jobs with processing times exceeding the slack (Vepsalainen 1984). In computational tests an exponential function of the slack was found to be somewhat more efficient[4] (Rachamadugu and Morton 1981). The priority index of the rule shown in Figure 1, here called the Apparent Tardiness Cost (ATC), is for single-stage as follows:

$$\mathrm{ATC}(t) = \frac{v_i}{p_i}\exp\left(-\frac{[d_i - p_i - t]^+}{kp}\right). \quad (7)$$

Again, k is a look-ahead parameter that scales the slack (measured in units of average processing time[5]) according to the expected number of competing jobs. A fixed value $k = 2$ has been used in static flow shops, including one-machine studies (Rachamadugu 1982, Rachamadugu and Morton 1981, Vepsalainen 1984, Vepsalainen and Morton 1985), and $k = 3$ is a reasonable "average" for dynamic job shops. This local look-ahead can be adjusted on the basis of expected number of competing jobs (usually within $1.5 < k < 4.5$) to reduce weighted tardiness costs in extremely slack or congested shops (Ow 1985, Rachamadugu 1982), whereas nonparameterized rules, such as Critical Ratio and MOD, are inherently myopic.

The ATC rule is superior to other sequencing heuristics and close to optimal for the weighted-tardiness performance in single machine problems (Rachamadugu and Morton 1981). Similarly to the MOD rule, it trades off job's urgency (slack) against machine utilization, but due to the more complex weighted criterion, an additional look-ahead parameter is needed to assimilate the competing jobs which may have different weights. Another issue evident in multi-stage shops is how to coordinate the local rule with appropriate operation due dates, i.e., how to derive the complete multi-operation ATC rule in Table 1.

[3] With equal weights, condition (5) reduces to the MDD rule discussed above.

[4] Intuitively, the exponential look-ahead works by ensuring timely completion of short jobs (steep increase of priority close to due date), and by extending the look-ahead far enough to prevent long tardy jobs from overshadowing clusters of shorter jobs (Vepsalainen 1984).

[5] In simulation, a long-run average, or a standard processing time, can be used instead of the actual queue mean.

2.4. *Coordinating Local Rules with Operation Due Dates*

When jobs have multiple operations, the performance of due date (and slack) oriented rules can be improved by assigning local due dates. If, for instance, the initial flow allowance is allocated to operation leadtimes in proportion to their processing times to assign local due dates, then the use of the Operaton Due Date (ODD) rule can reduce mean tardiness in comparison to the global EDD discipline (Baker 1984). This method of setting operation due dates is widely used in practice but obviously suspect for handling rush orders with high delay penalties but late release dates. A more practical approach has been suggested (Vepsalainen 1984, Vepsalainen and Morton 1985), whereby operation due dates could be obtained from the job due date by subtracting some reasonable estimates of the expected leadtimes on the subsequent machines:

operation due date = job due date − estimate of subsequent leadtime or

$$d_{ij} = d_i - \sum_{q=j+1}^{m_i} (W_{iq} + p_{iq}). \tag{8}$$

These efficient operation due dates now depend explicitly on the expected waiting times, W_{ij}, but not on the release date. Accordingly, in the multi-operation ATC index (Table 1), the global slack is first allocated to the remaining leadtime. The local "resource-constrained" slack, measured from the established operation due date, is then compared to the competing work to evaluate the "apparent" tardiness cost.

2.5. *Leadtime Estimation for Priority Rules*

Leadtime predictions and standards are often used in communication with customers, but also as a part of priority rules, as discussed above: either the leadtimes estimates are compared to the remaining slack (as in the COVERT and Critical Ratio rules), or they are used to find reasonable operation due dates (as with the ODD, MOD, and ATC rules). Conventional studies and scheduling systems have relied on the *ad hoc* estimates based on prorated initial flow allowance, or on some accounting leadtimes based on past observations or desired values; these standards are then used in all load situations. In simulation studies (Carroll 1965, Conway et al. 1967), it was found convenient to estimate the waiting time of job i at operation j, W_{ij}, as a multiple of the corresponding processing time p_{ij}:

$$W_{ij} = bp_{ij}. \tag{9}$$

The leadtime estimation parameter b should reflect the anticipated machine utilization. We will test these standard estimates, used in many scheduling applications and Materials Requirements Planning systems as well, with the new rules. Other leadtime estimation methods exist, however, with great promise (Vepsalainen 1984, Vepsalainen and Morton 1985).

3. A Computational Study

3.1. *Job Shop Problems*

In this set of experiments (Vepsalainen 1984, Vepsalainen and Morton 1985), we simulate a job shop with 10 machines, jobs arriving continuously with interarrival times taken from a truncated Poisson distribution. The jobs have one to 10 operations with a randomly assigned routing through the machines, and the process is observed until all of 2,000 jobs have been completed. Three kinds of processing time distributions are tested. In a uniform shop, the sizes of the jobs are assumed constant $s = 15$, and the processing times are drawn from a uniform distribution $p \sim U[1, 30]$. In a proportionate shop, the jobs are first assigned size $s_i \sim U[5, 25]$, and the processing times are generated from $p_i \sim U[0.33*s_i, 1.67*s_i]$. Thus the processing times of a job are correlated, or almost proportionate, over the operations. The third kind of shop is a uniform shop with bottleneck machines; the relative speeds of three machines are up to 30% faster than average, and three up to 20% slower.

Job weights are drawn from $v_i \sim U[1, 2*s_i]$. The load of the shop is determined by the arrival rate to yield five approximate utilization levels: 80%, 85%, 90%, 95% and 97% of the bottleneck capacity. Due dates are assigned randomly over a full range of flow allowances, with an average of six (three) times the mean job processing time for relatively loose (tight) due date setting. The full factorial design consists of 3 shop and load types; 2 due date settings; and 5 utilization levels, or 30 different problems with 2,000 jobs each.

3.2. *Measures of Performance*

The primary objective is to minimize weighted tardiness. Since the numerical measures of weighted tardiness depend on problem size (number of jobs and machines) and the distribution of processing times and delay penalties, we have found it useful to normalize the results:

$$\text{Normalized } WT = \frac{WT}{nmpv} \tag{10}$$

where WT is given by equation (1), n = number of jobs, m = average number of operations per job, p = average processing time of an operation, and v = average delay penalty per unit time. The normalized results allow intuitive interpretation and comparisons across studies[6] without having to rescale the "live" data provided by managers.

In practice, there are often other goals equally important for scheduling applications. We have recorded several secondary criteria to examine the robustness of the individual rules and to demonstrate the trade-offs of rule selection. Number of tardy jobs is an important measure of service level, normalized to portion of tardy jobs, or percentage tardy.

Maintaining low in-process inventory is also a traditional yardstick for scheduling rules. We record two different measures: Work-In-Process (WIP) is the cost of holding the jobs from the start of the first operation, a_i, to the completion of the last operation, C_i:

$$\text{WIP} = \frac{\sum_{i=1}^{n} s_i(C_i - a_i)}{nmps}. \tag{11}$$

The holding cost is determined by the size, s_i, correlated with job i's work and value content. The WIP measure is normalized using the total processing time multiplied by the average holding cost, s. In the interest of reducing the variance of completion times from the due dates and not rewarding earliness, we record also the Work-In-System (WIS) holding cost:

$$\text{WIS} = \frac{\sum_{i=1}^{n} s_i(\max\{C_i, d_i\} - a_i)}{nmps}. \tag{12}$$

WIS charges inventory until the due date, d_i, assuming that the customer would not accept early shipment due to "just in time" materials management.

3.3. *Priority Rules Tested*

The problems are simulated with the priority rules shown in Table 1. The rules incorporate varying degrees of available information: the FCFS rule uses one of a job's

[6] Only approximate comparisons between problems of different size and number of operations are possible (Vepsalainen 1984)—appropriate adjustments for these and other attributes (distribution of flow allowances, weights, etc.) can be made by experience.

attributes, the arrival date, indirectly; the EDD index consists of the due date; the S/RPT rule also includes processing time and current time information. The WSPT rule, on the other hand, includes only weight and processing requirements. The weighted COVERT rule combines the effects of WSPT and S/RPT, plus expected leadtime information (similar "retrofitting" of the other nonweighted rules, such as DCR and MOD, is not possible). Our new ATC rule (the multi-operation version of the myopic Rachamadugu-Morton rule in equation (7)) derives a job's priority from comparison with its contenders, coordinating local tradeoffs on the basis of operation due dates defined in equation (8).

This set of rules allows us to study the impact of the additional information upon the efficiency and robustness of dispatching. Although the quality of leadtime estimates can also affect performance, we test the ATC and COVERT rules only with the standard estimates given in equation (9).

The slack evaluation parameters b and k are now fixed throughout the experiment. The waiting time estimates W_{ij} are generated with $b = 2.0$ in equation (9) for all shops and load conditions. A look-ahead parameter $k = 2$ is used for COVERT as discussed above. For the ATC rule, look-ahead is extended to $k = 3.0$ (from $k = 2.0$ in flow shops (Rachamadugu 1982, Vepsalainen 1984), partly to compensate for longer average queue lengths in job shops.

These parameters could be adjusted *a priori* on the basis of queue length estimates, for instance.[7] Such interpretation of rule parameters as coordination variables has not been given for the previous combination rules but their weights have been determined by trial and error (Baker 1974, Conway et al. 1967, Elvers and Taube 1983). Since both b and k affect the length of the total look-ahead, they need to be mutually adjusted: in varying load, b can be underestimated, on average, to reduce realized leadtimes, if this bias is then compensated for by slightly overestimating k to smooth out priorities. Hence the slight inconsistency of the parameter values above, i.e., $k > b$, verified through extensive testing.

3.4. *Results and Discussion*

The weighted tardiness results are shown in Table 2 for the tight and slack due date settings separately. Here the results for the three different shop layouts (uniform, proportionate, and bottleneck) are averaged because of very similar response. From these results, graphed in Figure 2, it is obvious that the rules with more information outperform the simpler ones: COVERT is as good as the best of the simple rules, and often much better, and the new ATC rule outperforms even COVERT consistently.

WSPT is an overall third, but at 40–240% higher weighted-tardiness costs with tight due dates (and 230–650% higher with loose due dates) the simplicity of WSPT seems perhaps bought at a high price. The EDD and S/RPT rules are reasonable only with easy due dates and low utilization, failing when shop load exceeds 90%. S/RPT dominates EDD, outperforming even the standard COVERT in light load with loose due dates. FCFS comes distant last, with at least seven times the tardiness cost of the new rules.

The portion of tardy jobs, shown in Table 2 and also in the graphs of Figure 3, exhibits qualitatively similar behavior: ATC is clearly superior over the standard rules, with COVERT second. WSPT is close, however, crossing over COVERT at 95% with loose due dates and at 85% with tight due dates, and beating even ATC at 97%.[8] As to the other rules, the results agree with nonweighted scheduling (Elvers and Taube 1983),

[7] Systematic fine tuning of the parameters yields minor but consistent improvements of weighted tardiness (usually less than 3% on average, see Vepsalainen 1984).

[8] With $k = 4$, however, ATC outperforms WSPT also in congested shops (Vepsalainen 1984, Vepsalainen and Morton 1985).

TABLE 2

Comparison of the Rules in Terms of Normalized Weighted Tardiness (Portion of Tardy Jobs % in Parentheses).

(a) Loose due date allowances
Normalized Weighted Tardiness (Portion of Tardy Jobs):

	Approximate Load: 80%	85%	90%	95%	97%
Rule:					
FCFS	0.278 (16.5)	0.546 (23.0)	1.173 (34.3)	2.692 (51.2)	3.390 (52.3)
EDD	0.022 (5.1)	0.073 (8.7)	0.197 (20.4)	1.222 (45.3)	1.899 (51.0)
S/RPT	0.018 (5.3)	0.034 (9.6)	0.078 (19.7)	0.919 (48.6)	1.503 (54.2)
WSPT	0.110 (12.7)	0.208 (16.1)	0.348 (20.1)	0.617 (24.3)	0.710 (24.1)
COVERT	0.018 (6.7)	0.030 (8.9)	0.056 (13.4)	0.199 (23.1)	0.294 (25.7)
ATC	0.016 (4.4)	0.029 (6.5)	0.046 (8.6)	0.191 (17.9)	0.291 (20.3)

(b) Tight due date allowances
Normalized Weighted Tardiness (Portion of Tardy Jobs):

	Approximate Load: 80%	85%	90%	95%	97%
Rule:					
FCFS	0.753 (36.8)	0.922 (42.1)	2.329 (60.9)	3.033 (64.5)	5.984 (76.4)
EDD	0.354 (31.4)	0.444 (38.2)	1.662 (67.2)	2.360 (69.7)	4.465 (80.5)
S/RPT	0.269 (36.1)	0.338 (43.0)	1.586 (74.5)	2.062 (74.9)	4.063 (85.1)
WSPT	0.247 (24.3)	0.296 (26.3)	0.556 (32.1)	0.666 (34.1)	1.079 (36.8)
COVERT	0.106 (22.1)	0.121 (24.6)	0.340 (36.3)	0.432 (38.7)	0.777 (44.3)
ATC	0.103 (17.0)	0.121 (20.3)	0.332 (30.2)	0.419 (33.0)	0.765 (37.1)

except that EDD is better than S/RPT, and that FCFS outperforms both of these rules in congested shops and with short due date allowance.

Of the other secondary criteria, the WIP inventory holding costs shown in Table 3 exhibit somewhat unexpected behavior with increasing load; the EDD and ATC rules cross over WSPT in quite low shop utilization (85% to 90%), followed by S/RPT and COVERT. The ATC rule outperforms COVERT again in all problem classes.

The WIS costs are similar for all due date oriented rules, S/RPT and EDD leading with some margin the ATC and COVERT rules (especially in high load). The WSPT priorities, not aligned with the due dates, yield WIS costs even higher than FCFS.

Finally, the economic performance of the rules can be summarized in terms of the combined costs of tardy orders, inventory holding and rush shipping (Vepsalainen 1984). The ATC rule is strong across the criteria, and its composite costs, averaged over the three shop types, are consistently the lowest. For realistic range of relative cost parameters, the rules rank: ATC–COVERT–WSPT–S/RPT–EDD–FCFS, with the exception that S/RPT and EDD cross over WSPT in shops with slack due dates and low utilization, and EDD sometimes beats COVERT in very light load.

4. Summary and Conclusions

The strategic costs of late orders have not been considered in the legion of previous studies of priority scheduling. We have modelled these costs using a weighted-tardiness criterion, and worked out suitable dispatching rules. The new rules, Apparent Tardiness Cost (ATC) and weighted COVERT, not only subsume many existing rules but extend them with additional information by "looking ahead"; the same expected leadtimes can serve as reference for slack evaluation and operation due dates.

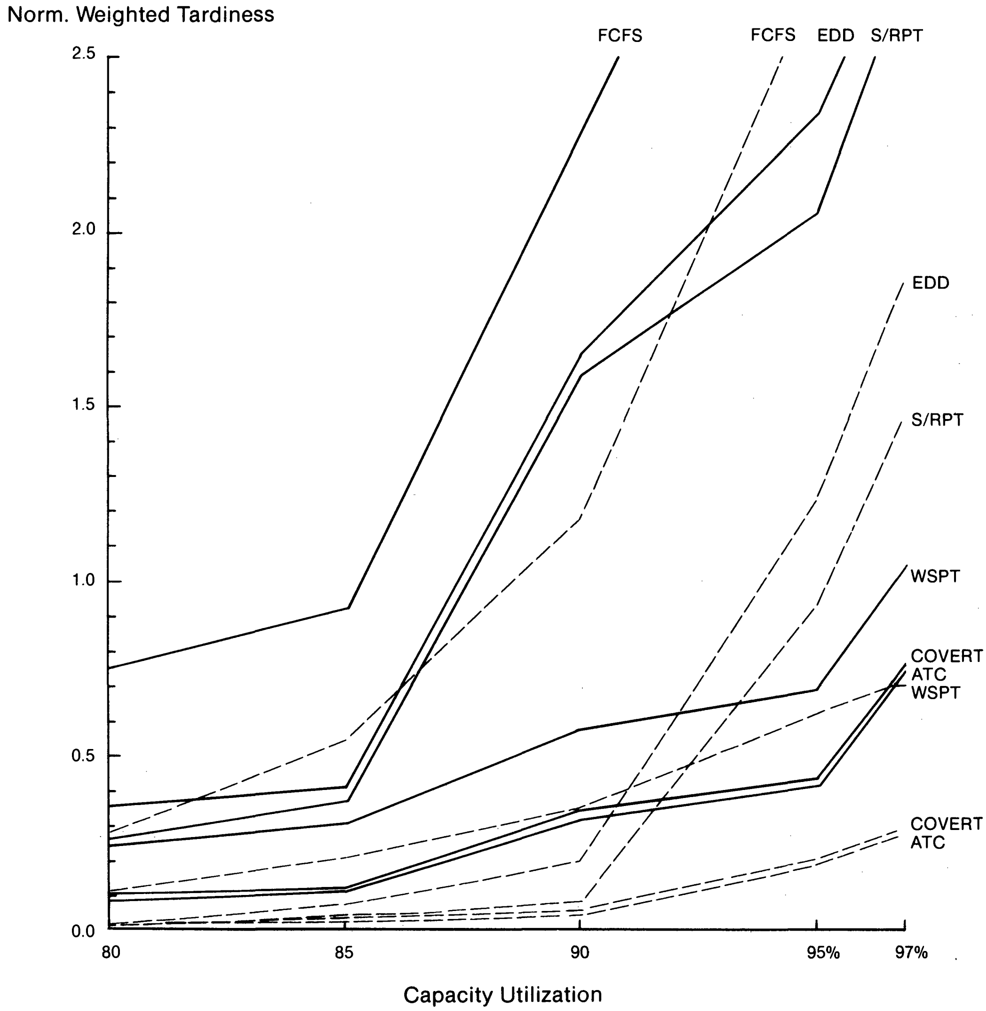

FIGURE 2. A Graph of Normalized Weighted Tardiness Results for Different Levels of Load; Loose Due Dates (Dashed Line) and Tight Due Dates (Solid Line).

Large-scale computational experiments in job shops have confirmed that the more complete rules dominate the previous heuristics in terms of weighted-tardiness performance. Overall, the ATC rule ranks first in all load conditions, exhibiting robustness not achieved by any rule in previous studies testing mean tardiness. The COVERT rule is a close second and reliable especially in congested shops and with tight due dates. The superior performance of these rules carries over to the portion of tardy jobs, and the well-coordinated schedules naturally eliminate high in-process inventories.

This demonstrated versatility for several objectives in wide variety of different shops and load conditions highlights a sound approach to priority dispatching. The ATC priority index maximizes the expected benefits (avoidance of tardiness costs) of sequencing a job at the current position per the opportunity cost of the machine, or more generally:

$$\text{Priority Index} = \frac{\text{(expected cost of delaying activity)}}{\text{(opportunity cost of resources)}}.$$

The expected rates of delay penalties are determined, in part, by the slack evaluation

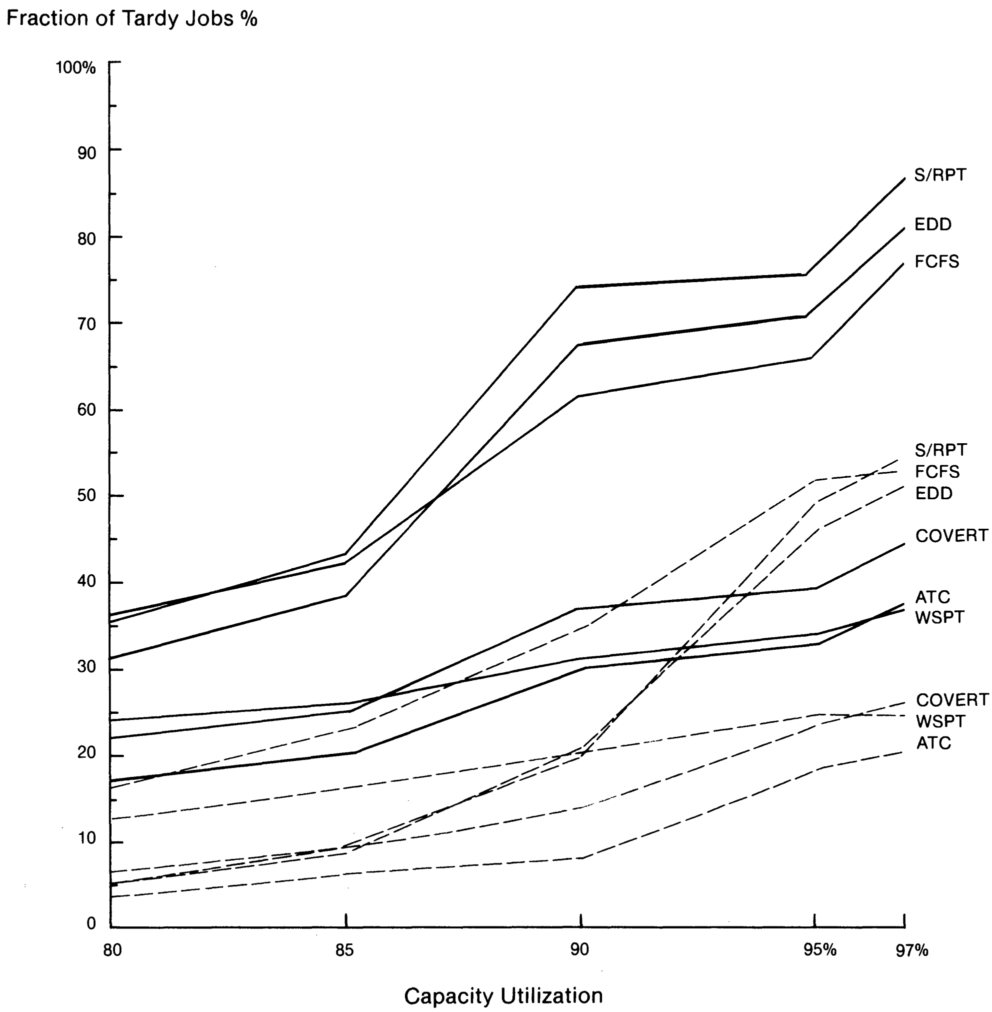

FIGURE 3. A Graph of the Average Portion of Tardy Jobs for Different Levels of Load: Loose Due Dates (Dashed Line) and Tight Due Dates (Solid Line).

parameters. In the case of the ATC rule, the look-ahead and leadtime estimation parameters should reflect the future load in the shop, akin to dual variables for intertemporal coordination of the dispatching process. In our tests the ATC and COVERT rules outperformed the competing rules even with fixed "average" parameter values, and contrary to practitioners' previous suspicion, improving adjustments can be made for extreme load conditions on the basis of expected queue lengths, for instance (Vepsalainen 1984). Furthermore, the new rules can be enhanced by applying more appropriate leadtime estimation methods (Vepsalainen 1984, Vepsalainen and Morton 1985).

The extension of this greedy priority approach to more general scheduling problems appears straightforward. Other cost factors besides tardiness penalties could be integrated into the benefit side explicitly, for instance inventory holding and earliness costs (Ow and Morton 1984), and rush shipping costs. In the machine-constrained case above, job processing times suffice as the surrogates of resource externalities. Multiple resources and their opportunity costs could be considered as well, including machines, robots, labor, tools and fixtures (Sathi et al. 1984). Sequence-dependent set-up costs,

TABLE 3

The Performance of the Rules in Terms of Normalized Work-in-Process Inventories (Work-in-System Holding Costs Shown in Parentheses).

(a) Loose due data allowances
Normalized WIP (WIS):

Rule:	Approximate Load: 80%	85%	90%	95%	97%
FCFS	2.55 (6.85)	3.20 (6.88)	4.46 (7.38)	6.32 (8.35)	6.80 (8.91)
EDD	2.30 (6.46)	2.78 (6.05)	3.47 (5.91)	4.37 (5.85)	4.69 (6.26)
S/RPT	2.62 (6.58)	3.12 (6.30)	3.86 (6.08)	4.72 (5.80)	4.96 (6.24)
WSPT	2.23 (6.98)	2.74 (7.14)	3.59 (7.83)	4.86 (8.78)	5.35 (9.37)
COVERT	2.48 (6.59)	2.96 (6.34)	3.86 (6.26)	4.74 (6.18)	5.11 (6.61)
ATC	2.30 (6.51)	2.75 (6.25)	3.49 (6.14)	4.35 (6.06)	4.77 (6.53)

(b) Tight due date allowances
Normalized WIP (WIS):

Rule:	Approximate Load: 80%	85%	90%	95%	97%
FCFS	2.85 (4.29)	3.08 (4.36)	4.64 (5.45)	5.20 (5.97)	7.40 (7.90)
EDD	2.43 (3.64)	2.54 (3.59)	3.34 (3.97)	3.75 (4.25)	4.74 (5.08)
S/RPT	2.59 (3.62)	2.70 (3.61)	3.58 (3.96)	3.77 (4.15)	4.63 (4.88)
WSPT	2.43 (4.31)	2.63 (4.47)	3.78 (5.46)	3.97 (5.62)	5.57 (7.10)
COVERT	2.60 (3.84)	2.72 (3.84)	3.77 (4.42)	4.06 (4.69)	5.34 (5.78)
ATC	2.43 (3.84)	2.60 (3.87)	3.46 (4.54)	3.83 (4.62)	5.11 (5.73)

multiple routings, and project scheduling pose further analytical challenges to this approach. The implementation of these extended rules requires flexible computational support, such as knowledge-based systems with facilities for formal integration of managerial judgment to the scheduling process (Fox 1983, Sathi et al. 1984, Vepsalainen and Greenberg 1983).[9]

[9] This article is based on research sponsored, in part, by the Air Force Office of Scientific Research under contract F49620-82-K-00017, the Westinghouse Corporation, and the Robotics Institute at Carnegie-Mellon University. The authors thank Paul Kleindorfer and two anonymous referees for helpful comments.

References

ADAM, N. R. AND J. SURKIS, "Priority Update Intervals and Anomalies in Dynamic Ratio Type Job Shop Scheduling Rules," *Management Sci.*, 26, 12 (December 1980), 1227–1237.

BAKER, K. R., *Introduction to Sequencing and Scheduling*, John Wiley & Sons, Inc., New York, 1974.

———, "Sequencing Rules and Due-Date Assignments in a Job Shop," *Management Sci.*, 30, 9 (September 1984), 1093–1104.

——— AND J. W. M. BERTRAND, "A Dynamic Priority Rule for Sequencing Against Due-Dates," *J. Oper. Management*, 3, 1 (November 1982), 37–42.

BERRY, W. L. AND V. RAO, "Critical Ratio Scheduling: An Experimental Analysis," *Management Sci.*, 22, 2 (October 1975), 192–201.

BUFFA, E. S. AND J. G. MILLER, *Production-Inventory Systems: Planning and Control*, 3rd Ed., Richard Irwin Inc., Homewood, Ill., 1979.

CARROLL, D. C., "Heuristic Sequencing of Jobs with Single and Multiple Components," Ph.D. Thesis, Sloan School of Management, MIT, 1965.

CONWAY, R. W., W. L. MAXWELL AND L. W. MILLER, *Theory of Scheduling*, Addison-Wesley Inc., Reading, Mass., 1967.

ELVERS, D. A. AND L. R. TAUBE, "Time Completion for Various Dispatching Rules in Job Shops," *OMEGA*, 11, 1 (1983), 81–89.

FOX, M. S., "Job Shop Scheduling: An Investigation in Constraint-Directed Reasoning," Ph.D. Thesis, Department of Computer Science, Carnegie-Mellon University, April 1983.

GRAVES, S. C., "A Review of Production Scheduling," *Oper. Res.*, 29, 4 (July–August 1981), 646–675.

MUTH, J. F. AND G. L. THOMPSON (EDS.), *Industrial Scheduling*, Prentice-Hall, Englewood Cliffs, N.J., 1963.

OW, P. S., "Focused Scheduling in Proportionate Flowshops," *Management Sci.*, 31, 7 (1985), 852–869.

——— AND T. E. MORTON, "Single Machine Early/Tardy Problem," Working Paper #42-83-84, Graduate School of Industrial Administration, Carnegie-Mellon University, 1984.

PANWALKAR, S. S. AND W. ISKANDER, "A Survey of Scheduling Rules," *Oper. Res.*, 25, 1 (1977), 45–61.

RACHAMADUGU, R. V., "Myopic Heuristics in Job Shop Scheduling," Ph.D. thesis, Graduate School of Industrial Administration, Carnegie-Mellon University, 1982.

——— AND T. E. MORTON, "Myopic Heuristics for the Single Machine Weighted Tardiness Problem," Working Paper #28-81-82, Graduate School of Industrial Administration, Carnegie-Mellon University, 1981.

SATHI, A., M. S. FOX, M. GREENBERG AND T. E. MORTON, "CALLISTO: An Intelligent Project Management System," Technical report, The Robotics Institute, Carnegie-Mellon University, 1984.

VEPSALAINEN, A. P. J., "State-Dependent Priority Rules for Scheduling," Ph.D. thesis, Graduate School of Industrial Administration, Technical Report CMU-RI-TR-84-19, The Robotics Institute, Carnegie-Mellon University, 1984.

——— AND M. GREENBERG, "Knowledge Representation in a Project Scheduling System," in 14th Pittsburgh Conference on Simulation and Modeling, University of Pittsburgh, Pittsburgh, PA, 1983.

——— AND T. E. MORTON, "Leadtime Estimation for Priority Scheduling with Strategic Tardiness Costs," Working Paper #85-09-03, Department of Decision Sciences, University of Pennsylvania, 1985.