J SYST SCI SYST ENG
Vol. 32, No. 5, October 2023, pp. 553–570
DOI: https://doi.org/10.1007/s11518-023-5574-8

ISSN: 1004-3756 (paper), 1861-9576 (online)
CN 11-2983/N

# Rule based vs Optimization based Workload Control with and without Exogenous Lead Times: An Assessment by Simulation

**Mingze Yuan,**[a, b] **Ting Qu,**[b, c, d] **Matthias Thürer,**[e] **Lin Ma,**[f] **Lei Liu**[a]

[a]School of Management, Jinan University, Guangzhou 510632, China
liuleifeichuan@foxmail.com
[b]GBA and B&R International Joint Research Center for Smart Logistics, Jinan University, Zhuhai 519070, China
mingze@stu2020.jnu.edu.cn
[c]School of Intelligent Systems Science and Engineering, Jinan University, Zhuhai 519070, China
[d]Institute of Physical Internet, Jinan University, Zhuhai 519070, China
quting@jnu.edu.cn
[e]Chair of Factory Planning and Intralogistics, Chemnitz University of Technology, Chemnitz 09125, Germany
matthiasthurer@workloadcontrol.com (✉)
[f]Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China
Malin15102939217@163.com

**Abstract.** Order release is a key production planning and control function, specifically in high variety contexts. A large literature on release methods that balance the workload consequently emerged. These Workload Control methods can be rule based, using a simple greedy heuristic, optimization based or optimization based with lead times that are exogenous. Although all three types of methods have the same objective, their performance has never been compared. Using simulation, this study shows that a better on time delivery performance of jobs can be achieved by the two optimization based release methods. Most importantly, optimization based methods that assume lead times to be exogenous significantly outperform alternative methods in terms of tardiness performance. Rule based and optimization based Workload Control without exogenous lead times overemphasize average lateness reduction, which leads to sequence deviations that offset performance improvements through balancing. In contrast, Workload Control methods that assume lead times to be exogenous limit sequence deviations, which leads to a significant reduction in dispersion of lateness. This has important implication for the future design of order release methods, and managerial practice.

**Keywords:** Workload control, order release, production planning, production control, make-to-order production.

## 1. Introduction

The degree of coordination between demand, material flows and capacity is a key determinant of business success. Order release is a key production planning and control function to realize this coordination. When order release is applied, then orders are not immediately released to the shop upon their arrival. Jobs are withheld in a backlog or pool from where they are released to realize certain performance metrics, such as stable Work-In-Process (WIP) levels and due date adherence. This study focuses on Workload Control (WLC), e.g. Wiendahl et al. (1992), Bechte (1994), Land and Gaalman (1998), Cigolini and Portioli-Staudacher (2002), Stevenson et al. (2005), Neuner and Haeussler (2021), Haeussler et al. (2023), which is arguably the only group of order release methods that allows for workload balancing, i.e. the equal distribution of workload across workstations. This makes it specifically suited for high-variety production characterized by stochastic demand fluctuations. We argue that the recent WLC literature can be subdivided into three strands: (i) rule based WLC methods; (ii) optimization based WLC methods; and (iii) optimization based WLC methods that assume that lead times are exogenous, i.e. lead times are considered as part of the release method. Lead times are a planning parameter that estimates the throughput time.

Rule based WLC methods use a simple greedy heuristic to decide which job to release when from the pool, e.g. Bechte (1988), Hendry and Kingsman (1991), Oosterman et al. (2000), Sabuncuoglu and Karapinar (2000), Land (2006), Thürer et al. (2012 2014), Liu et al. (2022). Jobs in the pool are first sequenced according to a priority measure (e.g. urgency) and then subject to a selection decision in sequence. This selection decision typically checks whether an order can be released without violating a preestablished workload limit (or norm) at stations.

Optimization based WLC methods without exogenous lead times integrate the sequencing and selection decision of rule based WLC into a single decision, e.g. Irastorza and Deane (1974), Yan et al. (2016), Portioli-Staudacher and Tantardini (2012), Fernandes et al. (2020), Kundu et al. (2021). Most of the literature used an Integer Linear Programming (ILP) model that integrates load and urgency into one objective function setting different weights.

Optimization based WLC methods with exogenous lead times use the lead time as a parameter, e.g. Haeussler and Netzer (2020), Haeussler et al. (2020). Release is, for example, operationalized using an Input/Output Control (IOC) model with specified fixed integer lead times for all workstations instead of a workload limit. The WIP at each workstation is limited by the constraint that it can be cleared within the lead times (Pürgstaller and Missbauer 2012).

All three types of WLC methods are similar. Rule based WLC, optimization based WLC, and optimization based WLC with exogenous lead times create a pool from where jobs are released based on the workload situation. This is different from due date based release methods such as, for example, incorporated in Material Requirements Planning (MRP). The main difference across WLC methods is that rule based WLC uses a greedy heuristic and the workload limit as the main constraint, optimization based uses optimization and the workload limit is a main constraint, and optimization based with exogenous lead times uses optimization without an explicit workload limit. But despite their similarities, no study to date compared all three WLC methods. Pürgstaller and Missbauer (2012) compared optimization based with exogenous lead times and rule based WLC. They concluded that the former largely outperforms rule based WLC in their simulations, except in the case of a largely constant demand and product mix. But Pürgstaller and Missbauer (2012) did not use more advanced rule based release methods, such as the one presented in Thürer et al. (2012). Fernandes et al. (2020), concluded that optimization based can outperform rule based WLC, especially in terms of percentage tardy and mean tardiness. But Fernandes et al. (2020) did not consider optimization based WLC with exogenous lead times. Managers are consequently still left alone with the decision which WLC method to choose for their shop. In response this study asks:

*Which WLC method (rule based, optimization based or optimization based with exogenous lead times) leads to the best tardiness performance?*

Discrete event simulation will be used to answer this question. In the next section, we introduce the order release methods to be included in our study. The simulation model used to assess performance is then described in Section 3, before results are presented and analyzed in Section 4. A discussion is presented in Section 5. Finally, we conclude in Section 6, where we also summarize managerial implications and limitations.

## 2. Background

This section introduces the specific WLC methods that will be considered in this study. Rule based WLC is reviewed in Section 2.1, optimization based WLC Section 2.2, and optimization based WLC with exogenous lead times in Section 2.3.

### 2.1 Rule based WLC

A broad set of different rule based WLC methods has been presented in the literature. In this paper, the LUMS COR (Lancaster University Management School Corrected Order Release) method is used given its good performance in previous studies (Thürer et al. 2012). Although new methods, such as Continuous Release (Fernandes et al. 2017) have the potential to outperform LUMS COR, we decided for LUMS COR to keep it closer to optimization based release. Continuous Release was included in the study by Fernandes et al. (2020), who compared OPTimization release method (OPT) and LUMS COR. So, we are confident that our findings also hold for this method. LUMS COR uses a periodic release procedure to keep the corrected workload $W_s$ that is released to workstation $s$ and not yet completed within a pre-established workload limit or norm $N_s$ as follows:

1) All jobs in the set of jobs $J$ in the pre-shop pool are sorted according to the planned release date.

2) The job $j \in J$ with the highest priority is considered for release first.

3) Take $r_j$ to be the ordered set of operations in the routing of job $j$. If job $j$'s processing time $P_{ij}$ at the $i^{th}$ operation in its routing – corrected for station position $i$ – together with the workload $W_s$ released to station $s$ (corresponding to operation $i$) and yet to be completed fits within the workload norm $N_s$ at this station, that is $\frac{p_{ij}}{i} + W_s \leq N_s,\ \forall i \in r_j$, then the job is selected for release. That means it is removed from $J$ and its load contribution is included, i.e. $W_s := W_s + \frac{p_{ij}}{i},\ \forall i \in r_j$. Otherwise, the job remains in the pool and its processing time does not contribute to the station load.

4) If the set of jobs $J$ in the pool contains any jobs that have not yet been considered for release, then return to Step 2 and consider the job with the next highest priority. Otherwise, the release procedure is complete, and the selected jobs are released to the shop floor.

Since a released job contributes to $W_s$ until its operation at this station is completed, the load contribution to a station in LUMS COR is calculated by dividing the processing time of the operation at a station by the station's position in a job's routing (Oosterman et al. 2000). In addition to the above periodic release mechanism, LUMS COR incorporates a continuous workload trigger. If the workload of any station falls to zero, the next job in the pool sequence with that station as the first in its routing is released irrespective of whether this would exceed the workload norms of any station in a bid to avoid premature station idleness.

## 2.2 Optimization Based WLC

This study focuses on methods that use ILP models. Irastorza and Deane (1974) were the first to apply linear programming to order release, integrating lower-workload, upper-workload, and urgency of orders into an objective function. Since then, there has been increasing interest, e.g. Yan et al. (2016), Fernandes et al. (2020), Kundu et al. (2021). The main difference among the different studies is the weight given to each element in the objective function. Fernandes et al. (2020) recently argued that the use of weights is not meaningful given that urgency and workload represent different measures. They consequently suggested that OPT should only focus on the workload and that urgency should be integrated by creating two classes: urgent and non-urgent orders. Urgent orders are released first according to OPT before non-urgent orders are released according to rule based WLC. The method proposed in Fernandes et al. (2020) will be considered in this study since it yielded the best performance.

The OPT release for urgent orders is executed according to the following model:

$$\min f = \sum_s (N_s - W_s) \tag{1}$$

$$W_s = A_s + \sum_i Y_j \times \frac{p_{ij}}{i}, \forall i \in r_j \tag{2}$$

$$W_s \leq N_s, \forall s \in S_j \tag{3}$$

$$Y_j = \begin{cases} 0, & \text{if } j \text{ is not selected for release} \\ 1, & \text{if } j \text{ is selected for release} \end{cases} \quad \forall j \tag{4}$$

$$W_s, A_s \geq 0, \forall s, j \tag{5}$$

Where $A_s$ is the 'corrected' aggregate load at station $s$ before the release decision is made, $S_j$ is the set of stations required for the processing of order $j$, while other variables are as defined in Section 2.1. Meanwhile, $Y_j$ is the decision variable. The release decision is executed by setting this decision variable to 1 (selected for release) or 0 (not selected for release) for each of the jobs in the pre-shop pool. The optimization procedure sets these variables so that the objective function is minimized given the constraint defined in Equation (3). Once the release decision has been made, the load of the released jobs is credited to the current stations' workload, and the jobs are released and removed from the pool list.

## 2.3 Optimization Based WLC with Exogenous Lead Times

This study only focuses on methods that use simple IOC models to ensure similar complexity (and consequently applicability in practice) across the methods compared. This means that optimization methods using non-integer lead times (e.g. Hackman and Leachman (1989), Kacar et al. (2016)) or iterative mechanisms to determine lead times (e.g. Leachman and Carmon (1992), Kim and Kim (2001), Hnaien et al. (2008)) are neglected. If methods using simple IOC models outperform other methods, then this will likely also hold for these more complex methods.

IOC subdivides the production plan into time periods and determines the amount of work to be released per period. It specifies fixed, integer lead times for all workstations, and then limits the WIP at each workstation so that the WIP can be cleared within the lead times (e.g. Pürgstaller and Missbauer (2012), Haeussler et al. (2020)). It is thus closely aligned with the idea of input/output control Wight (1970), Plossl and Wight (1973) that underlies all WLC methods. The original model, as introduced by Pürgstaller and Missbauer (2012), is not provided here given space constraints. The model used in this study integrates some variables of the original model into a set. Using the notation in Table 1, and assuming that the production system consists of several workstations with input and output buffers, and that there is a finished goods inventory at the end of the production system, the IOC model can be formulated as follows:

$$\min \sum_s \sum_j \sum_t (\omega_{js} W_{jts} + \omega_{js} F_{jts} + \theta_j B_{jts}) + \sum_j \sum_t (\varphi_j F_{jt} + \theta_j B_{jt}) \tag{6}$$

$$W_{jts} = W_{j,t-1,s} + R_{jts} - X_{jts}, \forall j, t, s \tag{7}$$

$$F_{jts} - B_{jts} = F_{j,t-1,s} - B_{j,t-1,s} + X_{jts} - R_{j,t-\mu_j,s} - D_{jts}, \forall j, t, s \neq \max(r_j) \tag{8}$$

$$F_{jt} - B_{jt} = F_{j,t-1} - B_{j,t-1} + R_{j,t-\mu_j,n} + \bar{R}^f_{jts} - D_{jt}, \forall j, t, s = \max(r_j) \tag{9}$$

$$\sum_j \alpha_{js} W_{jts} \le \sum_j \sum_{l=t+1}^{t+\tau_{ts}} \alpha_{js} X_{jls}, * \forall s; \text{for } t = 1, \cdots, T - \tau_{ts} \tag{10}$$

$$\sum_j \alpha_{js} W_{jts} \le \sum_j \sum_{l=t+1}^{T} \alpha_{js} X_{jls} + (t + \tau_{ts} - T)C_s, * \forall s; \text{for } t = T - \tau_{ts} + 1, \cdots, T \tag{11}$$

$$\sum_j p_{js} X_{jts} \le C_s, \forall j, t, s \tag{12}$$

$$W_{jts}, F_{jts}, X_{jts}, R'_{jts} \ge 0, \forall j, t, s \tag{13}$$

If workstation $s$ is the last processing operation in the routing of job $j$, $F_{jts}$ can be transferred immediately to the finished goods inventory $F_{jt}$ and is removed. Meanwhile, $D_{jts}$ denotes the quantity of job $j$ that was released before the current period time and should be transferred from workstation $s$ to another workstation or finished goods inventory in period $t$. Equation (5) presents the resource balance constraint for WIP inventory at the input buffer of each workstation. Equation (6) and (7) define WIP balance and inventory balance for different workstations. Equation (6) defines resource balance when workstation $s$ is not the last workstation in the routing of order $j$. Equation (7) defines resource balance when workstation $s$ is the last workstation in the routing of order $j$. Constraints (8) and (9) limit the WIP at the input buffer of workstations during the whole planning horizon $T$. The length of time $t$ in constraint (9) ensures the equality of timely cleared orders. Constraint (10) ensures restricted production resource usage during a period at each workstation. Finally, constraint (11) assures non-negative values of the deci-

**Table 1** Symbols Used in the IOC Model

(a) Indices Used in Model

| Indices | |
|---|---|
| $t$ | Period index ($t = 1, \cdots, T$) |
| $s, n$ | Workstation index ($s, n = 1, \cdots, S^*$) |
| $j$ | Product index ($j = 1, \cdots, J$) |

(b) Decision Variables Used in Model

| Decision variables | |
|---|---|
| $W_{jts}$ | WIP of product $j$ at the input buffer or processing at station $s$ at the end of period $t$ |
| $F_{jt}$ | Finished quantity of product $j$ at the end of period $t$ |
| $F_{jts}$ | Finished quantity of product $j$ at the end of period $t$ at the output buffer of workstation $s$ |
| $B_{jt}$ | Backlog quantity of product $j$ at the end of period $t$ |
| $B_{jts}$ | Backlog quantity of product $j$ at workstation $s$ at the end of period $t$ |
| $R_{jt}$ | Released quantity of product $j$ in period $t$ |
| $R_{jts}$ | Released quantity of product $j$ to station $s$ in period $t$ |
| $X_{jts}$ | Quantity of product $j$ that entered the output buffer at workstation $s$ in period $t$ |

(c) Parameters Used in Model

| Parameters | |
|---|---|
| $\bar{R}^n_{jts}$ | Quantity of product $j$ that flows from workstation $s$ to workstation $n$ in period $t$ |
| $\bar{R}^f_{jts}$ | Quantity of product $j$ that flows from workstation $s$ to finished goods in period $t$ |
| $C_s$ | Capacity at workstation $s$ |
| $D_{jt}$ | Demand of product $j$ in period $t$ |
| $\omega_{js}$ | Unit cost of product $j$ at workstation $s$ |
| $\gamma_{js}$ | Unit cost of product $j$ at output buffer of workstation $s$ |
| $\varphi_j$ | Unit cost of product $j$ in finished goods inventory |
| $\theta_j$ | Unit backlog penalty cost of product $j$ if tardy finished |
| $\alpha_{js}$ | Processing time of product $j$ at workstation $s$ |
| $\tau_{ts}$ | Lead time at workstation $s$ at the period $t$ |
| $\mu_{js}$ | Lead time of order $j$ at the total upstream workstation of workstation $s$ |

sion variables.

## 3. Simulation Model

The simulated shop and job characteristics are first summarized in Section 3.1. How we operationalized the three WLC methods is then outlined in Section 3.2, before the dispatching rules considered are summarized in Section 3.3. Finally, a description of the experimental design and the performance measures is given in Section 3.4. The production environment and the release methods are implemented in Python.

### 3.1 Shop and Job Characteristics

We consider two shop types: a Pure Job Shop (PJS) and a General Flow Shop (GFS). Both contain six workstations with equal, constant capacity. Following Oosterman et al. (2000), the routing length of jobs varies uniformly from one to six operations for both shops. The routing length is first determined before the routing sequence is generated randomly without replacement, i.e. re-entrant flows are prohibited as in e.g. Fernandes et al. (2020). This leads to the routing vector for the PJS. For the GFS, this routing vector is sorted such that the routing becomes directed and there are typical upstream and downstream stations. Operation processing times follow a 2-Erlang distribution with a truncated mean of 1 time unit and a maximum of 4 time units. The workload limit in LUMS COR and OPT should be larger than the maximum possible processing time. To still allow for tight control, operation processing times are truncated. The inter-arrival times of orders follow an exponential distribution with a mean of 0.648 time units, which deliberately results in an utilization of 90%. We ensured that there is an equal utilization for all methods tested. This means all methods produce the same amount of work and given the use of common random number streams also the same set of jobs. Due dates are established by adding a random allowance, uniformly distributed between 30 and 45 time units, to the entry time of orders.

### 3.2 WLC Order Release

As in previous simulation studies on WLC, it is assumed that all jobs are accepted, materials are available, and all necessary information regarding shop floor routings, processing times, etc. is known. Jobs flow directly into the pre-shop pool at their arrival and await release by the release methods. The parametrization of each rule is presented next. As a baseline measure, experiments without controlled order release have also been executed, i.e. jobs are released onto the shop floor immediately upon arrival.

#### 3.2.1 Rule Based WLC - LUMS COR

Eight settings for the workload norm $N_s$ are considered, from 5 to 12 time units. The periodic release interval was set to 4 time units. The planned release date of a job is given by its due date minus an allowance for the operation throughput time for each operation in its routing. The allowance for the operation throughput time at each station was set to 3 time units based on preliminary simulation experiments.

#### 3.2.2 Optimization Based WLC– OPT

OPT uses the same setting for the workload norm and the periodic time interval as LUMS COR. To define urgent and non-urgent jobs, the same planned release dates are used. Urgent jobs are released using the OPT model.

#### 3.2.3 Optimization Based WLC with Exogenous Lead Times – IOC*

The optimization model is triggered every release period, i.e. a rolling horizon release planning with periodic re-planning of orders is simulated. The planning horizon $T$ was set to 12 time units based on previous literature, e.g. Haeussler et al. (2020). Lead times are estimated backward as follows:

$$L^{i}_{js} = p_{js} f, i = 1 \tag{14}$$

$$L^{i}_{js} = L^{i-1}_{js'} + p_{js} f, i > 1 \tag{15}$$

where $L^{i}_{js}$ denotes the lead time estimate for job $j$ at workstation $s$, i.e. the $i^{th}$ operation in the routing of job $j$, which is determined by the lead time of the previous operation in the routing, i.e. the $(i-1)^{th}$ operation, and the processing time of job $j$ at workstation $s$, and f is the flow factor of job $j$, which is the ratio of a job's lead time to the sum of all processing times of all operations.

We further adopted the rounding method to obtain integer lead time estimates from the non-integer values observed in the simulation. The overall integer lead time of a job is determined by the sum of each operation's non-integer lead time. Note that we also conducted additional experiments using non-integer lead times. These experiments showed no significant performance differences across the methods compared in this study.

Meanwhile, the IOC* method considers several objectives. The objective function has two parts: costs incurred before the order is completed and costs incurred after the order is completed. The first part includes WIP cost at the input buffer of the workstation, inventory cost at the output buffer of the workstation, and backlog cost at the workstation. The second part includes the inventory cost and backlog cost of finished orders. The term backlog refers to orders that are tardy, i.e. for which the due date already passed the current date. These costs function equivalent to a weighting factor. For example, relative increases in finished goods inventory cost (compared to the other costs) can prevent early production, while increases in the backlog cost can reduce percentage tardy and mean tardiness. We consequently consider a criticality factor $q$ in this study, which is the ratio of finished inventory unit cost and backlog unit cost. We consider three levels: 0, 1/5, and 2/5, as summarized in Table 2. Theses cost ratios are based on previous literature Haeussler and Netzer (2020).

### 3.3 Dispatching

We consider two rules to choose the job that is processed next from the queue in front of a station on the shop floor in this study: First-Come-First-Served (FCFS) and Operation Due Date (ODD). FCFS was chosen as a benchmark. ODD was chosen since it performed well in job shops (Kanet and Hayya 1982). IOC* includes an indication of the operation due dates. For rule based WLC and OPT, the operation due date for the last operation in the routing of a job is equal to the due date while the operation due date of each preceding operation is determined by successively subtracting an allowance for the operation throughput time from the operation due date of the next operation. In this study, the allowance for the operation throughput time at each station is set to 3 time units based on preliminary simulation

**Table 2** Cost Parameters Considered in this Study

| | IOC*I | IOC*II | IOC*III |
|---|---|---|---|
| WIP unit cost ($\omega_{js}$) | 15 | 15 | 15 |
| Finished Inventory unit cost ($\varphi_j$) | 0 | 15 | 30 |
| Backlog unit cost ($\theta_j$) | 75 | 75 | 75 |
| Cost criticality factor ($q$) | 0 | 1/5 | 2/5 |

experiments.

### 3.4 Experimental Design and Performance Measures

The experimental factors are: (i) the two shop types; (ii) the three different order release methods (LUMS COR, OPT and IOC*) and their parametrization; and (iii) the two dispatching rules. This results in 84 (2 x 21 x 2) scenarios using a full factorial design. Each scenario was replicated 100 times, and for each replication data was collected for 10,000 time units, being the warm-up period set to 3,000 time units.

There is no best method and better in our study refers to one method leading to better results for a given measure and in a very specific context. Our focus is on tardiness performance and a broad set of measures is considered to capture this performance. The five principal performances measures considered in this study are as follows: *the gross throughput time*, i.e. the completion time of the order minus its entry time; *the shop floor throughput time*, i.e. the gross throughput time minus the queuing time in the pre-shop pool; *the percentage tardy*, i.e. the percentage of orders delivered after the customer due date; *the mean tardiness*, where $T_j = \max(0, L_j)$ indicates the tardiness of job $j$, with $L_j$ being the lateness of job $j$ (i.e. the actual delivery date minus the due date of job $j$); and *the mean earliness*, where $E_j = \mid \min(0, L_j) \mid$ indicates the earliness of job $j$. We consider earliness since we have a finished goods inventory. In contrast, most of the WLC release literature considers a pure make-to-order context where orders are directly delivered to the customer and, most importantly, the customer values this early delivery.

## 4. Results

### 4.1 Performance Assessment

Since there is no 'best' parameter setting across the different measures, we opted for performance curves. Rather than showing a single value for each method, we illustrate performance over a range of norm levels which allows for qualitative comparison of methods across different performance measures. Each performance curve represents one release method, and each data point on a curve represents a specific norm level for LUMS COR and OPT or a specific factor $q$ for the IOC* model. The main simulation results for the PJS are illustrated in Figure 1a and Figure 1b for FCFS and ODD dispatching, respectively. In addition, the results for immediate release (IMM) are given by an isolated point. Meanwhile, shop floor throughput times are commonly used as instrumental variable in the order release literature since tightening the norms reduces shop

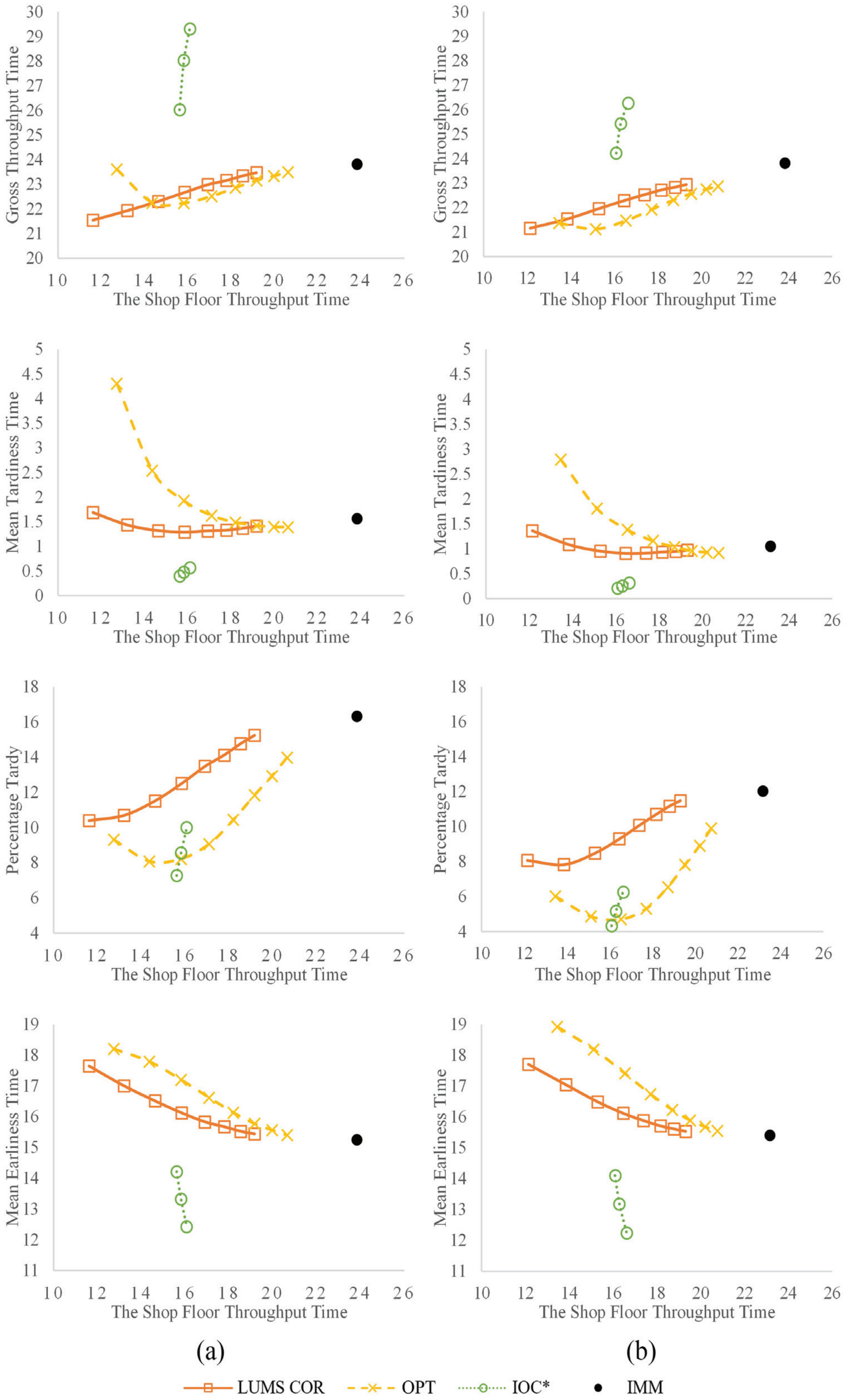


**Figure 1** Results of LUMS COR, OPT, and IOC* Release Methods in the PJS: (a) FCFS Dispatching Rule; (b) ODD Dispatching Rule

floor throughput times. This is not true for IOC, since it does not use a norm, which explains the vertical line. We still kept the shop floor throughput time at the x-axis given common convention in the WLC literature.

The performance of the different order release methods can be evaluated by comparing the performance curves in each figure. Results confirm Fernandes et al. (2020) – OPT outperforms LUMS COR in terms of percentage tardy and it can provide similar mean tardiness levels if workload norms are set appropriately. This is achieved by a reduction in gross throughput time (and thus mean lateness), leading tighter norms to an increase in the dispersion of lateness, as can be observed from the mean tardiness and mean earliness performance. Results extend Pürgstaller and Missbauer (2012) showing that IOC* also outperforms more advanced rule based WLC methods. Most importantly, IOC* leads to the best performance in terms of percentage tardy, mean tardiness and mean earliness. It significantly reduces the dispersion of lateness, which offsets the loss in gross throughput time (and thus mean lateness) performance. This has important implication for the WLC literature, focusing the majority of studies on rule based order release or optimization based order release without exogenous lead times.

The performance of the two different dispatching rules can be evaluated by comparing results in Figure 1(a) (FCFS dispatching) and Figure 1(b) (ODD dispatching). As expected, ODD leads to better tardiness performance. Most importantly, the qualitative performance effect of the different release methods is not affected by the choice of dispatching rule. Finally, similar conclusions on the relative performance of the release methods and dispatching rules to those in the PJS can be drawn from our results in the GFS, i.e. when routings are directed. This is illustrated in Figure 2(a) and 2(b), which present the results for FCFS and ODD dispatching, respectively.

## 4.2 Performance Analysis

To better understand performance differences across methods, we collected the overtime development of the corrected aggregate workload at an arbitrary station for the PJS. Results, as given in Figure 3, were collected for a norm level of six for LUMSCOR and OPT and a cost criticality factor of 0 for IOC*. Results show that OPT keeps the corrected aggregate load closer to the workload norm than LUMS COR. This can be observed, for example, around 800 time units. Meanwhile, there is only a difference between both methods when not all urgent orders can be released. This is in high load periods when many jobs become tardy. In these periods OPT is able to better fill up the norm with urgent jobs. LUMSCOR further allows for exceeding the norm through its starvation avoidance trigger. IOC* allows for the highest load. This in turn gives the sequencing rule the capability to reduce the dispersion of lateness.

We only focused on operational performance measures so far. These remain rather separated and an integrative view is missing. Mezzogori et al. (2022) recently proposed to quantify the main benefits of WLC in economic terms, suggesting this as the easiest way to compare different and even conflicting performance measures. Costs and incomes are iden-

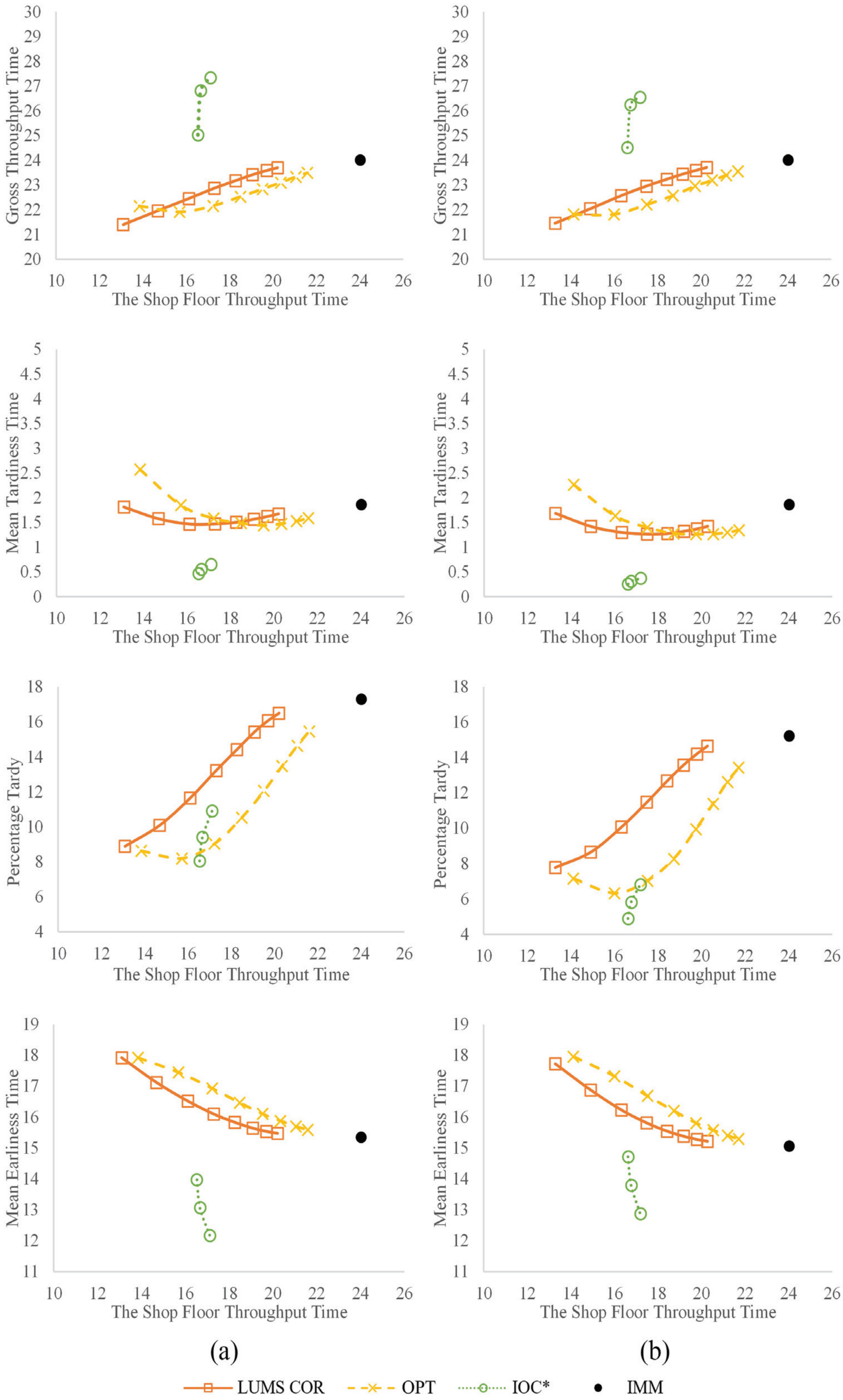


**Figure 2** Results of LUMS COR, OPT, and IOC* Release Methods in the GFS: (a) FCFS Dispatching Rule; (b) ODD Dispatching Rule

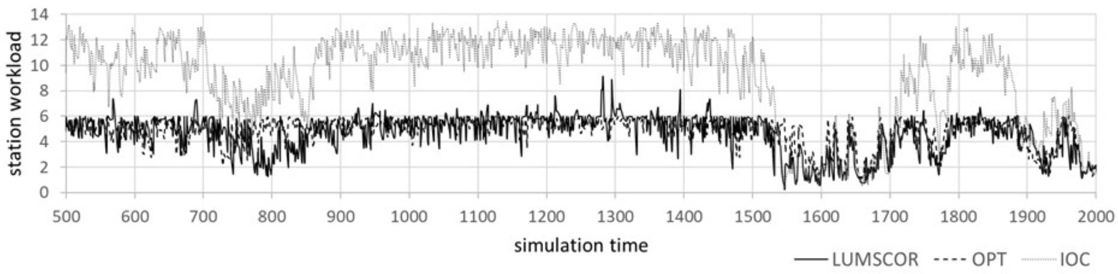


**Figure 3** Overtime Development of Station Workload

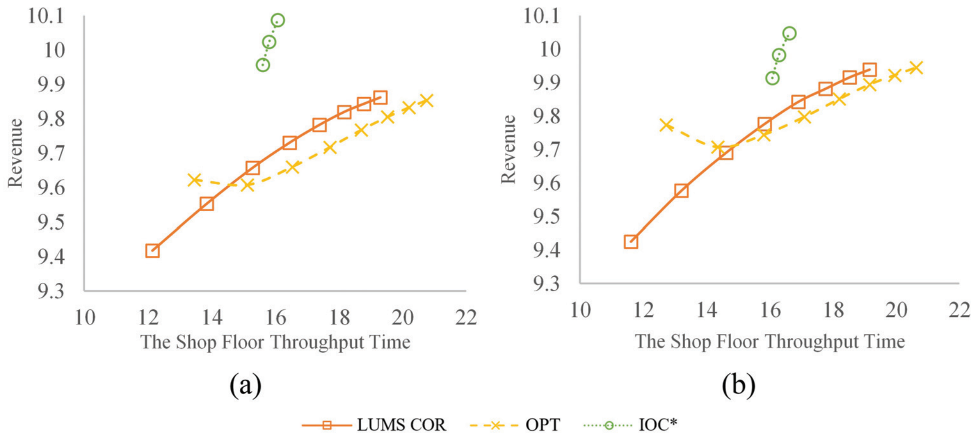


**Figure 4** Revenue per Job of LUMS COR, OPT, and IOC* Release Methods in the PJS: (a) FCFS Dispatching Rule; (b) ODD Dispatching Rule

tified and used to develop an overall economic measure that can be used to evaluate and fine tune the operating features of WLC. Results for this measure are given in Figure 4a and 4b, which present the results for FCFS and ODD dispatching, respectively. These results further confirm our findings.

## 5. Discussion

This study extends Pürgstaller and Missbauer (2012), Haeussler and Netzer (2020), who already showed that optimization based WLC methods are better than rule based WLC methods. This study extends these studies by showing that WLC methods that consider the lead time to be exogenous also have the potential to perform better than more advanced OPT methods, and rule based WLC methods that only later emerged.

The literature distinguishes between WIP regulating release methods, such as Constant WIP (ConWIP), and methods that balance the workload (Lödding 2012). WIP regulating methods control the number of jobs, which means they neglect the actual workload of each job. This leads to unbalanced workload situations when the workload varies across jobs as is typical for high variety contexts. Load balancing method, such as WLC, consider the workload. But this also means that a job that is more urgent may be delayed if its workload does not fit the workload limit. If the number of jobs is controlled, these sequence deviations are not introduced. The tighter the workload norm the larger the sequence deviations, which leads to some large jobs being delayed for a long time. Both LUMS COR and OPT suffer from sequence deviations, as can be observed from the increase in mean tardiness and mean earliness (and thus the dispersion of lateness) when norms are tightened. A main weakness of both LUMS COR and OPT is the focus on reducing average lateness introduced by the workload norm. The issue is the underlying model that underlies both. IOC* ensures that all jobs are cleared within the lead time. This avoids very large sequence deviations as can be observed from the reduction in mean tardiness and mean earliness (and thus the dispersion of lateness). As a drawback, IOC* is likely more complex.

Delivery performance can be improved by improved load balancing, which reduces the mean lateness, or improved timing, which reduces the dispersion of lateness (Soepenberg et al. 2012). Our results highlight that LUMS COR and OPT focus too much on load balancing. IOC* realizes a better trade-off in terms of load balancing and timing. However, this depends on due date tightness. Tighter due dates simply require the average lateness to be reduced, with improved balancing being one way of achieving this (Thürer et al. 2015). But for most practical situations, ensuring a mean lateness and then reducing the dispersion of lateness by appropriated weighting of earliness and tardiness is likely to lead to better performance.

## 6. Conclusion

WLC realizes a key production planning and control function: order release. A large literature on WLC release methods consequently emerged. WLC release methods can be either rule based, i.e., use a simple greedy heuristic, optimization based, or optimization based with lead times that are exogenous. Despite the large literature on WLC, these three types of methods have never been compared. Literature only compared a subset and remained inconclusive. This leaves managers alone in their task to choose an appropriate WLC order release method for their shop. In response, this study started by asking: Which WLC method (rule based, optimization based or optimization based with exogenous lead times) leads to the best tardiness performance? Simulation results show that WLC order release that considers lead times to be exogenous (IOC*) outperforms WLC order release methods that do not consider lead times as part of the planning procedure, which may be either rule based (LUMS COR) or optimization based (OPT). This has important implication for the future design of order release methods.

### 6.1 Managerial Implications

Our results indicate that a focus on WIP may not be the best if company performance is measured in earliness and tardiness. The shop floor goal of stable, short queues in front of stations may result in missing these company goals if sequence deviations are introduced. Measuring the queue does not indicate how long each individual job waits in the queue, but this is what is of importance to the customer. Workload Control release that assumes lead times to be exogenous, avoids large sequence deviations. Previous literature highlighted that delivery performance improvement requires a careful analysis whether better load balancing or improved sequencing is required. This study shows that sequencing can be improved by relaxing load balancing. It is a question of trade-offs rather than continuous reduction.

### 6.2 Limitations and Future Research

A main limitation of our study is the limited environmental setting. While we consider this to be justified to keep our study focused, future research could assess whether our findings also hold in other shop structures or under different levels of processing time variability. Meanwhile, we also only focused on input control, neglecting other manufacturing control functions to keep our study focused. Future research should include other manufacturing control functions, which were shown to significantly improve the performance of release methods (e.g. Mezzogori et al. (2021)). A further limitation is the neglect of more advanced optimization based release methods that consider lead times to be exogenous, such as methods that assume load dependent lead times. The main reason are the complexities involved, which are still not fully resolved and thus question the applicability in practice compared to the simpler methods compared in this study. Future research is needed to further develop these advanced methods, specifically in the light of our results that highlight their performance potential. This also includes simplifications and explanations that ensure that managers in practice understand the methods they will use, a key requirement for application.

## Acknowledgments

This work is financially supported by National Key Research and Development Program of China (2021YFB3301701), 2019 Guangdong Special Support Talent Program – Innovation and Entrepreneurship Leading Team (China) (2019BT02S593), 2018 Guangzhou Leading Innovation Team Program (China) (201909010006), and the Science and Technology Development Fund (Macau SAR) (0078/2021/A).

We also appreciate the sponsorships from the industry, including but not limited to Carpoly Chemical Group Co., Ltd. GBA and B&R International Joint Research Center for Smart Logistics is a provincial research lab sponsored by the Department of Science and Technology of Guangdong Province, thanks to which the international collaboration has been effectively conducted.

Finally, we gratefully thank the editor and all reviewers for their time spend making their constructive remark and useful suggestions, which has significantly raised the quality of the paper. Each suggested revision and comment, brought forward by the reviewers was accurately incorporated and considered.

## Data Availability

The datasets generated during and/or analysed during the current study are not publicly available but are available from the corresponding author on reasonable request.

## Conflicts of Interest

The authors declare no conflict of interest.

### References

Bechte W (1988). Theory and practice of load-oriented manufacturing control. *The International Journal of Production Research* 26(3): 375-395.

Bechte W (1994). Load-oriented manufacturing control just-in-time production for job shops. *Production Planning & Control* 5(3): 292-307.

Cigolini R, Portioli-Staudacher A (2002). An experimental investigation on workload limiting methods within ORR policies in a job shop environment. *Production Planning & Control* 13(7): 602-613.

Fernandes N O, Thürer M, Silva C, Carmo-Silva S (2017). Improving workload control order release: Incorporating a starvation avoidance trigger into continuous release. *International Journal of Production Economics* 194: 181-189.

Fernandes N O, Thürer M, Pinho T M, Torres P, Carmo-Silva S (2020). Workload control and optimised order release: An assessment by simulation. *International Journal of Production Research* 58(10): 3180-3193.

Hackman S T, Leachman R C (1989). A general framework for modeling production. *Management Science* 35(4):478-495.

Haeussler S, Netzer P (2020). Comparison between rule- and optimization-based workload control concepts: A simulation optimization approach. *International Journal of Production Research* 58(12): 3724-3743.

Haeussler S, Stampfer C, Missbauer H (2020). Comparison of two optimization based order release models with fixed and variable lead times. *International Journal of Production Economics* 227: 107682.

Haeussler S, Neuner P, Thürer M (2023). Balancing earliness and tardiness within workload control order release: An assessment by simulation. *Flexible Services and Manufacturing Journal* 35(2): 487-508.

Hendry L, Kingsman B (1991). A Decision Support System for Job Release in Make-to-order Companies. *International Journal of Operations & Production Management* 11(6): 6-16.

Hnaien F, Dolgui A, Ould Louly M A (2008). Planned lead time optimization in material requirement planning environment for multilevel production systems. *Journal of Systems Science and Systems Engineering* 17(2): 132-155.

Irastorza J C, Deane R H (1974). A loading and balancing methodology for job shop control. *AIIE Transactions* 6(4): 302-307.

Kacar N B, Mönch L, Uzsoy R (2016). Modeling cycle times in production planning models for wafer fabrication. *IEEE Transactions on Semiconductor Manufacturing* 29(2):153-167.

Kanet J J, Hayya J C (1982). Priority dispatching with operation due dates in a job shop. *Journal of Operations Management* 2(3): 167-175.

Kim B, Kim S (2001). Extended model for a hybrid production planning approach. *International Journal of Production Economics* 73(2): 165-173.

Kundu K, Land M J, Portioli-Staudacher A, Bokhorst J A (2021). Order review and release in make-to-order flow shops: Analysis and design of new methods. *Flexible Services and Manufacturing Journal* 33: 750-782.

Land M (2006). Parameters and sensitivity in workload control. *International Journal of Production Economics* 104(2): 625-638.

Land M J, Gaalman G J (1998). The performance of workload control concepts in job shops: Improving the release method. *International Journal of Production Economics* 56: 347-364.

Leachman R C, Carmon T F (1992). On capacity modeling for production planning with alternative machine types. *IIE Transactions* 24(4): 62-72.

Liu J, Land M J, Bokhorst J A, Chen Q (2022). Improving coordination in assembly job shops: Redesigning order release and dispatching. *Flexible Services and Manufacturing Journal*: 1-29.

Lödding H (2012). *Handbook of Manufacturing Control: Fundamentals, Description, Configuration.* Springer Science & Business Media, Germany.

Mezzogori D, Romagnoli G, Zammori F (2021). Defining accurate delivery dates in make to order job-shops managed by workload control. *Flexible Services and Manufacturing Journal* 33(4): 956-991.

Mezzogori D, Romagnoli G, Zammori F (2022). A new perspective on workload control by measuring operating performances through an economic valorization. *Scientific Reports* 12(1): 14599.

Neuner P, Haeussler S (2021). Rule based workload control in semiconductor manufacturing revisited. *International Journal of Production Research* 59(19): 5972-5991.

Oosterman B, Land M, Gaalman G (2000). The influence of shop characteristics on workload control. *International Journal of Production Economics* 68(1): 107-119.

Philipoom P R, Malhotra M K, Jensen J B (1993). An evaluation of capacity sensitive order review and release procedures in job shops. *Decision Sciences* 24(6): 1109-1134.

Plossl G, Wight O W (1973). Capacity planning and control. *Production and Inventory Management* 14(3): 31-67.

Portioli-Staudacher A, Tantardini M (2012). A lean-based ORR system for non-repetitive manufacturing. *International Journal of Production Research* 50(12): 3257-3273.

Pürgstaller P, Missbauer H (2012). Rule-based vs. optimisation-based order release in workload control: A simulation study of a MTO manufacturer. *International Journal of Production Economics* 140(2): 670-680.

Sabuncuoglu I, Karapinar H Y (2000). A load-based and due-date-oriented approach to order review/release in job shops. *Decision Sciences* 31(2): 413-447.

Soepenberg G D, Land M J, Gaalman G J C (2012). A framework for diagnosing the delivery reliability performance of make-to-order companies. *International Journal of Production Research* 50(19): 5491-5507.

Stevenson M, Hendry L C, Kingsman B G (2005). A review of production planning and control: The applicability of key concepts to the make-to-order industry. *International Journal of Production Research* 43(5): 869-898.

Thürer M, Land M J, Stevenson M, Fredendall L D, Godinho Filho M (2015). Concerning workload control and order release: The pre-shop pool sequencing decision. *Production and Operations Management* 24(7): 1179-1192.

Thürer M, Qu T, Stevenson M, Maschek T, Godinho Filho M (2014). Continuous workload control order release revisited: An assessment by simulation. *International Journal of Production Research* 52(22): 6664-6680.

Thürer M, Stevenson M, Silva C, Land M J, Fredendall L D (2012). Workload control and order release: A lean solution for make-to-order companies. *Production and Operations Management* 21(5): 939-953.

Wiendahl H P, Glaessner J, Petermann D (1992). Application of load-oriented manufacturing control in industry. *Production Planning & Control* 3(2): 118-129.

Wight O (1970). Input/output control a real handle on lead time. *Production and Inventory Management* 11(3): 9-31.

Yan H, Stevenson M, Hendry L C, Land M J (2016). Load-oriented order release (LOOR) revisited: Bringing it back to the state of the art. *Production Planning & Control* 27(13): 1078-1091.

**Mingze Yuan** is a master student in management science and engineering at the School of Management, Jinan University (Guangzhou, China). Her main research interests include workload control, order release mechanisms for high variety make-to-order shops and Industry 4.0.

**Ting Qu** is a full professor at the School of Intelligent Systems Science and Engineering, Jinan University (Zhuhai, PR China). He received his Ph.D. degree from the Department of Industrial and Manufacturing Systems Engineering of The University of Hong Kong. His research interests include IoT-based smart manufacturing systems, logistics and supply chain management, and industrial product/production service systems. He has undertaken over twenty research projects funded by government and industry and has published nearly 200 technical papers in these areas, half of which have appeared in reputable journals. He serves as director or board member for several academic associations in industrial engineering and smart manufacturing.

**Matthias Thürer** is professor at Chemnitz University of Technology, Germany. He contributed to the improvement, simplification and integration of material flow control systems, and their integration with Industry 4.0. Apart from operations management, Matthias is also interested in social and philosophical issues including system theory, cybernetics, causality, and philosophy of science.

**Lin Ma** is a postdoctoral fellow in the Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong (PR China). He holds a Master from Xi'an University of Science and Technology, Xi'an (PR China) and a PhD from Jinan University, Guangzhou (PR China). Production Bottleneck Management for high variety make-to-order shops is one of his main research interests, he is also interested in intelligent manufacturing, including digital twins and production-logistics synchronization. He has published seven papers in journals such as the International Journal of Production Research, Production Planning & Control.

**Lei Liu** is currently a Ph.D. student in Management Science and Engineering at the School of Management, Jinan University (Guangzhou, PR China). He received a Master degree in industrial engineering from Chongqing University of Post and Telecommunications (Chongqing, PR China) in 2019 and a bachelor degree in manufacturing and automation from Northwest A&F University (Yangling, PR China) in 2014. Lei Liu worked as a big data R&D engineer from 2019 to 2020 at GREE Electric Appliances, a domestic appliance manufacturer and the world's largest air conditioner producer. His research interests include data and knowledge-driven operation decision making and optimisation for complex industrial systems, intelligent algorithms, and intelligent manufacturing for production-logistics synchronisation, production workload control, modular manufacturing systems and digital twins.