Flexible Services and Manufacturing Journal (2023) 35:487–508
https://doi.org/10.1007/s10696-021-09440-9

Balancing earliness and tardiness within workload control order release: an assessment by simulation

Stefan Haeussler[1] · Philipp Neuner[1] · Matthias Thürer[2]

Accepted: 15 December 2021 / Published online: 3 January 2022

Abstract

Most Workload Control literature assumes that delivery performance is determined by tardiness related performance measures only. While this may be true for companies that directly deliver to end-customers, for make-to-stock companies or firms that are part of supply chains, producing early often means large inventories in the finished goods warehouse or penalties incurred by companies downstream in the supply chain. Some earlier Workload Control studies used a so-called time limit, which constrains the set of jobs that can be considered for order release, to reduce earliness. However, recent literature largely abandoned the time limit since it negatively impacts tardiness performance. This study revisits the time limit, assessing the use of different adaptive policies that restrict its use to periods of either low or high load. By using a simulation model of a pure job shop, the study shows that an adaptive policy allows to balance the contradictory objectives of delaying the release of orders to reduce earliness and to release orders early to respond to periods of high load as quick as possible. Meanwhile, only using a time limit in periods of high load was found to be the best policy.

Keywords Workload control · Order release · Job shop · Simulation

✉ Philipp Neuner
philipp.neuner@uibk.ac.at

Stefan Haeussler
stefan.haeussler@uibk.ac.at

Matthias Thürer
matthiasthurer@workloadcontrol.com

[1] Department of Information Systems, Production and Logistics Management, University of Innsbruck, 6020 Innsbruck, Austria

[2] School of Intelligent Systems Science and Engineering, Jinan University, Zhuhai Campus, 519070 Zhuhai, People's Republic of China

1 Introduction

This study uses simulation to explore the impact of Workload Control order release in production contexts where earliness is a constraint. The motivation for this study arose from a project that sought to improve process flow within one plant of a leading paint manufacturer in the Pearl River Delta of the People's Republic of China. The main problem of the company was limited space in the finished goods warehouse given products were produced too early. The manufacturer produces a high variety of different products on a make-to-order basis. It has seven production bases and more than 2500 employees, realizing a net annual sales income of approximately 2.3 billion RMB. While the present study was originally motivated by this single company, the problem of balancing earliness and tardiness within Workload Control order release mechanisms is also highly relevant for make-to-stock manufacturers (Hutter et al. 2018).

Workload Control is a production planning and control concept which was developed for high-variety make-to-order contexts, but can also be applied to make-to-stock contexts (Zaepfel and Missbauer 1993b; Stevenson et al. 2005; Hutter et al. 2018). A key production control function of Workload Control is order release, which reduces queues and waiting times by imposing a cap on the workload released to the shop floor (Mezzogori et al. 2021). When order release control is applied, jobs do not enter the shop floor directly. Instead, they are retained in a pre-shop pool and released using criteria that allow the shop to achieve certain performance targets whilst limiting the workload on the shop floor.

A common assumption in the Workload Control literature is that orders are delivered as soon as they are completed. Delivery performance is consequently mainly evaluated by tardiness related performance measures, such as the percentage tardy or the mean tardiness; see, for example, the review of performance measures used in the Workload Control Literature in Hendry et al. (2013). It is assumed that the primary objective of Workload Control is to reduce tardiness, while earliness is largely neglected. Earliness may even be considered positive for the customer. Yet, while this may be true for some companies in a make-to-order context that directly deliver to an end-customer, in a make-to-stock or a supply chain context, delivering orders early may create either unwanted finished goods inventory or unwanted supply for the downstream company. In this context, there are often costs for both earliness and tardiness.

Only some of the earlier papers on Workload Control have taken earliness into account (e.g., Zaepfel and Missbauer 1993a; Wiendahl 1995). The focus of these papers is mainly on the Load Oriented Order Release mechanism (Bechte 1988) where earliness is usually controlled by altering the so-called time limit. The idea is that only jobs whose planned release date is earlier than the current time plus the time limit can be released. So, the set of orders that can be considered for release is constrained by the time limit, which avoids too early release of orders. However, the use of the time limit diminished over the years and in recent years most studies consider an infinite time limit (e.g. Land 2006; Thürer et al. 2012).

The main reason for neglecting the time limit is its negative impact on tardiness performance, as for example shown in Land (2006). There is an apparent contradiction between the need to delay the release of orders to reduce earliness and the need to respond to periods of high load as quick as possible (Zaepfel and Missbauer 1993a; Land et al. 2015).

To solve this quandary, this study conceptualizes a switching mechanism to balance these contradictory objectives. It is argued that the use of the time limit should be restricted to either high or low load periods. Both options are feasible. Restricting the time limit to low load periods reduces extreme earliness, which is more likely to occur in these periods. Meanwhile, restricting the time limit to periods of high load reduces earliness whilst ensuring a quicker response to the high load situations characteristic of these periods. Discrete event simulation will be used to assess different policies for the use of the time limit. The objective is to identify the adaptive policy that strikes the best balance between earliness and tardiness, and to give guidance on how to set parameters to realize a specific trade-off for a company. This provides management a tool to determine the trade-off and it provides guidance on how to create the best trade-off.

Finally, while the focus is on Workload Control order release, findings are also of importance to other production control systems. For example, Constant Work in Process (Spearman et al. 1990; Jaegler et al. 2018; Liberopoulos 2020), for which previous literature considered two parameters, a limit on the WIP and a work-ahead window (Jodlbauer and Huber 2008), which is equivalent to a time limit. After a review of literature in Sect. 2, the simulation model used to assess performance will be described in Sect. 3. This includes a description of the release method being considered and the different time limit policies applied. Results are then presented and discussed in Sect. 4 before final conclusions are provided in Sect. 5 together with limitations, and future research directions.

2 Literature review

This section shortly reviews how the literature on Workload Control order release took earliness and tardiness into account and outlines the research question that motivates this study.

Given its importance, a broad literature has emerged to assess the performance of Workload Control order release methods both through simulation (e.g. Perona and Portioli 1998; Cigolini and Portioli-Staudacher 2002; Thürer et al. 2012, 2016; Portioli-Staudacher and Tantardini 2012; Fernandes et al. 2014; Neuner and Haeussler 2020; Neuner et al. 2020; Kundu et al. 2021) and in practice (e.g. Hendry et al. 2013; Silva et al. 2015; Hutter et al. 2018). Most of this literature focused on showing how order release can improve tardiness performance. Yet in practice both earliness and tardiness may be a concern, and inventory control theory and order release models that use optimization models (Silver et al. 1998; Zipkin 2000; Haeussler et al. 2020; Haeussler and Netzer 2020) also typically consider earliness and tardiness.

Only some of the earlier papers on Workload Control have taken earliness into account (e.g., Bechte 1988; Zaepfel and Missbauer 1993a; Wiendahl 1995). These papers propose a so-called time limit to control earliness, i.e., only jobs whose planned release date is earlier than the current time plus the time limit can be released. But the time limit leads to a quandary (Zaepfel and Missbauer 1993a): in case of low capacity utilization the time limit should be low to avoid high finished goods inventory, but in case of high capacity utilization the time limit should be high since this is expected to enable production smoothing, i.e. shifting workload from high load to low load periods. Although Zaepfel and Missbauer (1993a) conjecture that a dynamic time limit might resolve this quandary, they do not provide any guidance on how to conceptualize a dynamic time limit.

In a different context, Land et al. (2015) introduce a dynamic planning approach to guide output control decisions. They argue that control methods should speed up the average lead time during periods of high load, whilst tending to reduce the standard deviation of lateness in low load periods. Land et al. (2015) find that it is this switching behaviour that explains the superior performance of Modified Operation Due Date dispatching (MODD; e.g. Baker and Kanet 1983) and showed that the same switching behaviour can be used for guiding capacity adjustments. In this study this switching behaviour is adopted and it is assessed whether it can be used to resolve the trade-off between earliness and tardiness. In other words, the use of the time limit is limited to either high or low load periods. This leads to the following research question (RQ):

RQ: Can a dynamic time limit reduce earliness without compromising tardiness?

To answer this question, the performance of Workload Control order release under different adaptive and static time limit policies is explored by use of discrete event simulation.

3 Methodology

Although motivated by a real-world company, the study does not rely on a detailed model that accurately represents all aspects of this real production system. In this regard, small systems provide a better insight into the role of operating variables, and, in practice, large systems can often be decomposed into several smaller systems (Bokhorst et al. 2004). The models by Melnyk and Ragatz (1989) and Oosterman et al. (2000) provide a generalized model that matches the main aspects of the real production system of interest. While any individual high-variety shop in practice will certainly differ from the considered stylized model, it captures the high routing variability, processing time variability, and arrival variability that defines this context in practice. Using this model also ensures comparability with earlier Workload Control research that often used this kind of hypothetical job shop model (Paredes et al. 2021).

Section 3.1 first describes how the shop was modelled before the Workload Control order release mechanism applied is outlined in Sect. 3.2. The different policies for the use of the time limit are then outlined in Sect. 3.3, before Sect. 3.4 introduces the dispatching rule used to prioritize jobs on the shop floor. Finally, Sect. 3.5 summarizes the experimental set-up and the main performance measures considered.

3.1 Shop and job characteristics

The simulation model has been implemented in FlexSim. The shop consists of six stations, where each station is a single, constant capacity resource. The routing length of jobs varies uniformly from one to six operations and is determined before the routing sequence is randomly generated without replacement. Thus, re-entrant flows are not considered in this paper. Operation processing times follow a truncated 2-Erlang distribution with a mean of 1 time unit after truncation and the maximum is 4 time units. The inter-arrival time of jobs to the shop follows an exponential distribution with a mean of 0.648 time units, which deliberately results in 90% utilization. The due dates are set exogenously by adding a uniformly distributed random allowance factor to the job entry time which was set arbitrarily between 28 and 38 time units.

3.2 Workload control order release and time limit policies

Several order release approaches are described in the Workload Control literature; for examples, see the reviews by Wisner (1995), Land and Gaalman (1996), Bergamaschi et al. (1997), and Fredendall et al. (2010). In this paper, the LUMS COR (Lancaster University Management School Corrected Order Release) method is used given its good performance in pure job shops (e.g. Thürer et al. 2012). LUMS COR uses both a periodic and a continuous element. The periodic release procedure keeps the workload released to a station within a pre-established workload norm as follows.

A pool sequencing rule first sorts all jobs in the set of jobs J in the pre-shop pool according to the planned release dates (PRDs). If a time limit is applied, then the set of jobs J is restricted by only including jobs that fall within the time limit, i.e. for which $PRD \leq current\ time + time\ limit$. Jobs are then selected from the set J for release as follows:

(1) The first job to be considered for release is the job with the earliest PRD.
(2) If for all operations in the routing of job j, j's load contribution to a station together with the workload already released and yet to be completed fits within the workload norm at this station, then the job is selected for release. That means it is removed from J and its load contributions are included into the station loads. Otherwise, the job remains in the pool and its load contributions are not added to the station loads. A released job contributes to a station load until its operation at this station is completed. The load contribution to a station is therefore 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).
(3) If the set of jobs J in the pool contains any jobs that have not yet been considered for release, then return to Step 1 and consider the job with the next highest priority. Otherwise, the release procedure is finished, and the selected jobs are released to the shop floor.

In addition to the periodic release mechanism, LUMS COR incorporates a continuous workload trigger. If the direct load 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 to avoid premature station idleness (see, e.g. Kanet 1988; Land and Gaalman 1998). However, similar to the periodic release element, a job is only released if its planned release date is within the time limit if a time limit is applied.

As in previous simulation studies on Workload Control (e.g. Land and Gaalman 1998; Fredendall et al. 2010; Thürer et al. 2012, 2017; Fernandes et al. 2017), it is assumed that all materials are available, and all necessary information regarding shop floor routings, processing times, etc. is known. Arriving jobs flow directly into the pre-shop pool and await release according to the release procedure described. Seven workload norms – from 4 to 10 time units – are considered. Finally, the PRD of a job is given by its due date minus an allowance for the operation throughput time for each operation in its routing. It is calculated as follows (e.g., Thürer et al. 2012):

$$\text{PRD} = \text{Due Date} - (\text{Total Processing Time} + \text{Routing Length} * \text{k}) \tag{1}$$

For all experiments, k was set to 2 time units based on preliminary simulation experiments.

3.3 Workload control order release and time limit policies

Two adaptive policies are tested: "Underload", which applies the time limit only in low load periods; and "Overload", which applies the time limit only in high load periods. As a benchmark, also two static policies are tested: "Never", which applies an unlimited time limit; and "Always", which always applies the time limit. The adaptive policies are inspired by the so-called Schmitt-Trigger which was invented by Otto Schmitt (Schmitt 1938). This trigger was developed in the field of electronics and converts an input signal into a high or a low output depending on whether the input signal exceeded an upper bound or dropped below a lower bound, as illustrated in Fig. 1.

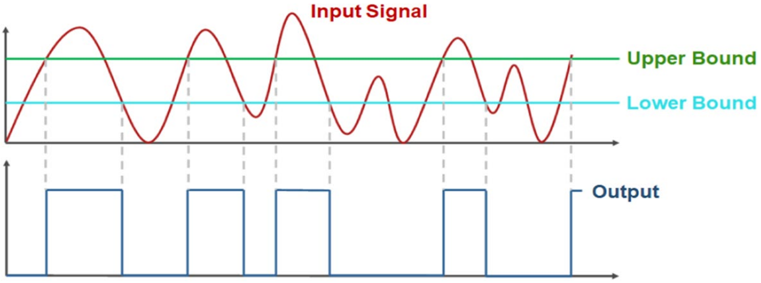


Fig. 1 Functionality of the Schmitt Trigger (Schmitt 1938)

Different to Fig. 1, this study does not differentiate between an upper and lower bound but as in Land et al. (2015) only uses one bound, namely the "load level" to discriminate between low load and high load periods. Preliminary experiments have shown that using a second bound does not significantly influence the performance of the adaptive time limit policies. The load level is measured as total corrected shop load, which is defined as the sum of the corrected station loads (Oosterman et al. 2000).

Note the difference between the Always and adaptive time limit policies (Underload and Overload): While the Always policy restricts the set of jobs in the pre-shop pool by only including jobs that fall within the time limit, a continuous review of the corrected shop load is required for the Underload and Overload policies. Let's assume that, at the beginning of the periodic release procedure, the corrected shop load is lower than the load level and the respective time limit is applied. Jobs in the pre-shop pool are then released until the workload norm of at least one station is violated. If during the periodic release procedure, the load level is exceeded, the time limit is changed according to the underlying logic of the respective adaptive time limit policy. Thus, for Underload, the time limit might be switched from tight to unlimited, and for Overload, it might change from unlimited to a tight time limit.

3.4 Shop floor dispatching rule

Once released, jobs enter the shop floor and move to the queue at the first station in their routing. Jobs in queues are prioritized according to Operation Due Dates (ODDs). The ODD for the last operation in the routing of a job is equal to the due date, while the ODD 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 ODDs are calculated similar to the PRD's:

$$\text{ODD} = \text{Due Date} - (\text{Remaining Processing Time} + \text{Remaining Operations} * \text{k}) \tag{2}$$

Again, k was set to 2 time units.

3.5 Experimental design and performance measures

The experimental factors are: (i) the four different time limit policies ("Never", "Always" "Underload" and "Overload"), (ii) the five different load levels to discriminate between low and high load periods (from 15 to 35 time units); (iii) the 5 different time limits (from 10 to 30 time units); and, (iv) the seven levels of the workload norm (from 4 to 10 time units). A full factorial design was used for each different time limit policy. This results in 392 scenarios as follows: 7 workload norms (Never) + 5 time limits × 7 workload norms (Always) + 5 load levels × 5 time limits × 7 workload norms (for both Underload and Overload). Each scenario was replicated 100 times, while results were collected over 10,000 time units following a warm-up period of 3,000 time units.

Delivery performance will be measured by: the percentage tardy, that is the percentage of jobs completed after the due date; the mean tardiness, that is $T_j = \max(0, L_j)$, 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, that is $E_j = \min(0, L_j)$. In order to evaluate load balancing performance, also the mean throughput time, that is the mean of the completion date minus the pool entry date across jobs, and the average shop floor throughput time are measured. While the throughput time includes the pool time (i.e., the time that an order waits before being released), the shop floor throughput time only measures the time after an order has been released to the shop floor.

4 Results

This section is structured into four parts: First, a statistical analysis summarizes the general findings and highlights the interactions between the time limit policies and the performance measures, second a detailed performance assessment is provided. Third, a sensitivity analysis is conducted and finally the section is concluded by a discussion of the results.

4.1 Statistical analysis

To obtain first insights into the impact of the tested experimental factors, a statistical analysis has been conducted by using an ANOVA (Analysis of Variance). The ANOVA is based on a block design, which is typically used to account for known sources of variation in an experiment. In the ANOVA, the workload norm is treated as blocking factor. This allows for capturing the main effects of this factor and the main and interaction effects of the three time limit related factors (time limit policy, time limit size, and load level for discriminating between low and high load periods). The main and interaction effects are presented in Table 2. All main effects, two-way and three-way interactions were shown to be statistically significant at $\alpha = 0.05$.

The Scheffé multiple-comparison procedure was used to further prove the significance of the differences between the results of the different time limit policies. Results in Table 1 highlight the quandary that this study seeks to solve. On the one hand, the Never time limit policy performs best in terms of percentage tardy orders and mean tardiness, but worst in terms of mean earliness. On the other hand, the Always time limit policy yields the highest percentage of tardy orders and mean tardiness but performs best in terms of mean earliness. As expected, the performance of the two adaptive time limit policies lies somewhere in-between these two extremes, with the Overload policy resulting in the best balance between reducing earliness and tardiness.

Another noteworthy result is that the adaptive Overload policy results in the lowest shop floor throughput time and the second lowest values for the total throughput

Table 1 Multiple comparisons for the different time limit policies

Time limit policy (x)	Time limit policy (y)	% Tardy		Tardiness		Earliness		Shop floor throughput time		Throughput time	
		Lower[a]	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Never	Always	−11.05	−9.85	−1.17	−0.93	4.10	4.37	−0.10*	0.04	−5.52	−5.05
	Underload	−7.15	−6.04	−0.75	−0.53	2.98	3.24	−0.34	−0.20	−3.97	−3.54
	Overload	−2.33	−1.22	−0.29	−0.07	0.52	0.78	0.26	0.40	−1.05	−0.61
Always	Never	9.85	11.05	0.93	1.17	−4.37	−4.10	−0.04*	0.10	5.05	5.52
	Underload	3.59	4.12	0.35	0.46	−1.19	−1.06	−0.28	−0.21	1.43	1.64
	Overload	8.40	8.94	0.82	0.92	−3.65	−3.53	0.32	0.39	4.35	4.56
Underload	Never	6.04	7.15	0.53	0.75	−3.24	−2.98	0.20	0.34	3.54	3.97
	Always	−4.12	−3.59	−0.46	−0.35	1.06	1.19	0.21	0.28	−1.64	−1.43
	Overload	4.66	4.97	0.43	0.49	−2.50	−2.43	0.58	0.62	2.87	2.99
Overload	Never	1.22	2.33	0.07	0.29	−0.78	−0.52	−0.40	−0.26	0.61	1.05
	Always	−8.94	−8.40	−0.92	−0.82	3.53	3.65	−0.39	−0.32	−4.56	−4.35
	Underload	−4.97	−4.66	−0.49	−0.43	2.43	2.50	−0.62	−0.58	−2.99	−2.87

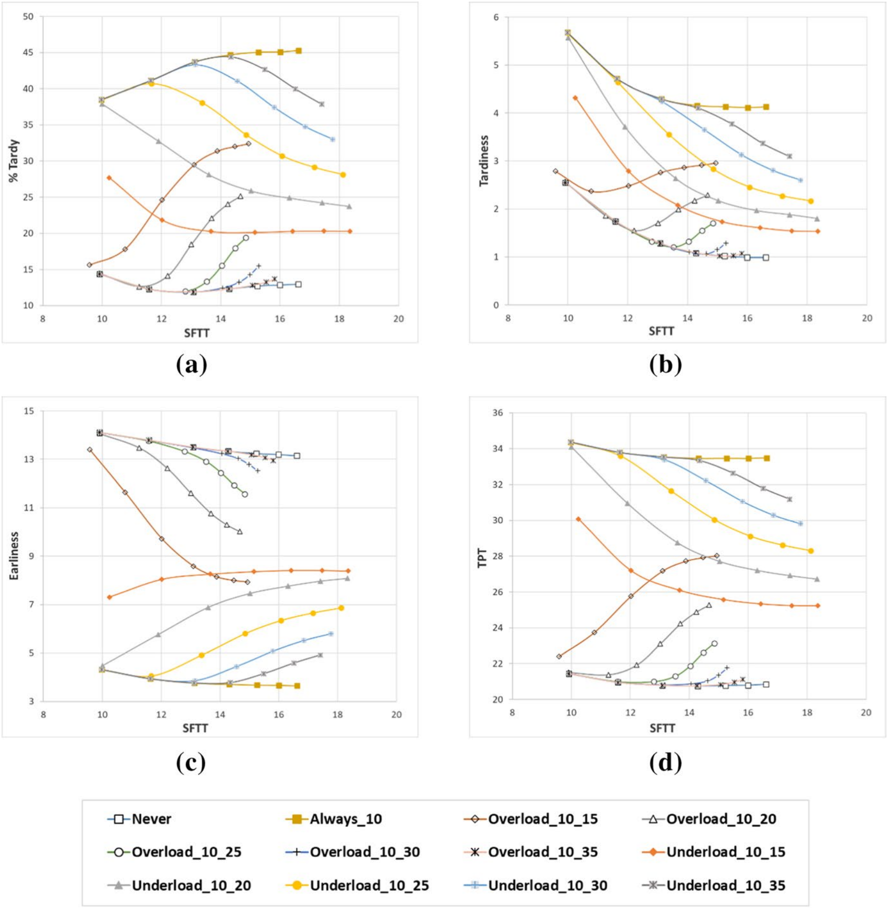
*not significant at $\alpha = 0.05$

[a] 95% confidence interval

times. This means that this policy improves load balancing. In comparison to Never, the Overload policy (i) releases orders later, (ii) orders spend less time on the shop floor, and (iii) orders spend less time in the finished goods inventory. A more detailed performance assessment is presented in the following subsection.

4.2 Detailed performance assessment

Results are presented in the form of performance curves. For brevity, and since the qualitative effect across the different time limit sizes was equivalent, only the performance curves for a time limit of 10 time units are presented. Performance curves for a time limit of 20 time units are given in Fig. 4, whilst all remaining results



(a)

(b)

(c)

(d)

Fig. 2 Performance curves for the different time limit policies and a time limit of 10 time units. **a** Percentage Tardy Jobs vs. Shop Floor Throughput Time, **b** Tardiness vs. Shop Floor Throughput Time, **c** Earliness vs. Shop Floor Throughput Time, and **d** Throughput Time vs. Shop Floor Throughput Time

can be found in a data repository under: http://dx.doi.org/10.17632/4n83dt4tzm.2. Figures 2a through 2d show the percentage of tardy orders (% Tardy), mean tardiness (Tardiness), mean earliness (Earliness) and throughput times (TPT) on the vertical axes and the shop floor throughput times (SFTT) on the horizontal axes. The left-hand starting point of the curves represents the tightest workload norm (4 time units) and the workload norm increases stepwise by moving from left to right in each graph, with each data point representing one workload norm (from 4 to 10 time units). Loosening the workload norm increases the workload on the shop floor and, as a result, the shop floor throughput times. Basically, in all figures the preferred direction is towards the origin of the coordinate system. Meanwhile the different curves for Underload and Overload represent the different settings of the load level to discriminate between low load and high load periods.

For brevity, a single component is used to denote the Never scenarios, a double to denote the Always scenarios and a triple to denote the Underload and Overload scenarios. The first component represents the application of the time limit: “Never”, “Always”, only in “Underload” (i.e. low load) periods, or only in “Overload” (i.e. high load) periods. The second component indicates the time limit, which is 10 time units for all scenarios presented in Fig. 2. The third and last component represents the load levels used to indicate low and high load periods that trigger the switching behavior for the two adaptive policies (Underload and Overload).

Figures 2a–d clearly show the two extreme sides of the two static approaches, Never and Always, which the study seeks to balance with the adaptive approaches (Underload and Overload). Never yields the lowest and Always the highest total throughput time for a given workload norm. This is quite intuitive since introducing a time limit leads to later releases of orders. While the Never time limit policy yields low mean tardiness and percentage tardy it yields the highest earliness and thus might tie up a significant amount of capital in the finished goods inventory. Contrary to that, the Always policy results in the exact opposite performance by yielding low earliness but the worst tardiness performance. These results are in line with earlier studies analysing the effect of static time limits (e.g., Land 2006; Zaepfel and Missbauer 1993a).

Figures 2a–d also show that, with decreasing workload norms, the performance curves of the two adaptive order release mechanisms (Overload and Underload) approach the two static mechanisms (Never and Always) respectively. As a consequence, some Underload scenarios are equivalent to the respective Always scenario since the load level is seldom reached and the time limit (almost) never switched to unlimited, and some Overload scenarios are equivalent to the Never scenario since again the load level is seldom reached, but in this case, the policy (almost) never switches to applying the time limit. Similarly, increasing the time limit (see Appendix B and http://dx.doi.org/10.17632/4n83dt4tzm.2) reduces the gap between the two static mechanisms Never and Always, since Always approaches the Never policy which is unaffected by the time limit. Thus, the time limit parameter determines the range within which the adaptive policies balance the trade-off between earliness and tardiness where the Never and Always policies represent the two extremes.

With regard to balancing earliness and tardiness by using the two adaptive policies, Figs. 2a, b show that Underload results in similar or lower percentages of tardy orders and tardiness compared to Always, but higher percentages of tardy orders and tardiness than Never. Likewise, Overload yields similar or higher percentages of tardy orders and tardiness compared to Never but lower percentages of tardy orders and tardiness than Always. Focusing on earliness, which is depicted in Fig. 2c, Overload yields similar or lower earliness compared to Never but higher earliness than Always. Finally, Fig. 2d shows the balancing performance of the tested policies. While the Underload policy yields similar or higher shop floor throughput times compared to the Never and Always time limit policy, the Overload policy is able to improve the load balancing performance since, for a given workload norm, the respective Overload scenarios lead to lower shop floor throughput times compared to all other time limit policies.

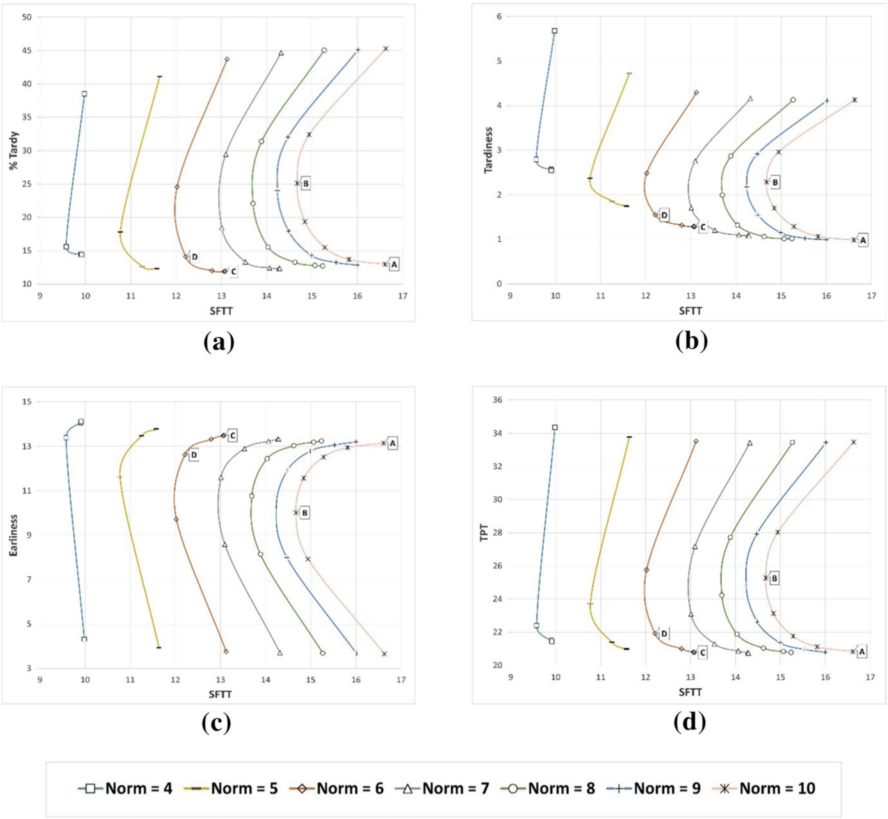
4.3 Sensitivity analysis

This section assesses the performance impact of setting different load levels to discriminate between low and high load periods. This sensitivity analysis is limited to the Overload policy since, overall, it yields the best performance of the two tested adaptive policies. A time limit of 10 time units is used, since it provides a reasonable and illustrative bandwidth between the two extremes Never and Always.

Figures 3a–d show the balancing and timing performance of the two static approaches Never and Always and the dynamic Overload policy. Here, each curve shows the different scenarios among the policies for a given workload norm level. Since there are seven settings for the workload norm, there are seven curves in each figure. The uppermost point of each of the seven curves in Figs. 3a, b and d represents the respective Always scenario and in Fig. 3c the corresponding Never scenario. Likewise, the lowermost point of each curve in Figs. 3a, b and d indicates the respective Never scenario and in Fig. 3c the corresponding Always scenario.

When moving from one specific Never scenario to the related Always scenario, the load level is stepwise reduced from 35 to 15 time units in steps of 5 time units, i.e. there are five related Overload scenarios. There are consequently seven data points that constitute each of the seven curves. For lower workload norms, some Overload scenarios are equivalent to the respective Never scenario since the load level is seldom reached. Thus, it seems that for lower workload norms less data points are shown in Figs. 3a through 3d, however, these data points are just overlapping.

Nevertheless, once one has determined the time limit parameter such that a preferable bandwidth between the two extremes Never and Always is reached (in this case 10 time units), the next step is to balance the trade-off between earliness and tardiness by step wise reducing the load level parameter. For a given norm level and starting with the Never policy, lowering the load level results in lower earliness but higher tardiness measures. Which load level is set depends on one's preferences with



(a)

(b)

(c)

(d)

Fig. 3 Sensitivity analysis for the different time limit policies and a time limit of 10 time units. **a** Percentage Tardy Jobs vs. Shop Floor Throughput Time, **b** Tardiness vs. Shop Floor Throughput Time, **c** Earliness vs. Shop Floor Throughput Time, and **d** Throughput Time vs. Shop Floor Throughput Time

regard to earliness and tardiness, while Overload scenarios at the same time have a positive effect on the load balancing performance.

For example, consider a company that applies a workload norm of 10 time units and no time limit, i.e. the Never scenario indicated by position A in Figs. 3a through 3d. In order to balance the trade-off between earliness and tardiness, the company decides to introduce a time limit using an adaptive policy. Moving along the curve leads them to their preferred load level for discriminating between low and high load periods – 20 time units, which corresponds to position B. This move from A to B enabled the company to reduce its earliness at the detriment of higher tardiness. However, this also helped the company to achieve the lowest shop floor throughput time at their specified workload norm level of 10 time units. The same holds when moving from C to D along the curve for a workload norm of 6 time units. But here the trade-off is more conservatively balanced. Nevertheless, it shows that given the workload norm of a company, reducing the load level that identifies low and high load periods allows to balance the trade-off between earliness and tardiness, while allowing for an improved load balancing performance.

4.4 Discussion of results

Applying an adaptive time limit allows for reducing earliness whilst limiting its detrimental effect. While it may appear more reasonable to only use the time limit in low load periods, when extreme earliness is more likely to occur, the results show that the time limit should only be applied during high load periods. An explanation for the worse performance of the Underload policy can be found in the self-reinforcing effect, which in system dynamics is known as a positive feedback loop (see Sterman 1989, 2000). Using the Underload policy, in low load periods the release quantities are restrained by the time limit since only the most urgent jobs in the pre-shop pool are considered for potential release. But this leads to a build-up of jobs in the pre-shop pool. As soon as demand increases and the system enters a high load period, the Underload policy switches to an unlimited time limit and all jobs in the pre-shop pool are considered for release. This "shock" of orders leads to higher release quantities and thus a higher workload compared to the Overload policy. On the contrary, the adaptive Overload policy smooths demand peaks by producing in advance since it switches to an unlimited time limit in low load periods. In other words, the pre-shop pool is cleared in low load periods such that urgent orders arriving during upcoming high load periods fit within the given workload norms and can thus be timely released. Meanwhile earliness is reduced in high load periods, where the large set of jobs in the pool means that many jobs fall into the time limit, which in turn allows for load balancing.

5 Conclusions

The literature on Workload Control order release typically focuses on tardiness for measuring delivery performance. This means that companies with high holding costs or limited space in their finished goods inventory find little to no guidance in the literature whether to implement Workload Control, and how to implement Workload Control. Only some earlier studies take earliness into account by using a so-called time limit, which constrains the set of jobs that can be considered for release, to improve earliness (Zaepfel and Missbauer 1993a; Wiendahl 1995). However, earlier studies also have shown that the time limit negatively impacts tardiness performance and consequently its use has been widely abandoned (Land 2006).

This study revisits the time limit. Based on the ideas of Zaepfel and Missbauer (1993a) and Land et al. (2015), it conceptualizes a switching mechanism that limits the use of the time limit to either low or high load periods. In response to the research question – *Can a dynamic time limit reduce earliness without compromising tardiness?* – simulation results show that the dynamic time limit policies allow to balance the contradictory objectives of delaying the release of orders to reduce earliness and releasing orders early to respond to periods of high load as quick as possible. Meanwhile, only using a time limit in periods of high load was found to be the best policy because it smooths demand peaks by producing in advance. It thus not only reduces earliness but also allows for a better load balancing.

5.1 Managerial implications

Many shops in practice have some constraint on earliness. For example, make-to-order companies that have limited space in their warehouse, make-to-stock manufacturers, and companies that are part of larger supply chains where products must be delivered just-in-time. In this context, producing early is often related to higher holding costs or penalties incurred by companies downstream the supply chain. The study at hand shows that Workload Control including an adaptive time limit is also applicable to these companies. The best performing policy (Overload) did not only balance earliness and tardiness, but also yielded the best load balancing performance. This simple extension to Workload Control order release makes the concept and its performance benefits more applicable to a broader range of companies.

5.2 Limitations and future research

This study provides important insights for research and practice, but also has its limitations. First, the results are limited to the simulated production environment, and the validity of the results for flow shops, or unbalanced shops, such as semiconductor fabs, should be assessed in future studies. Second, adding further experimental factors, such as different demand patterns, or including different scheduling or pool sequencing rules, would be beneficial. Furthermore, future studies could focus on optimizing the parameters, such as the load levels for given workload norms, or setting load levels and time limits adaptively by using some machine learning algorithm; see e.g. Schneckenreither and Haeussler (2019) or Schneckenreither et al. (2020) for a first step in this promising future research direction.

Appendix A

See Table 2

Table 2 . Main and interaction effects

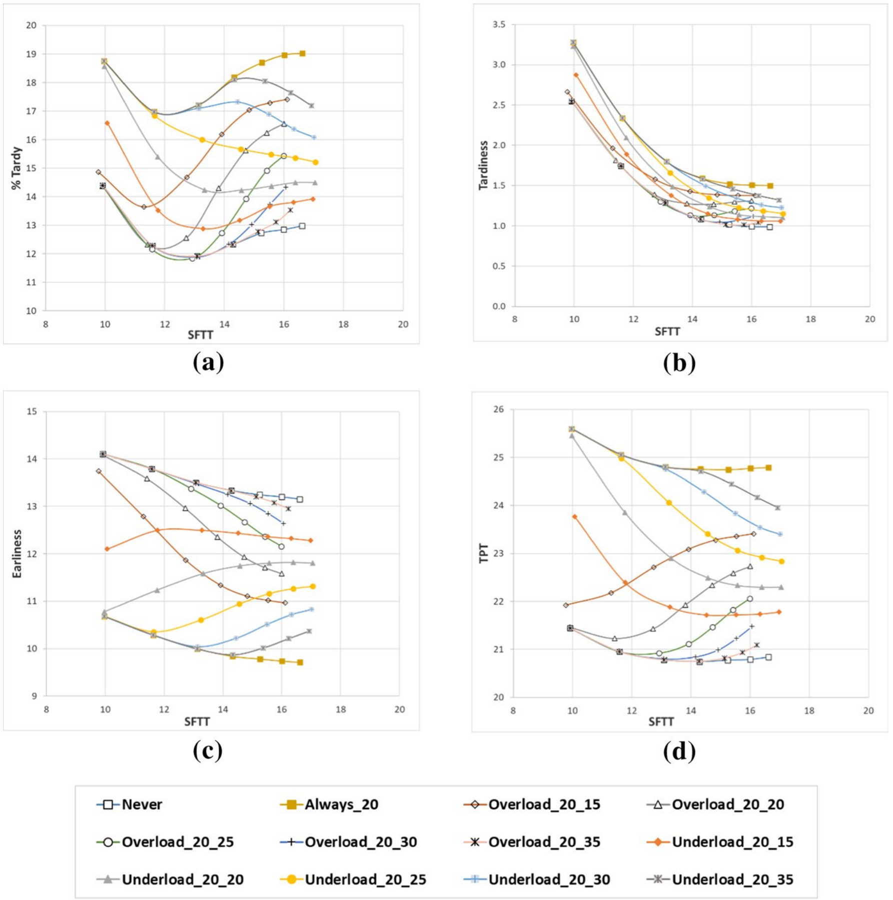
	Source of variance	Sum of squares	Degree of freedom	Mean squares	F-ratio	*p*-Value
Tardiness	Norm	13,674.32	6	2279.05	2204.67	0.00
	Time Limit Policy (TLP)	1862.23	1	1862.23	1801.45	0.00
	Time Limit (TL)	10,929.70	4	2732.43	2643.25	0.00
	Load Level (LL)	53.15	4	13.29	12.85	0.00
	TLP * LL	1519.57	4	379.89	367.49	0.00
	TLP * TL	2423.96	4	605.99	586.21	0.00
	TL * LL	78.13	16	4.88	4.72	0.00
	TLP * TL * LL	1592.33	16	99.52	96.27	0.00
	Residual	40,458.46	39,138	1.03		
Earliness	Norm	2757.09	6	459.52	324.37	0.00
	Time Limit Policy (TLP)	53,185.92	1	53,185.92	37,543.52	0.00
	Time Limit (TL)	125,861.71	4	31,465.43	22,211.20	0.00
	Load Level (LL)	365.92	4	91.48	64.58	0.00
	TLP * LL	18,964.31	4	4741.08	3346.69	0.00
	TLP * TL	43,394.57	4	10,848.64	7657.97	0.00
	TL * LL	196.07	16	12.25	8.65	0.00
	TLP * TL * LL	10,630.53	16	664.41	469.00	0.00
	Residual	55,444.73	39,138	1.42		

Table 2 . (continued)

	Source of variance	Sum of squares	Degree of freedom	Mean squares	F-ratio	*p*-Value
Percentage Tardy Orders	Norm	13,199.78	6	2199.96	82.37	0.00
	Time Limit Policy (TLP)	202,802.54	1	202,802.54	7593.50	0.00
	Time Limit (TL)	1,137,881.18	4	284,470.29	10,651.37	0.00
	Load Level (LL)	6098.23	4	1524.56	57.08	0.00
	TLP * LL	15,7719.34	4	39,429.84	1476.36	0.00
	TLP * TL	286,041.50	4	71,510.38	2677.55	0.00
	TL * LL	8691.70	16	543.23	20.34	0.00
	TLP * TL * LL	166,665.00	16	10,416.56	390.03	0.00
	Residual	1,045,274.16	39,138	26.71		
Throughput Time	Norm	4369.11	6	728.19	176.45	0.00
	Time Limit Policy (TLP)	74,951.81	1	74,951.81	18,161.93	0.00
	Time Limit (TL)	209,226.17	4	52,306.54	12,674.65	0.00
	Load Level (LL)	695.92	4	173.98	42.16	0.00
	TLP * LL	31,185.16	4	7796.29	1889.16	0.00
	TLP * TL	65,896.53	4	16,474.13	3991.92	0.00
	TL * LL	486.38	16	30.40	7.37	0.00
	TLP * TL * LL	20,025.18	16	1251.57	303.27	0.00
	Residual	161,517.21	39,138	4.13		

Table 2 . (continued)

	Source of variance	Sum of squares	Degree of freedom	Mean squares	F-ratio	*p*-Value
Shop Floor Throughput Time	Norm	195,103.87	6	32,517.31	81,901.07	0.00
	Time Limit Policy (TLP)	3159.14	1	3159.14	7956.90	0.00
	Time Limit (TL)	8.48	4	2.12	5.34	0.00
	Load Level (LL)	24.20	4	6.05	15.24	0.00
	TLP * LL	480.34	4	120.08	302.45	0.00
	TLP * TL	2062.48	4	515.62	1298.69	0.00
	TL * LL	14.36	16	0.90	2.26	0.00
	TLP * TL * LL	450.48	16	28.16	70.91	0.00
	Residual	15,539.02	39,138	0.40		



(a)

(b)

(c)

(d)

Fig. 4 Performance curves for the different time limit policies and a time limit of 20 time units. **a** Percentage Tardy Jobs vs. Shop Floor Throughput Time, **b** Tardiness vs. Shop Floor Throughput Time, **c** Earliness vs. Shop Floor Throughput Time, and **d** Throughput Time vs. Shop Floor Throughput Time

Appendix B

Figures 4a through 4d show the results for a time limit of 20 time units. One can see that the relations, described in Sect. 4, also hold under a longer time limit, but in this case the performance difference between the Never and the Always time limit policy (representing the lower and upper bound for the considered performance measures) decreases. Nevertheless, the Underload and Overload policy still yield intermediate timing measures (i.e. percentage of tardy orders, mean tardiness and mean earliness) compared to the Never and Always policy, and the Overload policy still yields lower shop floor throughput times compared to all other policies. However, the performance deteriorates even more under the longer time limit when workload norms are increased.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10696-021-09440-9.

Acknowledgements The authors are grateful to the two anonymous reviewers for reading the manuscript carefully and providing constructive comments which greatly helped to improve this paper. The authors would like to thank Reha Uzsoy, Hubert Missbauer, as well as participants of the 21st International Working Seminar on Production Economics that were part of the workshop: 'Workload Control – Quo Vadis?' for very valuable suggestions that helped us to improve the paper. Finally, we want to acknowledge Julian Eckhart for conducting the initial simulation study that showed the potential of adaptive Workload Control policies.

Authors' contributions All authors contributed to the study conception and design, and data collection and analysis. The first draft of the theoretical part of the manuscript was mainly written by Matthias Thürer, while Philipp Neuner and Stefan Haeussler focused more on the results section. All authors commented on previous versions of the manuscript and read and approved the final manuscript.

Funding Open access funding provided by University of Innsbruck and Medical University of Innsbruck. This work was supported by the National Natural Science Foundation of China [Grant Number 71872072; 72150610504].

Availability of data and material The datasets analysed during the current study are available in the following Mendeley data repository: http://dx.doi.org/10.17632/4n83dt4tzm.2

Code availability Can be provided upon request by the corresponding author.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Baker KR, Kanet JJ (1983) Job shop scheduling with modified due dates. J Oper Manag 4(1):11–22

Bechte W (1988) Theory and practice of load-oriented manufacturing control. Int J Prod Res 26(3):375–395

Bergamaschi D, Cigolini R, Perona M, Portioli A (1997) Order review and release strategies in a job shop environment: a review and a classification. Int J Prod Res 35(2):399–420

Bokhorst JCA, Slomp J, Gaalman GJC (2004) On the who-rule in dual resource constrained (DRC) manufacturing systems. Int J Prod Res 42(23):5049–5074

Cigolini R, Portioli-Staudacher A (2002) An experimental investigation on workload limiting methods within ORR Policies in a job shop environment. Prod Plann Control 13(7):602–613

Fernandes NO, Land MJ, Carmo-Silva S (2014) Workload control in unbalanced job shops. Int J Prod Res 52:679–690

Fernandes NO, Thürer M, Silva C, Carmo-Silva S (2017) Improving workload control order release: incorporating a starvation avoidance trigger into continuous release. Int J Prod Econ 194:181–189

Fredendall LD, Ojha D, Patterson JW (2010) Concerning the theory of workload control. Eur J Oper Res 201(1):99–111

Haeussler S, Netzer P (2020) Comparison between rule-and optimization-based workload control concepts: a simulation optimization approach. Int J Prod Res 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. Int J Prod Econ 227:107682

Hendry LC, Stevenson M, Huang Y (2013) Workload control: successful implementation taking a contingency-based view of production planning & control. Int J Oper Prod Manag 33(1):69–103

Hutter T, Haeussler S, Missbauer H (2018) Successful implementation of an order release mechanism based on workload control: a case study of a make-to-stock manufacturer. Int J Prod Res 56(4):1565–1580

Jaegler Y, Jaegler A, Burlat P, Lamouri S, Trentesaux D (2018) The ConWip production control system: a systematic review and classification. Int J Prod Res 56(17):5736–5757

Jodlbauer H, Huber A (2008) Service-level performance of MRP, kanban, CONWIP and DBR due to parameter stability and environmental robustness. Int J Prod Res 46(8):2179–2195

Kanet JJ (1988) Load-limited order release in job shop scheduling systems. J Oper Manag 7(3):44–58

Kundu K, Land MJ, Portioli-Staudacher A, Bokhorst JA (2021) Order review and release in make-to-order flow shops: analysis and design of new methods. Flex Serv Manuf J 33:750–782

Land M (2006) Parameters and sensitivity in workload control. Int J Prod Econ 104(2):625–638

Land MJ, Gaalman GJC (1996) Workload Control concepts in job shops: a critical assessment. Int J Prod Econ 46–47:535–548

Land MJ, Gaalman GJC (1998) The Performance of workload control concepts in job shops: improving the release method. Int J Prod Econ 56–57:347–364

Land MJ, Stevenson M, Thürer M, Gaalman GJC (2015) Job shop control. In search of the key to delivery improvements. Int J Prod Econ 168:257–266

Liberopoulos G (2020) Comparison of optimal buffer allocation in flow lines under installation buffer, echelon buffer, and CONWIP policies. Flex Serv Manuf J 32:297–365

Melnyk SA, Ragatz GL (1989) Order review/release: Research issues and perspectives. Int J Prod Res 27(7):1081–1096. https://doi.org/10.1080/00207548908942609 (1989).

Mezzogori D, Romagnoli G, Zammori F (2021) "Defining accurate delivery dates in make to order job-shops managed by workload control." Flex Serv Manuf J, in press.

Neuner P, Haeussler S (2020) Rule based workload control in semiconductor manufacturing revisited. Int J Prod Res. https://doi.org/10.1080/00207543.2020.1797208

Neuner P, Haeussler S, Ilmer Q (2020) "Periodic workload control: a viable alternative for semiconductor manufacturing." Proceedings of the 2020 winter simulation conference, in press.

Oosterman B, Land MJ, Gaalman G (2000) The influence of shop characteristics on workload control. Int J Prod Econ 68(1):107–119

Paredes FJG, Godinho Filho M, Thürer M, Fernandes NO, Jabbour CJC (2021) "Factors for choosing production control systems in make-to-order shops: a systematic literature review." J Intel Manuf, in press

Perona M, Portioli A (1998) The impact of parameter setting in load oriented manufacturing control. Int J Prod Econ 55:133–142

Portioli-Staudacher A, Tantardini M (2012) A lean-based ORR system for non-repetitive manufacturing. Int J Prod Res 50(12):3257–3273

Schmitt OH (1938) A thermionic trigger. J Sci Instrum 15:24–26

Schneckenreither M, Haeussler S (2019) Reinforcement learning methods for operations research applications: the order release problem. In: Nicosia G, Pardalos P, Giuffrida G, Umeton R, Sciacca V (eds) Machine learning, optimization, and data science. LOD 2018. Lecture notes in computer science, vol 11331. Springer, Cham. https://doi.org/10.1007/978-3-030-13709-0_46

Schneckenreither M, Haeussler S, Gerhold C. (2020) "Order release planning with predictive lead times: a machine learning approach." Int J Prod Res, in press.

Silva C, Stevenson M, Thürer M (2015) A case study of the successful implementation of workload control. J Manuf Technol Manag 26(2):280–296

Silver E, Pyke DG, Peterson R (1998) Inventory management and production planning and scheduling. Wiley, New York

Spearman ML, Woodruff DL, Hopp WJ (1990) CONWIP: a pull alternative to Kanban. Int J Prod Res 28(5):879–894

Sterman J (1989) Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. Manage Sci 35(3):321–339

Sterman J (2000) business dynamics: systems thinking and modeling for a complex world with CD-ROM. McGraw-Hill Education.

Stevenson M, Hendry LC, Kingsman BG (2005) A review of production planning and control: the applicability of key concepts to the make to order industry. Int J Prod Res 43(5):869–898

Thürer M, Stevenson M, Silva C, Huang G (2012) The Application of workload control in assembly job shops: an assessment by simulation. Int J Prod Res 50:5048–5062

Thürer M, Stevenson M, Qu T (2016) Job sequencing and selection within workload control order release: an assessment by simulation. Int J Prod Res 54:1061–1075

Thürer M, Land MJ, Stevenson M, Fredendall LD (2017) On the integration of due date setting and order release control. Prod Plann Control 28(5):420–430

Wiendahl HP (1995) Load oriented manufacturing control. Springer, Berlin

Wisner JD (1995) A review of the order release policy research. Int J Oper Prod Manag 15(6):25–40

Zaepfel G, Missbauer H (1993a) Production planning and control (ppc) systems including load-oriented order release—problems and research perspectives. Int J Prod Econ 30–31:107–122

Zaepfel G, Missbauer H (1993b) New concepts for production planning and control. Eur J Oper Res 67:297–320

Zipkin PH (2000) Foundations of inventory management. McGraw-Hill, Boston

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Stefan Haeussler is currently Associate Professor at the Department of Information Systems, Production and Logistics Management at the University of Innsbruck. His main research focus is on quantitative methods for decision support in the areas of production planning and supply chain management. His research focuses on order release, lead time management, dispatching and their practical application. Methodologically, he focuses on discrete event simulation, optimization, economic experiments and machine learning methods.

Philipp Neuner is currently working as research assistant at the Department of Information Systems, Production and Logistics Management at the University of Innsbruck. He received his M.Sc. degree in Information Systems from the University of Innsbruck in 2019 and is currently studying for his PhD degree in Management at the University of Innsbruck. His research interests include manufacturing planning and control, simulation modeling, optimization and workload control.

Matthias Thürer is Distinguished Professor in Management Science and Engineering at Jinan University (Zhuhai, PR China). Before getting involved in academia, Matthias worked in several companies, did an apprenticeship and became a master craftsman ('Meister'). 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.