Order review and release in make-to-order flow shops

Kundu, Kaustav; Land, Martin J.; Portioli-Staudacher, Alberto; Bokhorst, Jos A. C.

Published in:
Flexible Services and Manufacturing Journal

DOI:
10.1007/s10696-020-09392-6

Publication date:
2021

Flexible Services and Manufacturing Journal (2021) 33:750–782
https://doi.org/10.1007/s10696-020-09392-6

Order review and release in make-to-order flow shops: analysis and design of new methods

Kaustav Kundu[1] · Martin J. Land[2] · Alberto Portioli-Staudacher[1] · Jos A. C. Bokhorst[2]

Published online: 19 July 2020

Abstract

Increased customization has strengthened the importance of make-to-order companies. The advent of lean management and the introduction of smart and flexible technologies has enabled many of these companies to create flow shop routings. Order review and release (ORR) research, which originally focused on job shops, has started paying attention to flow shops. However, the results have not provided clarity on the best ORR method for flow shops. This study aims at developing such a method by applying a modular design approach. It identifies the relevant elements of ORR methods for flow shops, combines them into new methods and evaluates them in a simulation study. The simulation results demonstrate that performance in pure flow shops can be strongly improved by applying the right combination of workload measures, load balancing, and order dispatching. Specifically, the results show that (1) classical workload measures are still as effective as novel measures that have been suggested for flow shops, (2) balancing workloads explicitly through optimization at the order release stage strongly improves performance, and (3) shortest processing time dispatching is highly effective in flow shops as it avoids starvation of stations. In-depth analyses have been executed to unravel the reasons of performance improvements. As such, the article provides clarity on the improvement potential that is available for ORR in flow shops, while the new modular methods provide a first step in exploiting this potential.

Keywords Flow shop · ORR · Workload balancing · Workload limiting · Simulation

✉ Alberto Portioli-Staudacher
alberto.portioli@polimi.it

Extended author information available on the last page of the article

1 Introduction

This paper aims at developing order review and release (ORR) methods for flow shops with variable processing times. Due to changes in customer needs and increased competition from low cost countries, more and more companies are strategically focusing on customization. This leads to low volume/high variety make-to-order production with highly variable processing times. Originally, this type of production has been addressed by adopting job shop configurations to accommodate the variety and variability, while order review and release (ORR), a popular production planning and control technique, has been used extensively to manage the performance of such a complex configuration (Baykasoğlu and Göçken 2011). New technologies and digitalization are leading to increased flexibility of systems (Theorin et al. 2017), and the advent of lean management and quick response manufacturing is fostering a product based view rather than a resource based one (Krishnamurthy and Suri 2009; Fullerton et al. 2014). This stimulates and enables companies to streamline their production processes, so products can be manufactured adopting a fixed sequence of more flexible resources. Consequently, we are witnessing a new generation of companies characterized by make-to-order production with high processing time variability in flow shop configurations.

In the field of production planning and control, flow shop literature has strongly focused on scheduling problems (Das and Khumawala 1991; Lin and Chen 2015), whilst research on control decisions such as ORR, has lagged behind. Ooosterman et al. (2000) were among the first to apply principles of workload control in ORR in shop configurations that included pure flow shops. They tested several workload measures and identified that a measure indicated as aggregate workload performed best in flow shops. Later, the results of Portioli-Staudacher and Tantardini (2012) suggested that a new workload measure would be more appropriate for flow shops. They designed an ORR method for flow shops including several novel elements. Besides the alternative workload measure, their method explicitly aimed at balancing the workloads among workstations, rather than limiting workloads at every release period. Their way of balancing allowed overloads for workstations if that would lead to a more evenly distributed workload among all workstations. The method was shown to strongly outperform the traditional workload limiting methods, but as multiple changes were combined in this method, the reasons behind the better performance were not clearly unfolded. More recently, Fernandes et al. (2020) showed that optimization-based order release has the potential to improve performance compared to workload limiting methods in flow shops. Their optimization method can be regarded as a mixed method with limiting and balancing properties. It aims at minimizing the differences between the released but not yet processed workload and a pre-established workload norm for all workstations, while violations of workload norms are not allowed. This raises the question how it compares to the balancing method used in Portioli-Staudacher and Tantardini (2012). Furthermore, Portioli-Staudacher and Tantardini (2012) combined their release method with a simple first-come-first-served

(FCFS) dispatching rule, arguing that the performance is less dependent on the dispatching rule when the order release rule is chosen well. Nevertheless, Thürer et al. (2015a) found that prioritizing the orders based on shortest processing times (SPT) is advantageous in flow shops with high variability in demand and processing times, particularly at upstream workstations to avoid starvation at more downstream workstations. Similarly, Fernandes et al. (2020) used FCFS dispatching as a baseline and showed that Modified Operation Due Dates (MODD) dispatching, combining SPT and Operation Due Dates, resulted in improvements for all performance measures under unrestrictive order release. To conclude, contradictory results in flow shop literature do not provide clarity on (1) the impact of new flow shop oriented load measures, (2) the advantages of load balancing compared to load limiting and (3) the effectiveness of combining controlled release methods with starvation avoiding dispatching rules such as SPT.

This research aims to better understand the functioning of ORR in flow shops and to develop new ORR methods that best suit the needs of flow shops with high processing time variability. By creating a modular design of ORR methods, this is the first study to identify the specific elements that should be combined in such a method for flow shops. The study focuses on periodic—daily—release decisions, which fit common company practices but have received less attention in recent ORR studies. The periodicity enables the release of optimized combinations of orders. Simulation based optimization methods in manufacturing have received attention in the recent literature (Fernandes et al. 2020; Lin et al. 2019; Lin and Chen 2015). With the advances in computing technology, the use of optimization techniques for planning and control decisions such as ORR will only increase in the near future.

This paper is organized as follows. The theoretical background is provided in Sect. 2, which analyses existing literature and identifies the elements of ORR methods that need further testing in flow shops. Section 3 describes and motivates the setup of the flow shop simulation model and outlines the main experimental design for which the identified elements of ORR methods are translated into a modular design of ORR methods, including two new ORR methods. It also specifies several extensions of the main experimental design to provide more in-depth insights and discusses the performance measures used in the simulation study. Section 4 presents the findings and further assesses and discusses the critical elements that are combined in the best performing ORR method. In Sect. 5, conclusions are drawn, along with managerial implications and recommendations for future research.

2 Theoretical background

The literature on ORR methods is reviewed with a particular focus to identify those elements that are potentially important in developing periodic ORR methods for pure flow shops with variable processing times. With reference to Bergamaschi et al. (1997), we confine ourselves to commonly used periodic atemporal ORR systems with input control and workloads measured in processing time units. We assume that the flow shop consists of a single production unit. Another category of ORR systems explicitly models the ORR decisions for individual

production units within a larger company by multi-period optimization models. We refer to Haeussler et al. (2020) and Missbauer (2020) for recent studies in this specific research stream and to Haeussler and Netzer (2019) for a comparison with the ORR systems addressed in this paper.

Section 2.1 starts by discussing workload measurement, i.e. workload aggregation approaches. Workload balancing and schedule visibility in ORR methods are discussed in Sects. 2.2 and 2.3, respectively. Next, order dispatching is discussed in Sect. 2.4. Finally, our research objectives are stated in detail in Sect. 2.5.

2.1 Workload aggregation

Recording the actual workload of each workstation in order to determine the new orders that can be released is a key element of load-based ORR methods. According to Oosterman et al. (2000), there are three relevant aggregation levels in measuring a workstation's workload. The direct workload reflects the load in front of the workstation, the aggregate workload additionally includes the load of the station that has been released but that still has to pass upstream workstations, and the shop workload further incorporates load already completed but not out of the shop yet. The simulation study of Oosterman et al. (2000) showed that none of the traditional methods based on the three levels in workload calculation is suitable for all shop floor configurations. Whereas direct loads play an important role in job shops, more focus on control of upstream loads is required in pure flow shops due to the impossibility to quickly affect the direct workloads of downstream stations. The study revealed that a focus on direct workload even led to an undesirable impact of order releases in flow shops. The methods using limits for aggregate workloads tested by Oosterman et al. (2000) were shown to perform the best in pure flow shops, and will therefore be used in the development of our proposed new rule.

While traditional aggregation methods include an order in the workload of workstation until completion of the operation at that workstation, Portioli-Staudacher and Tantardini (2012) suggested that the focus in flow shops should be on balancing only the workload that has been released recently. They do not consider the workload on the shop floor that already passed the first workstation. We will refer to this as the released workload. In the development of a new ORR method we will investigate both aggregation approaches, specified as follows:

The aggregate workload L_{kt}^{A} of the *kth* workstation at time t is defined by Eq. (1).

$$L_{kt}^{A} = \sum_{j} p_{jk} I(t)_{\left[t_j^R, t_{jk}^C\right\rangle} \tag{1}$$

with p_{jk}: the processing time of order j at workstation k, t_j^R: the time of release of order j, t_{jk}^C: the time of completion of the *kth* operation of order j, $I(t)_{[\ldots\rangle}$: an indicator function being 1 if t is in the specified interval, 0 elsewhere.

The released workload L_{kt}^{R} of the *kth* workstation at time t is defined by Eq. (2).

$$L_{kt}^{R} = \sum_{j} p_{jk} I(t)_{\left[t_j^R, t_{j1}^C\right\rangle} \tag{2}$$

Thus, an order j is part of the aggregate workload L_{kt}^{A} of a workstation k until the time t_{jk}^{C} when its operation at that specific workstation has been completed and part of the released workload L_{kt}^{R} for each workstation until completion of the operation at the first workstation at time t_{j1}^{C}. This study will evaluate the performance differences resulting from these two workload aggregation approaches.

2.2 Workload balancing in ORR methods

A large part of ORR methods restrict the workload present on the shop floor by releasing a set of orders that fit within pre-specified workload limits for each workstation. By setting these workload limits, the workload of the workstations is expected to become more balanced, since all workstations are supposed to receive a workload close to their limit (Land and Gaalman 1998; Sabuncuoglu and Karapinar 1999; Henrich et al. 2004; Land 2006). More explicitly balancing oriented ORR methods use optimization to balance the workload among workstations. Generally, their objective is to minimize the deviations from a target workload level rather than strictly limiting the workload. Thus, one or more workstations may have slightly higher loads, if doing so reduces the underload of other workstations.

However, the impact of balancing has mainly been researched in job shops. Some classical studies investigating balancing oriented ORR methods in job shops are by Irastorza and Deane (1974), Shimoyashiro et al. (1984), Onur and Fabrycky (1987), Van Ooijen (1998), Cigolini and Portioli-Staudacher (2002) and a more recent study by Yan et al. (2016). In these job shop studies there is no conclusive evidence that balancing oriented methods perform significantly better than limiting oriented methods.

In pure flow shops there are hardly any studies comparing limiting oriented methods and balancing oriented methods. An exception is the study of Portioli-Staudacher and Tantardini (2012) that compares a new balancing oriented method with one of the most common and best performing load limiting methods. The balancing oriented method outperforms the load limiting method. However, as several changes were combined in their method, the performance improvements cannot be fully attributed to the aspect of workload balancing. The load balancing method in Fernandes et al. (2020) used optimisation to minimize the underload only, which can be regarded as a mixed method containing limiting and balancing elements. They showed that this mixed method could improve the performance compared to load limiting methods in pure and general flow shops.

In this study, we will include the element of workload balancing in the development of a new ORR method in its pure form. The influences of workload balancing (explicit balancing) and workload limiting (implicit balancing) will be contrasted in the evaluation of different ORR methods. Additionally, the impact of not allowing overloads in the balancing objective (the mixed method) will be explored.

2.3 Schedule visibility

In ORR methods, the importance of schedule visibility for workload balancing is highlighted by a number of authors (Van Ooijen 1998; Cigolini and Portioli-Staudacher 2002; Portioli-Staudacher and Tantardini 2012). Extended schedule visibility relates to taking into account the current release period as well as the future ones. The majority of studies investigating extended schedule visibility relate to job shop systems (Cigolini et al. 1998; Van Ooijen 1998; Cigolini and Portioli-Staudacher 2002), whereas Portioli-Staudacher and Tantardini (2012) adopted the use of extended schedule visibility in a pure flow shop and assumed that it is one of the possible explanations for the improvement achieved by their ORR method. However, its effect has not been studied separately.

2.4 Order dispatching

Literature on priority dispatching at workstations is abundant. However, in flow shops orders all start at the same workstation and follow the same route, which means that the first workstation strongly controls the flow of orders through the rest of the shop (Thürer and Stevenson 2016). Portioli-Staudacher and Tantardini (2012) used the FCFS rule to avoid irregularities in the flow at successive steps. Thürer et al. (2015a) found that prioritizing the orders based on shortest processing times (SPT) is advantageous, particularly at upstream workstations to avoid starvation at more downstream workstations. They tested the use of SPT for dispatching at each workstation. This may cause some orders to become extremely late, which is solved in their study by replacing SPT by Modified Operation Due Dates (MODD) dispatching. Fernandes et al. (2020) also included MODD dispatching in their study on optimised order release in flow shops. This study will investigate the sensitivity of the ORR method to the choice of dispatching rule.

2.5 Research objective

Based on the literature, we identified four key elements that need to be considered in the development of a successful ORR method for pure flow shops with periodic release: (1) workload aggregation, (2) workload balancing, (3) schedule visibility, and (4) order dispatching. The aim of our study is to develop an improved ORR method for pure flow shops with periodic release, which will be based on a further elaboration of the four elements mentioned above.

3 Simulation model

A simulation model is built to test the influence of the relevant elements of periodic ORR methods for pure flow shops identified in Sect. 2. The simulation model is developed using the Python 3.4 Simpy module. Gurobi 6.5 is used to solve the optimization model incorporated in the simulation. First, the pure flow shop model

and the parameters used are motivated and described in Sect. 3.1. Then, the main experimental design is outlined in Sect. 3.2 followed by a specification of the ORR methods used in the main experimental design. Extensions of the main experimental design are elaborated in Sect. 3.3. Finally, the measures used to evaluate performance are presented in Sect. 3.4.

3.1 Simulation model and parameters

The pure flow shop model with five workstations of Portioli-Staudacher and Tantardini (2012) is adopted, as this study formed the trigger for most elements in the ORR methods. The maximum capacity of each workstation is 8 h per day. We assume that capacity is restricted by a single resource, contrary to recent studies by Thürer et al. (2019a, b) and Portioli-Staudacher et al. (2020), which focus on dual resource constraints. A Poisson distributed number of orders (mean 15) gets added in the pre-shop pool each day just before the release opportunity. Since exponential interarrival times create a Poisson distribution of the number of arrivals per time unit, this approach leads to the same distribution of the number of daily arrivals as in studies using continuous arrival processes with exponential interarrival times. However, our approach avoids the need for mechanisms responding to intermediate arrivals within a day as in Thürer et al. (2015a). The processing times at each workstation are assumed to be independent and follow a lognormal distribution with a mean of half an hour and a coefficient of variation of 0.8. This implies an average service rate of 16 orders during a day of 8 h. Given the arrival rate of 15 orders per day, this leads to an average utilization of 93.75%. We assume that sufficient space is available between the workstations to avoid blocking, contrary to a recent study of Thürer et al. (2020) that studies limited buffer sizes.

Due date setting is not always a straightforward process in practice and a wide variety of methods has been adopted (Caprihan et al. 2006; Zorzini et al. 2008; Thürer et al. 2019c). In order to keep the due date assignment method simple with a reasonable percentage of tardy orders, a constant delivery time allowance of 6 working days (48 h) is added to the order arrival date to calculate the due date. The constant allowance fits the standard routings in pure flow shops. The value of 6 working days is tuned to realise reasonable percentages tardy orders of around 2.5% for the best performing methods. The availability of new orders for release at the start of each day guarantees that all deadlines are automatically at the end of a working day. The latest release dates are calculated as the difference between the due date and the planned shop floor time. Based on tests in preliminary simulation experiments, the planned shop floor time is set to 3 working days. The period between release decisions is 1 day as in the study of Portioli-Staudacher and Tantardini (2012), who followed Perona and Portioli (1998). The ORR parameters are summarized in Table 1.

3.2 Experimental design

The main experimental design focuses on two of the four elements that we consider relevant in the development of a successful ORR method for pure flow shops

Table 1 Parameters and their values

Parameters	Values
Number of workstations	5
Demand distribution	Poisson ($\lambda = 15$)
Processing times	Lognormal (mean = 0.5, cv = 0.8)
Due date allowance	6 working days
Planned shop floor time	3 working days
Release period length	1 working day

with periodic release (see Sect. 2): workload aggregation and workload balancing. The balancing methods require long computational times, since an optimization model has to be solved before every successive release period in the simulation. Therefore, initial experiments have been executed to pre-test the factors and keep the final experimental design sufficiently compact. As hardly any performance differences could be observed for schedule visibility, this element has been excluded from the study. Furthermore, since SPT dispatching resulted in the best performance, this rule has been included as a fixed factor in the main experimental design and will be further explored in an extended analysis. As common in studies which focus on controlled release, the simulations are run with different workload targets, which depend on the ORR method that is used. Based on the above considerations, three experimental factors are included in the main experimental design, which is summarized in Table 2.

The combinations of the experimental factors 'workload aggregation' and 'workload balancing' result in four distinct ORR methods, which are further specified in Sects. 3.2.1 and 3.2.2 below. We aim at comparing the ORR methods at similar levels of workload reduction. Since for each method a different workload target value is required to realize this, it is not possible to define the tightness of the target in terms of the workload target values themselves. Instead, the realized average shop floor time is used as an intermediate variable and two target workload levels in different methods are assumed to be equally tight, if they result in the same average shop floor time. This is a common approach in nearly all recent studies on workload control (e.g. Thürer et al. 2014, 2015b, 2017; Fernandes et al. 2016; Yan et al. 2016; Fernandes et al. 2020). For each method, ten workload targets have been selected that best expose the range of realizable shop floor times for that method. This results in a full factorial design with 40 experiments.

Table 2 Main experimental design

Experimental Factors	Levels
Workload aggregation	Released workload, aggregate workload
Workload balancing	Limiting, balancing
Workload target level	10 levels leading to 10 shop floor times

This main design has been extended in several ways to deepen the findings and include additional factors that may relate to a specific ORR method. These extensions will be discussed in Sect. 3.3. The following part of this section specifies the ORR methods used in the main experimental design resulting from the combinations of the experimental factors 'workload aggregation' and 'workload balancing'. Section 3.2.1 specifies two existing ORR methods, whereas Sect. 3.2.2 proposes two new ORR methods.

3.2.1 Existing ORR methods

We discuss two existing ORR methods, meant particularly for pure flow shops, as below.

3.2.1.1 Aggregate workload limiting ORR method (AL) In the Aggregate workload limiting (AL) method, the aggregate workload L_{kt}^{A} for each workstation, as defined in Eq. (1), is evaluated. In the early work of Oosterman et al. (2000) this performed best in pure flow shops. In order to control the length of the queues in front of each workstation, workload targets reflecting the desired situation after each release period are set for each workstation. In a pure flow shop, each downstream workstation will need more aggregate workload on average than the preceding workstation, but there is no standard procedure to calculate the workload target for each workstation (Land and Gaalman 1996). In this paper a target workload for the first workstation ϑ_1 will be specified while the target workload of the *kth* workstation ϑ_k will be derived from this according to Eq. (3).

$$\vartheta_k = c + k(\vartheta_1 - c) \tag{3}$$

c is the capacity of a workstation for the period between two release decisions, which is 8 h for each workstation. This approach reflects the logic that, if possible, the first workstation should be loaded up to its capacity until the next release opportunity. Given the balanced capacities, an equal buffer $(\vartheta_1 - c)$ is then planned for each workstation. The release procedure is straightforward. The remaining aggregate workload L_{kt}^{A} is calculated for each workstation k just before release. Then all orders waiting for release in the pre-shop pool are successively considered in sequence of their latest release date τ_j. In our design with constant due date allowances and planned shop floor times, this sequence will be equal to the sequence of arrivals and due dates. If the release of the next order does not cause any workload L_{kt}^{A} to exceed the target ϑ_k, then the order is selected for release and added to the workload before the next order is considered. If any target would be exceeded, the order will wait in the pre-shop pool until the next release opportunity. A formalized algorithm for this stepwise procedure can be found in several articles (e.g. Oosterman et al. 2000; Thürer et al. 2011; Yan et al. 2016).

3.2.1.2 Released workload balancing ORR method (RB) The main purpose of the released workload balancing (RB) method is to balance released workload for each workstation, which is the workload from the moment of release until completion of

the first operation, as defined in Eq. (2). To respond to the shop floor situation, the RB method has to specify the target amounts to be released in a different way. This method follows the same logic as Portioli-Staudacher and Tantardini (2012). For a flow shop of K stations, we define the total workload L_t^T on the shop floor at time t as:

$$L_t^T = \sum_k \sum_j p_{jk} I(t)_{\left[t_j^R, t_{jK}^C\right\rangle} \tag{4}$$

and compare this total workload with a target value θ. The target for the amount of released workload for each workstation is then a *Kth* fraction of the deviation from this target. When the release decision is taken at time t, the load status will be determined just before this time and referred to as $L_{t^-}^T$.

The optimization model uses binary decision variables x_j that indicate whether an order j from the set of orders S in the pre-shop pool is selected for release $(x_j = 1)$ or not $(x_j = 0)$. We define $[y]^- \equiv \max\{0, -y\}$ and $[y]^+ \equiv \max\{0, y\}$ to specify the underload and overload, as in Eqs. (5) and (6) respectively, and minimize their sum according to Eq. (7).

$$U_k^R = \left[\sum_{j \in S} p_{jk} x_j - \frac{\left(\theta - L_{t^-}^T\right)}{K}\right]^- \quad for\ k = 1, \ldots, K \tag{5}$$

$$O_k^R = \left[\sum_{j \in S} p_{jk} x_j - \frac{\left(\theta - L_{t^-}^T\right)}{K}\right]^+ \quad for\ k = 1, \ldots, K \tag{6}$$

$$\text{minimize} \sum_{k=1}^{K} U_k^R + O_k^R \tag{7}$$

In the main design, we follow Cigolini and Portioli-Staudacher 2002) and Portioli and Tantardini (2012), by (1) treating overloads and underloads equally, and (2) forcing the release of orders by setting $x_j = 1$ for orders when the latest release date τ_j has been passed at the time of the release decision.

3.2.2 New ORR methods

Challenging the existing ORR methods and their use of "workload aggregation" and "workload balancing/limiting", we propose two new ORR methods that use a different combination of these factors. These new ORR methods are presented below.

3.2.2.1 Released workload limiting ORR method (RL) Portioli-Staudacher and Tantardini (2012) proposed to use released workloads in combination with workload balancing, whilst we will now test an ORR rule that combines released workload with workload limiting, the Released workload Limiting (RL) method. The release procedure is identical to that of the AL method, but instead of using the

aggregate workloads L_{kt}^{A}, the released workloads L_{kt}^{R} are considered. The released workloads are all compared with the same target ϑ_1, independent of the workstation.

3.2.2.2 Aggregate workload balancing ORR method (AB) The other new ORR rule that will be tested is the Aggregate workload Balancing (AB) method. In this rule ORR has an explicit workload balancing goal, as proposed by Portioli-Staudacher and Tantardini (2012). But rather than balancing released workloads, we will test balancing aggregate workloads. The main objective of this method is to minimize the deviation between the workstations' aggregate workloads over a certain period of time.

The optimization model now defines the underload U_k^A and overload O_k^A of a workstation k after the release decision by Eqs. (8) and (9), with all other variables as defined before.

$$U_k^A = \left[\sum_{j \in S} p_{jk} x_j - \left(\vartheta_k - L_{kt^-}^A\right)\right]^- \quad for\ k = 1, \ldots, K \tag{8}$$

$$O_k^A = \left[\sum_{j \in S} p_{jk} x_j - \left(\vartheta_k - L_{kt^-}^A\right)\right]^+ \quad for\ k = 1, \ldots, K \tag{9}$$

with these definitions translated into constraints, the AB method selects the orders by minimizing the sum of the overloads and underloads (Eq. 10).

$$\text{minimize} \sum_{k=1}^{K} U_k^A + O_k^A \tag{10}$$

If an order exceeds its latest release date, its release is forced to avoid extreme delays in the same way as discussed in Sect. 3.2.1. The load limiting methods AL and RL do not require these forced releases. These methods consider orders for release in sequence of their planned release date, which will automatically increase the release probability of an order as it gets closer to this date.

3.3 Extensions of the main experimental design

Additional experiments have been executed to provide more in-depth insights into mechanisms that explain the performance influences of the main experimental factors. Besides, we made specific choices in the main design of our four ORR methods to be consistent with Portioli-Staudacher and Tantardini (2012). These choices are relaxed in small extensions of the main design. Many of these factors relate to differences between recent studies of ORR in flow shops.

Table 3 summarizes these experimental factors that have been studied in additional experiments and the questions that have been answered.

Table 3 Extensions

Factor	Question
1. Underload/overload	Should loads be balanced around or within targets?
2. Dispatching rule	What is the impact of proposed dispatching rules on ORR?
3. Forced releases	Should release be forced after a latest release date is passed?
4 Sensitivity	How sensitive is performance to applied parameter values?

1. As explained in Sect. 2.2, the balancing method of Portioli-Staudacher and Tantardini (2012) allows both underloads and overloads when minimizing the deviations from a target load. Fernandes et al. (2020) use a mixed balancing and limiting method, by applying workload norms as strict upper bounds in their balancing approach, thus not allowing overloads. To deepen the insights in differences between load limiting and balancing, this mixed approach has been embedded by giving a sufficiently high weight to overloads in the objective functions (7) and (10).
2. To enable a focus on the ORR method a fixed dispatching rule is applied in the main experimental design. This is the relatively simple Shortest Processing Time (SPT) rule, suggested by Thürer et al. (2015a). As discussed in Sect. 2.2, Portioli-Staudacher and Tantardini (2012) applied FCFS, while Fernandes et al. (2020) suggest that further improvements can be realized by the use of a Modified Operation Due Dates. Both alternatives are evaluated in extended experiments.
3. The use of Modified Operation Due Dates for dispatching decisions may further reduce the need for considering urgency at the time of order release. While simply using FCFS for dispatching decisions, Portioli-Staudacher and Tantardini (2012) avoid extreme delays of orders as part in their load balancing ORR method. They force the release of an order as soon as the latest release date τ_j has been passed at the time of the release decision. Fernandes et al. (2020) also deal with urgency as part of the ORR method, by defining a class of urgent orders and a class of non-urgent orders. To test the importance of prioritizing urgent orders at release decisions in combination with different dispatching rules, additional experiments have been executed that restrain from using forced releases of urgent orders.
4. Finally, a sensitivity analysis has been executed, presenting the influence of (a) due date allowances and planned shop floor times, (b) utilization levels and (c) processing time variability.

3.4 Performance measures

Each experiment encompasses 50 replications. Based on initial experiments, the warm-up period is set to 1600 h and the length of each replication is set to 4000 h to overcome the initial transient and guarantee significance. The common random number technique has been used to reduce the variance among experiments. Results will be presented for three different performance variables. These are defined as follows:

Average gross throughput time The gross throughput time is calculated as the time of orders in the system from their arrival till delivery to the customer, which includes the time in the pre-shop pool and the time on the shop floor;
Standard deviation of lateness The lateness of each order is calculated as the difference between the time of completion and due date;
Percentage of tardy orders The percentage of tardy orders is determined as the percentage of orders with a lateness value exceeding zero.

Besides, the average shop floor time is used as an intermediate variable to represent the level of workload reduction realized by the different workload targets. The shop floor time is calculated as the time between the release of the order from the pre-shop pool to the shop floor and the delivery of the order.

Workload balancing related performance is best reflected in the capability to reduce the average gross throughput time at low workload levels (Germs and Riezebos 2010). However, as balancing could be obtained at the cost of reduced attention for the individual urgency of orders, the standard deviation of lateness will be monitored. The combination of both performance aspects should finally result in a low percentage of tardy orders, preferably at low levels of workloads.

4 Results and discussions

This presentation of results starts with a performance overview of the four ORR methods (AL, RB, RL, and AB) as specified in Sects. 3.2.1 and 3.2.2, combined with SPT dispatching. Sections 4.1 and 4.2 focus on the individual influences of workload aggregation and workload balancing, respectively. Within these subsections the findings of the main experimental design are presented first, followed by the extended analysis to deepen the insights into performance differences, and concluded with a discussion of the findings. Section 4.3 presents the results of the extended analysis regarding the effect of the dispatching rule and the forced release constraint. The results of the sensitivity analysis are presented in Sect. 4.4.

Figure 1 shows a performance overview of the four ORR methods. In each graph, the tightness of the workload target is indicated by the resulting average shop floor time, which is set on the horizontal axis. The three performance measures are set on the vertical axis in Fig. 1a–c, respectively. The points resulting from successive target levels are connected to obtain curves for each method.

At the infinite workload target levels the ORR method has no influence, as all the orders are immediately released to the shop floor. This point is positioned at an average shop floor time of around 13.33 h, with a similar average gross throughput time and a percentage tardy of around 2.92%.

Figure 1a shows that the gross throughput time increases for all ORR methods when reducing the average shop floor time (moving from right to left in the graph), with the largest increases for the workload limiting methods AL and RL. The percentage tardy however can be decreased with the workload balancing methods (to around 2% with AB) by setting tighter workload norms.

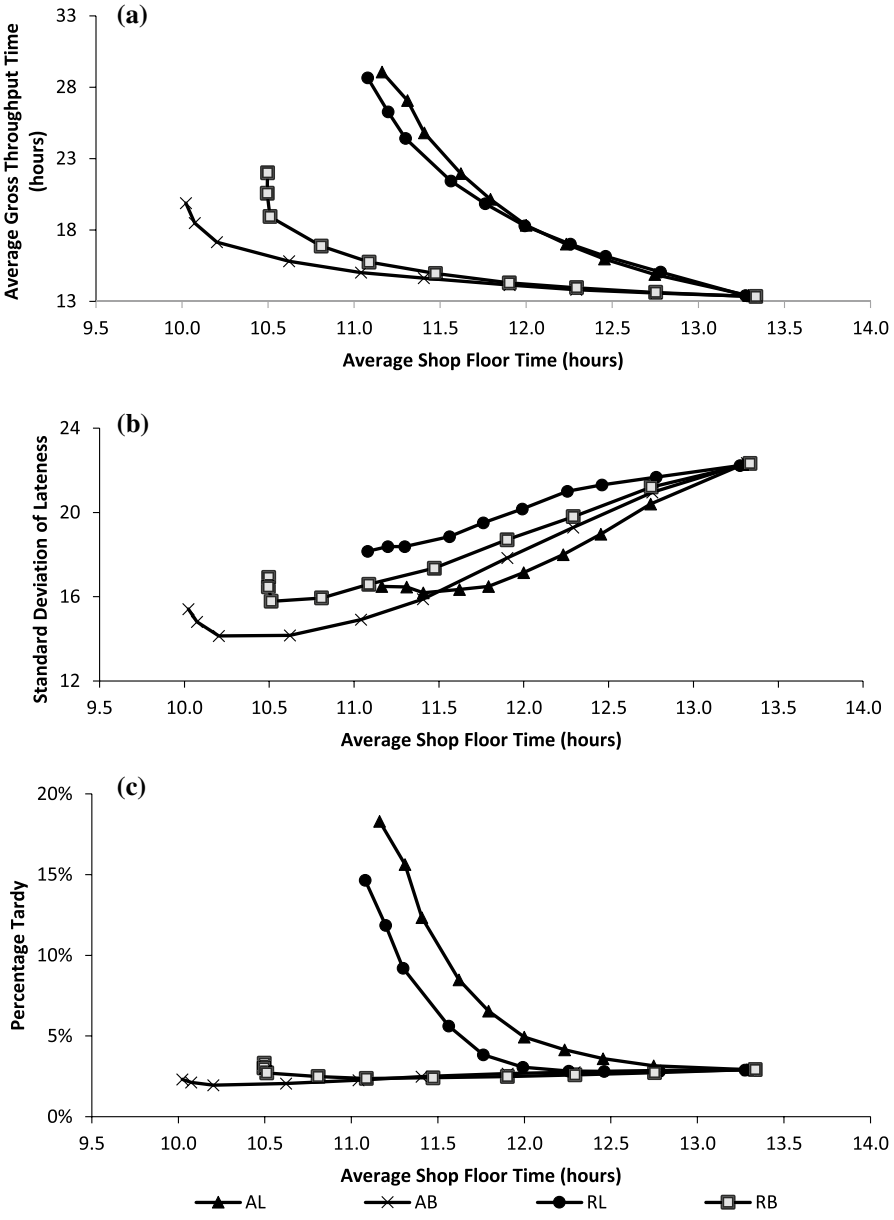

Fig. 1 Performance of the ORR methods

Overall, the AB method, which combines aggregate workloads and explicit workload balancing, leads to the best performing periodic ORR method for the pure flow shop in our main experimental design. The findings of each critical element are highlighted and studied more in-depth in the next subsections.

4.1 Workload aggregation

4.1.1 Main findings

Figure 1 shows that among the load limiting methods AL and RL, the aggregate workload method AL results in a lower standard deviation of lateness, while other measures worsen. However, among the load balancing methods AB and RB, the aggregate workload method AB performs consistently better for all performance measures, and especially at tighter workload targets. This suggests that for load balancing methods, traditional aggregate workload should be preferred over released workload, contrary to the suggestion of Portioli-Staudacher and Tantardini (2012). To understand this effect an extended analysis has been executed.

4.1.2 In-depth explanation and discussion

To explain the advantage of balancing based on aggregate workload, we measure the behaviour of flow times over time at each workstation for the two balancing methods RB and AB. For this purpose, we use the lowest workload target level, where the difference in performance is most striking between the two methods. Figure 2 presents these flow times for respectively RB (upper graphs) and AB (lower graphs) at workstation 1 (left) and workstation 5 (right) for each order during a sample from a single simulation run. The embedded numbers indicate the related coefficient of variation (CV). The figure shows that the flow times and thus also the workloads at the first workstation are less variable when the aggregate workload balancing method is applied, while for the fifth workstation the differences are small. The first

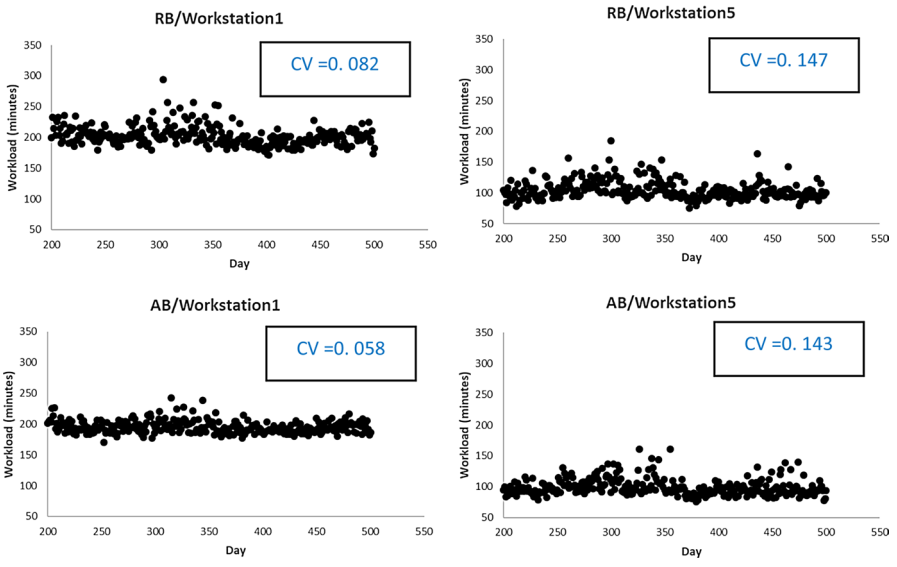

Fig. 2 Flow times at upstream/downstream workstations for RB and AB

workstation is a gateway in pure flow shops. Therefore, both AB and RB have more influence on the first workstation than on the subsequent workstations.

An important notion to proceed our analysis is that the workload measures for AB and RB are identical for the first workstation. As a consequence the disadvantage of the RB method cannot be in the workload measure itself and must be in the threshold with which the newly released workload is compared.

For aggregate workloads, the gap between the remaining aggregate workload and the target workload of each workstation $\left(\vartheta_k - L^A_{kt^-}\right)$ is used to create a threshold for the amount to be released. The threshold for the RB method, based on Portioli-Staudacher and Tantardini (2012), is less straightforward. To realize a feedback loop that forces the release decision to respond to the shop floor situation, Portioli-Staudacher and Tantardini (2012) chose to calculate the average workload gap across workstations. This causes the ORR method to respond strongly to work that is already downstream, while the intention of the method is to avoid responding to the downstream situation. Release decisions can only affect downstream workloads after a long delay.

To test whether this explains the performance difference between AB and RB, the threshold for RB is changed to $\left(\vartheta_i - L^R_{it^-}\right)$ for each workstation individually. This means that the workload gap is determined by considering the workload that resides at the first workstation for each downstream station, rather than averaging the workload in the whole flow shop.

Figure 3 shows the new comparison between AB and RB. The dashed line (RB′) shows the redesigned RB method, applying the new threshold setting. The curve of average gross throughput time (Fig. 3a) for RB′ now approaches the curve of AB more closely. At lower workload target levels the performance for RB′ still deteriorates slightly more. This relates to the fact that when workload target levels are tightened, RB′ loads for downstream stations have less stable levels (CV = 0.184 for workstation 5) than the aggregated load measures (CV = 0.143 for workstation 5; see Fig. 2).

The standard deviation of lateness (Fig. 3b) is highest for RB′. This in turn has interesting consequences for the resulting percentage tardy (Fig. 3c). This measure is now the lowest for RB at the highest workload targets, while it is the lowest for RB′ at some intermediate set of targets, and the lowest for AB at the tightest targets. AB realizes the lowest percentage tardy overall. In general, it shows that the choice of an appropriate threshold in defining the workload target is important when it comes to fine-tuning performance.

4.2 Workload balancing

4.2.1 Main findings

Germs and Riezebos (2010) define the workload balancing capability as the capability to reduce the average gross throughput time at tighter workload targets. However, due to reduced effectiveness of SPT dispatching at lower workload levels, we see in Fig. 1 that all methods show higher gross throughput times when workload targets

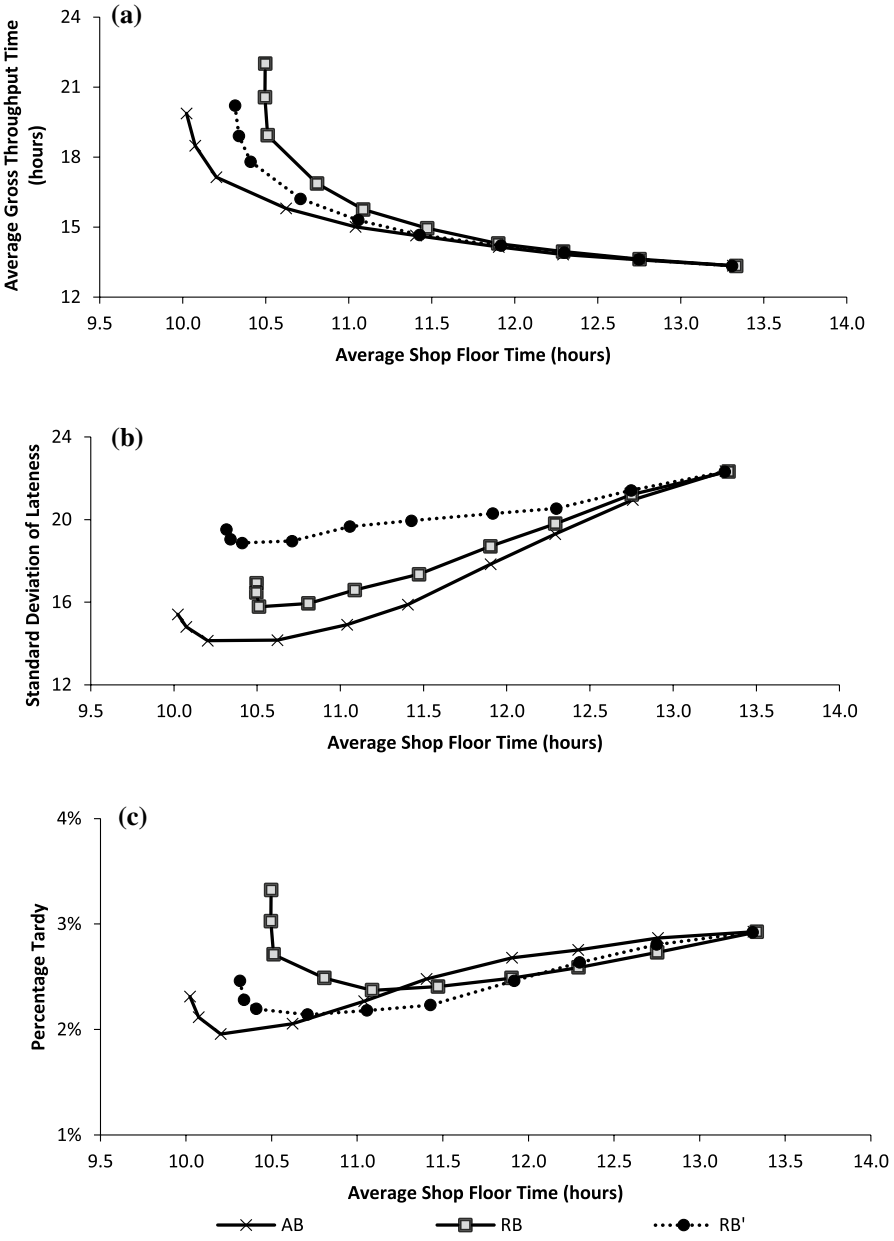

Fig. 3 Performance of AB, RB and RB′ with corrected load targets

get tighter. Instead, we should therefore look at the amount of increase in gross throughput times at low workload targets. Then Fig. 1 shows that the workload limiting methods AL and RL result in longer gross throughput times than the workload balancing methods AB and RB at all levels of workload. The limiting methods thus

have less balancing capabilities, as could be expected. Figure 1 also makes clear that the workload balancing element within the design of an ORR method is responsible for a larger part of the performance difference than the workload aggregation element.

4.2.2 Extended analysis: mixed load limiting and balancing

When workload targets act as a strict upper bound in balancing approaches, we can speak of a mixed balancing and limiting method. Fernandes et al. (2020) use such a mixed method. AB can simply be transformed in a mixed method by giving a sufficiently high weight to overloads in Eqs. (7) and (10). In Fig. 4 the resulting method AM is compared with AB and AL. Figure 4 shows that its gross throughput time performance, as the indicator of balancing capabilities, is better than that of AL, but weaker than AB. The mixed method is not able to reduce the shop floor times to similar levels as AB before performance starts to deteriorate strongly.

Our results confirm the conclusions of Portioli-Staudacher and Tantardini (2012) for flow shops regarding the advantages of explicit load balancing by optimization. This is much more effective than the implicit approach to load balancing that fits orders successively into workload limits (AL). Additionally, the results showed that in our setting the mixed method that minimizes underloads (AM) performs worse than explicit load balancing that allows overloads (AB). However, it is important to understand the background of the strong performance differences. This is addressed in the next subsection.

4.2.3 In-depth explanation and discussion

To explain the differences between the workload limiting and the workload balancing methods we measured the release lateness of each order as the difference between its actual release date and its latest release date, following the definition of Land (2006). Preferably, all orders released together should be equally urgent, as indicated by a similar release lateness. Figure 5 shows the difference in standard deviation of the release lateness for AL, AB, and AM. Figure 5 clearly reveals that, to realise a similar level of the shop floor time, the limiting method combines orders for release that vary much more in release lateness than the balancing method. The performance of the mixed method lies in between. To enable the same reduction of shop floor throughput time, this implies that limiting-oriented methods need a much larger pre-shop pool to combine orders before sufficient balance can be realized. This exactly reflects the essential problem of a weak balancing mechanism.

4.3 Order dispatching and the influence of forced releases for urgent orders

To show the effect of the dispatching rule the experimental design has been extended with a dispatching analysis for the best performing ORR method (AB). Figure 6 compares the SPT rule that we used in the previous experiments with the Modified Operation Due Date (MODD) rule and the First-Come-First-Served (FCFS) rule.

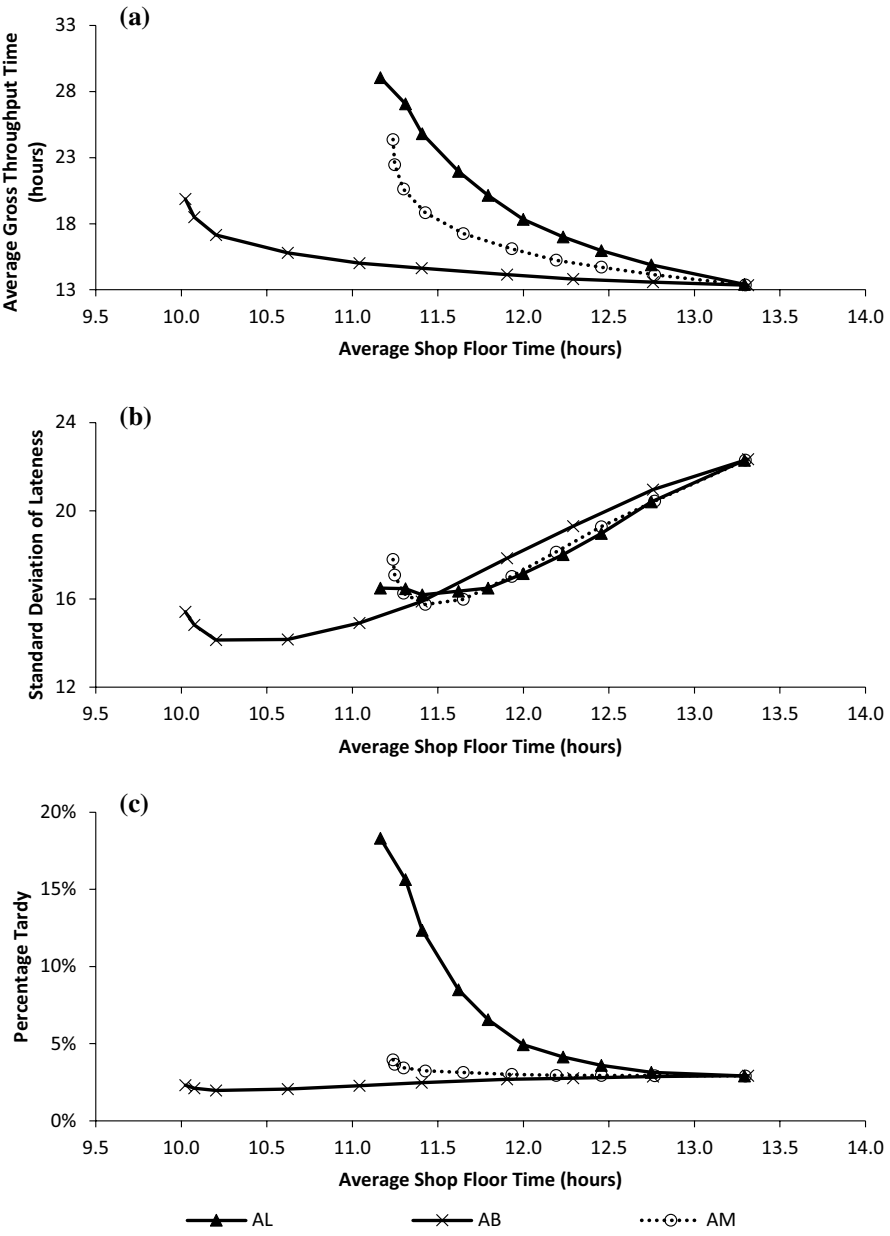

Fig. 4 Performance of mixed method AM in comparison to AB and AL

At time t, the MODD rule for station j will prioritize the order i that has the lowest priority number, given by the maximum of the operation due date δ_{ij} and earliest finish time (Baker 1984), i.e. max (δ_{ij}, $t+p_{ij}$), with p_{ij} being the processing time. This means that orders at a station will be sequenced according to their operation

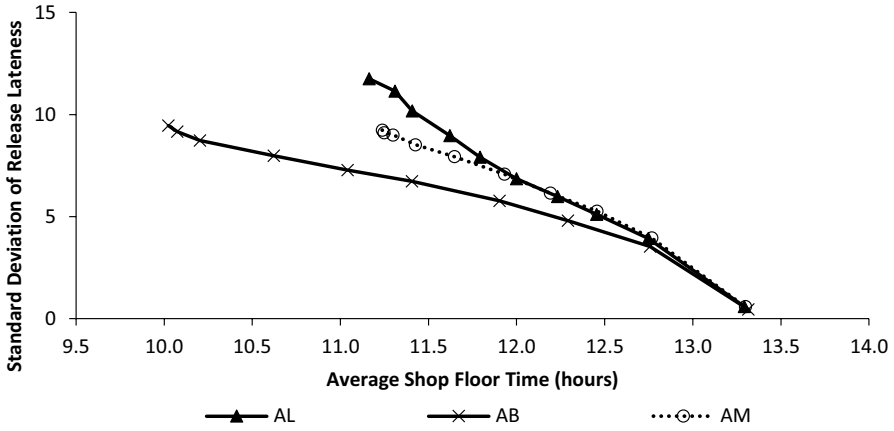

Fig. 5 Standard Deviation of Release Lateness for AL, AB and AM

due date, if the operation due date has not yet passed for any order. Among those orders that have passed their operation due date, it will first select the one with the shortest processing time (SPT). As soon as on average more orders risk to be completed tardy (i.e. busy periods), the main prioritization will be determined by SPT. Contrarily in quiet periods—with at the most one order exceeding its operation due date at the same station—priorities will be determined by the operation due date. The simplest method to determine operation due dates that can be combined with controlled order release is parameter free (Land et al. 2014). It determines the ODDs at the time of release of an order i and distributes the time remaining from the actual release time ρ_i until the due date δ_i* equally among all operations. This means that $\delta_{ij} = \rho_i + i \cdot (\delta_i^* - \rho_i)/5$ in our case with 5 operations for each order. For δ_i* *we* could either use the 'external' due date, or an internal due date, which leaves a small amount of slack until the external due date. Preliminary experiments using the external due date and internal due dates by subtracting 2, 4, 6, 8 and 16 h from the external due date showed that a subtracting 8 h results in the best performing MODD rule for our system. This is applied in Fig. 6. The figure shows that FCFS dispatching results in higher average shop floor times with higher gross throughput times and a higher percentage tardy compared to SPT and MODD. Furthermore, with SPT the average shop floor time can be further reduced than with MODD, resulting in a lower average throughput time and a slightly higher percentage tardy.

The inclusion of a due date orientation in the dispatching rule, as in MODD, could avoid the need to force the release of an order when the latest release date τ_j has been passed at the time of the release decision. These urgency oriented forced releases are part of the balancing methods in our study. Figure 7 shows the results of a further extension that analyzes the exclusion of these forced releases for both SPT and MODD dispatching. Excluding forced releases improves the average gross throughput time, as the ORR method is no longer restricted in its balancing objective. However, both standard deviation of lateness and the resulting percentage tardy deteriorate, which is not attractive. The impact of excluding forced releases for

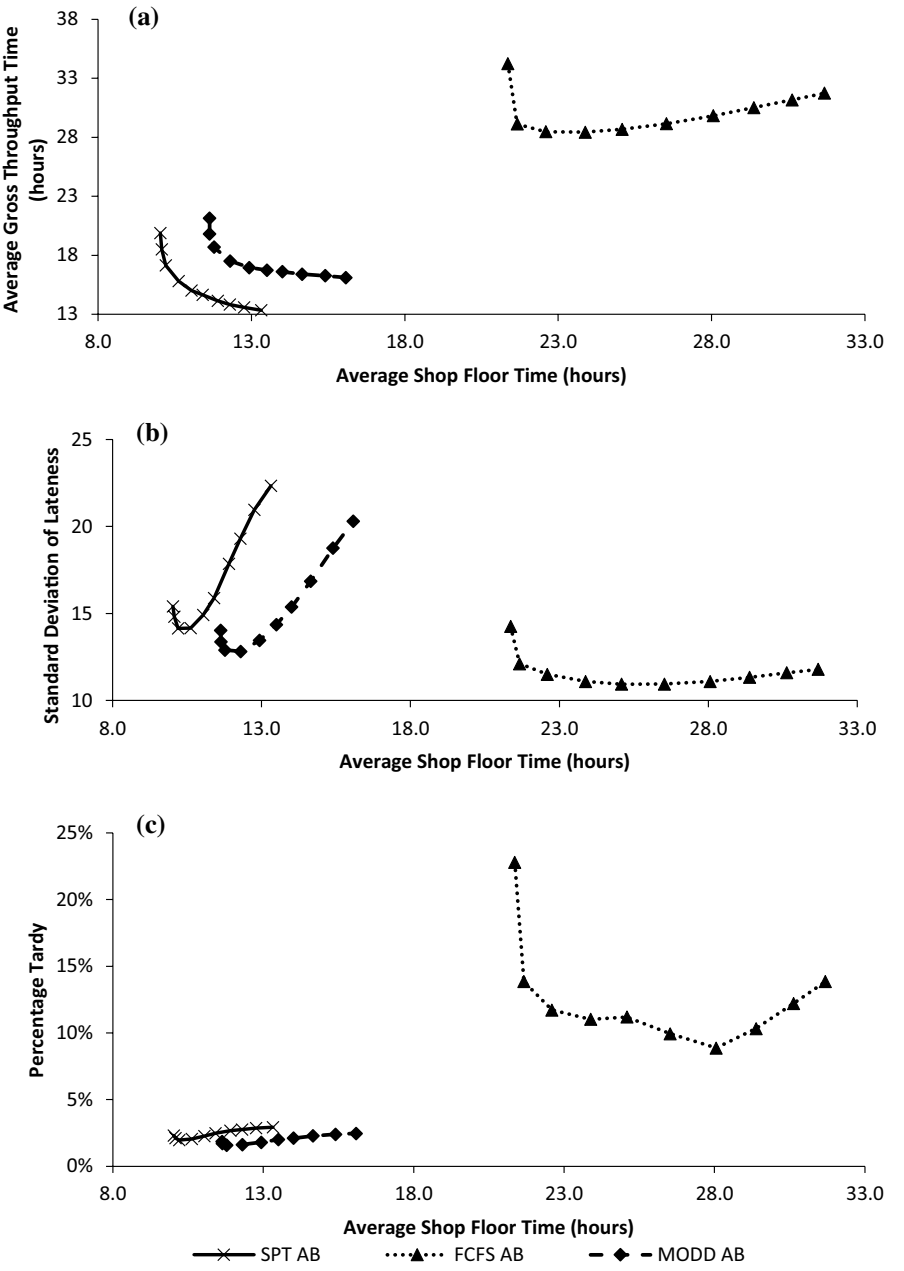

Fig. 6 Performance of AB for SPT, FCFS and MODD dispatching rules

urgent jobs hardly differs between SPT and MODD, even though the due date oriented dispatching rule MODD might partly compensate for not considering urgency at the time of release.

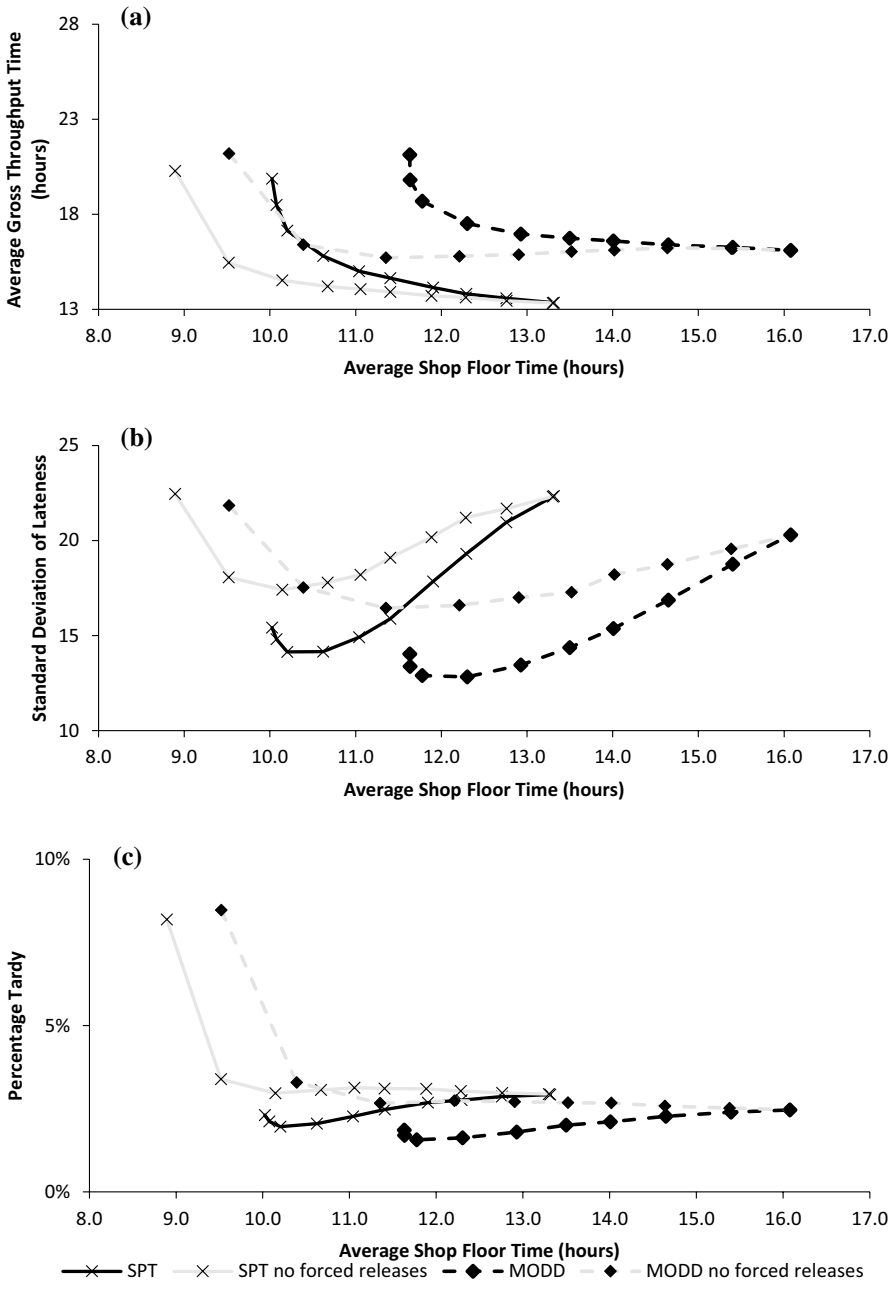

Fig. 7 Performance of AB with SPT and MODD dispatching rules with and without forced releases

4.4 Sensitivity analysis

A sensitivity analysis has been executed for a number of parameters that are fixed in the main experimental design. The influences of due date allowance combined with the planned shop floor time, utilization level and processing time variation are considered here. The results will be confined to the AB method combined with SPT dispatching and forced releases, since this combination was shown to give the best overall performance.

4.4.1 Due date setting and planned shop floor time

In the main experimental design, a due date allowance (DDA) of 6 working days and a planned shop floor time (PST) of 3 working days have been applied in order to create realistic ranges for the percentage of tardy orders for the full set of simulated ORR methods. In the sensitivity analysis we explore due date allowances of 4, 6, and 8 days, combined with planned shop floor times of 2, 3, and 4 days. We explored this full factorial, but exclude the most extreme combinations (DDA; PST) of (4; 4) and (8; 2) which are not logical.

Figure 8 presents the influence for each of the three performance measures. The average gross throughput time (Fig. 8a) increases with tighter due dates. With tighter due dates, more orders exceed their latest release date which forces their release to the shop floor. This reduces the balancing possibilities, and so the gross throughput times increase, especially at tighter workload target levels. Note that the latest release date τ_j is determined by the difference between the due date and the planned shop floor time which is the same for the combinations (8;4) and (6;2). The same applies to the combinations (6;4) and (4;2). Consequently these combinations have exactly the same gross throughput time (Fig. 8a) and standard deviation of lateness (Fig. 8b). Only the percentage tardy (Fig. 8c), which is mainly determined by the due date allowance, differs among all settings.

As could be expected, the percentage tardy (Fig. 8c) increases with tighter due dates and with later releases (shorter planned shop floor times). All figures show that combinations with the largest gap between due date allowance and planned shop floor time enable the largest reduction in realized shop floor times. This also relates to the fact that these combinations give the best opportunities to combine orders for balancing, before release is forced due to exceeding the latest release date. This effect becomes weaker for the highest due date allowances, as loose due dates make the ORR method less dependent on its balancing capabilities. While theoretically it might seem logical to apply tighter planned shop floor times at tighter workload targets, Fig. 8c shows that this relationship is highly complex in terms of due date performance influences. This explains the choice for a constant intermediate planned shop floor time level in the main experimental design.

4.4.2 Utilization level

In a pure flow shop setting relatively high utilization levels might be expected in comparison with job shops. The arrival rate of orders is set to 15 orders per day

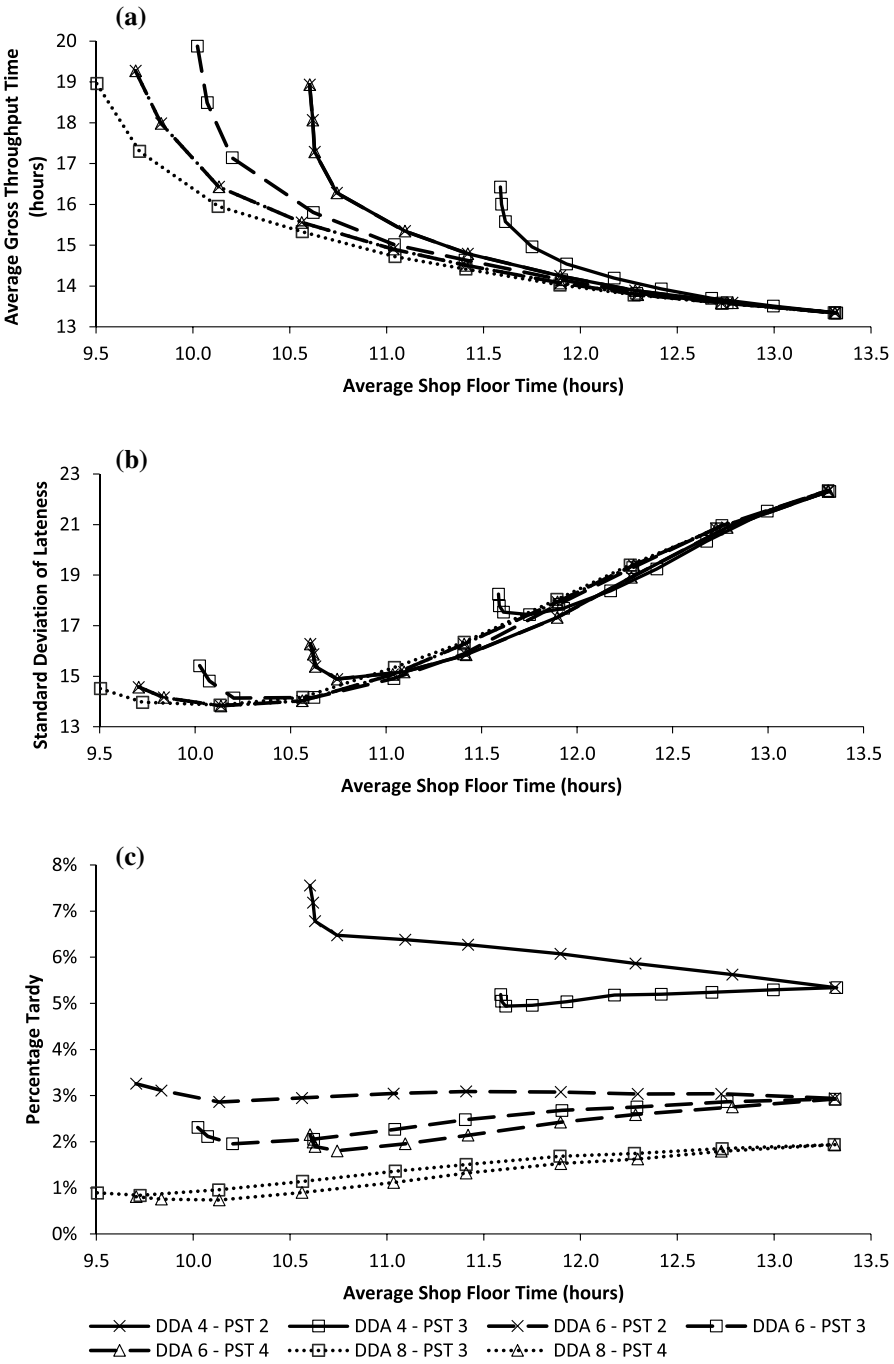

Fig. 8 Influence of the due date allowance and planned shop floor times

and the service rate is 16 orders per day, resulting in a high utilization level of 93.75% in the main experimental design. In the sensitivity analysis the impact of

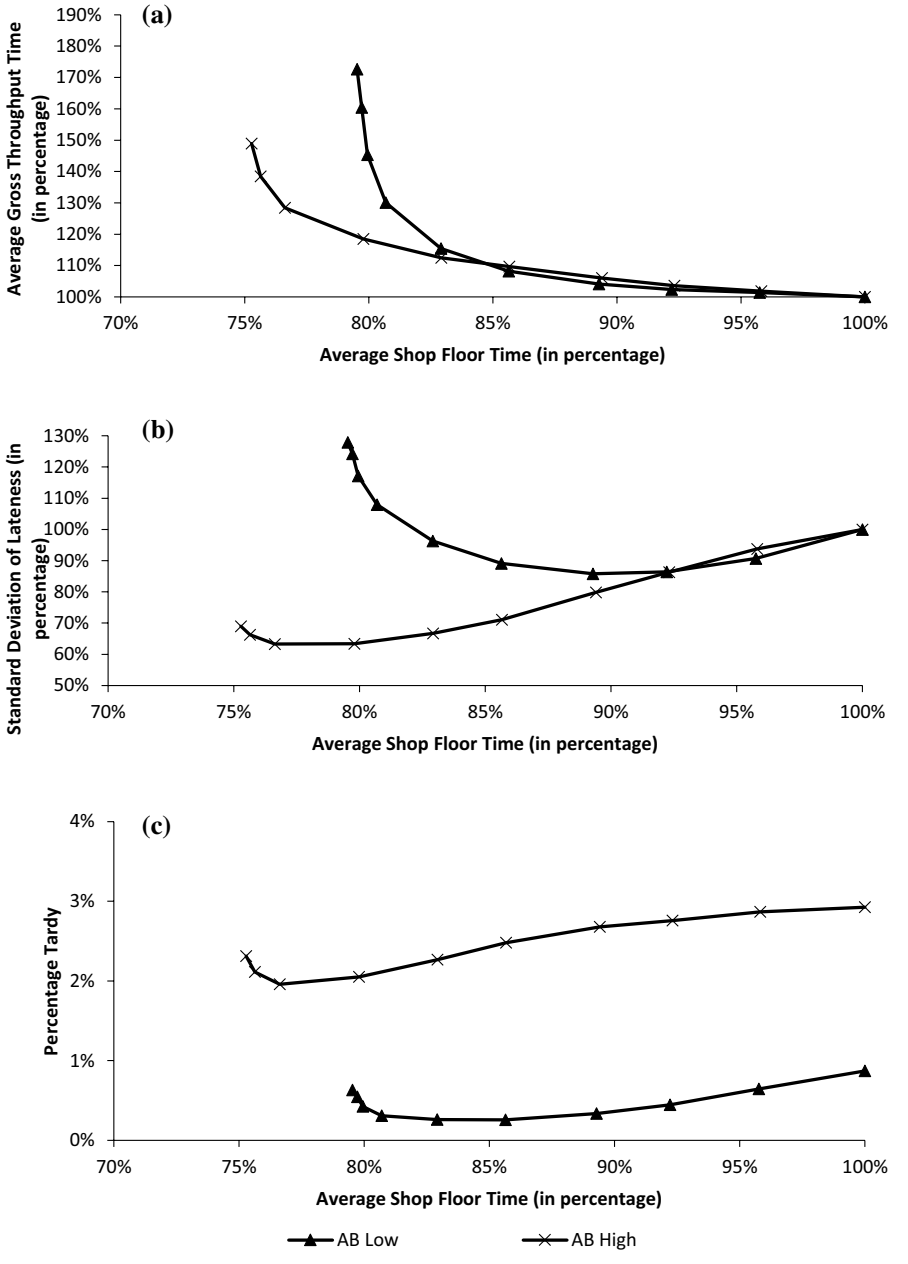

Fig. 9 Influence of the utilization level

decreasing the arrival rate to 14 orders per day is tested, which results in 87.5% utilization. This is indicated as 'low' in Fig. 9 that presents the results.

The absolute impact of utilization on gross throughput times is obviously large and particularly relative performance is relevant. Therefore, the average shop floor time, the average gross throughput time and the standard deviation of lateness have been normalized in the figures. This is done by dividing each value by the value realized at infinite workload target levels. The normalized values are thus expressed as a percentage on both the horizontal and vertical axis. Only the percentage of tardy orders is still expressed as the original non-normalized value. Figure 9a shows that AB is able to realize relatively better gross throughput time performance at high utilization levels. High utilization makes the method more dependent on its strong balancing capabilities. Of course more orders can be completed in time at low utilization levels. The percentage of tardy orders (Fig. 9c), which is always above 2 percent for method AB at 93.75% utilization, gets far below 1 percent for all workload targets when utilization decreases to 87.5%.

4.4.3 Processing time variability

In the main experimental design, the coefficient of variation of the processing times is 0.8. To analyse sensitivity a lower coefficient of variation of 0.6 and a higher coefficient of 1.0 have been tested. The same normalization approach has been applied as for the utilization level.

Figure 10 shows that increased processing time variability deteriorates all performance measures. This confirms the findings of Thürer et al. (2015a). The highest level of variability reduces the amount of possible shop floor time reduction. This is due to the increase of (non-normalized) throughput times at higher variability. With the same due date allowance, this triggers more forced releases of orders. As such, the impact is comparable with that of a lower due date allowance for the same level of variability, as discussed in Sect. 4.4.1.

4.4.4 Discussion of sensitivity analysis

The main experimental design showed the best overall performance for AB (explicit balancing of aggregate loads) combined with SPT dispatching. The sensitivity analysis first showed that this method requires careful setting of due date allowances and planned shop floor times. Higher values of the planned shop floor time lead to slightly lower percentages of tardy orders for the same due date allowance, as long as workload target levels are sufficiently tight. However, tight due dates and long planned shop floor times do not allow for the use of tight workload targets that strongly reduce shop floor times. Tight workload targets are shown to deteriorate performance as it forces the release of many orders that pass their latest release date, thereby reducing the balancing options.

AB performed best for demands that lead to high utilization levels. As practice will favour the heavy use of resources in high variety flow shops, this can be seen as an advantage. High utilization levels take more advantage of the balancing capabilities of the AB method at tight workload target levels. As could be expected, this

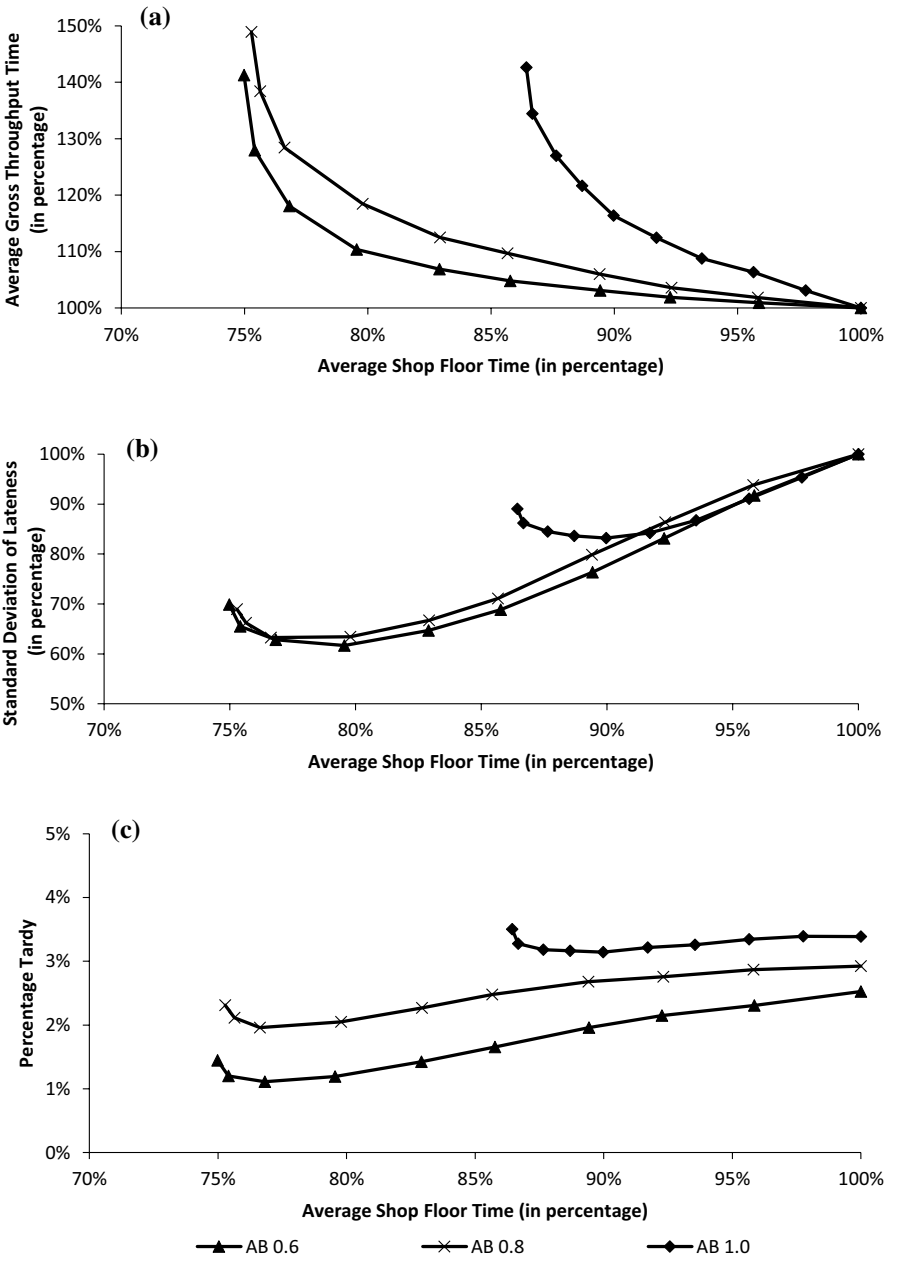

Fig. 10 Influence of processing time variability

will be at the cost of slightly higher percentages tardy, if due date allowances are not adapted.

A similar effect as for tight due dates was found if the variability of processing times exceeded a certain level within the same due date allowance setting. It forces the release of more orders that passed their latest release date, which disturbs the load balancing optimization at tight workload targets and corresponding short shop floor times.

5 Conclusion

Based on the increasing importance in practice of make-to-order flow shops with highly variable processing times, the aim of this study was to develop an improved ORR method for pure flow shops. To realise this aim, the elements of importance for flow shop performance as suggested by earlier studies have been investigated and combined in a modular ORR method design. A simulation study has been used to point out which choices should be made regarding each of the elements. It resolved the confusion resulting from the contradictory results of earlier studies.

In the first place, the findings show that the best performance results from an ORR method that balances workloads explicitly by optimization allowing both under and overloads. Performance deteriorates when only underloads are allowed, that is, using targets as strict limits in an optimization algorithm. In the flow shop context the weakest performance results from applying the traditional workload control heuristic, which balances workloads indirectly by fitting orders successively in workload limits. Secondly, balancing can simply be based on traditional aggregate workload measures, which include the processing times of all orders released and not yet completed at a workstation. Finally, in the specific environment of a flow shop with high processing time variability, dispatching based on SPT is found to be highly effective since it avoids starvation of the workstations. The advantage compared to FCFS dispatching applied in earlier studies is enormous. Use of Modified Operations Due Dates, which combines SPT advantages with a due date orientation, does not deliver clear advantages compared to basic SPT dispatching.

Earlier studies suggested a due date orientation in the ORR method by means of forced releases for urgent orders exceeding their latest release date. This was shown to contribute to all timeliness related performance measures despite an increase of throughput times.

The methods compared in this study had a modular structure. In-depth analyses of the main performance differences within each modular element have been executed to explain the reasons behind the differences observed. Performance differences related to the choice of workload aggregation measures could partly be attributed to the specification of load targets. Optimized balancing was shown to reduce the size of the pre-shop pool of orders that was needed to find a balanced set of orders to combine for release. Consequently, it could keep waiting times before release shorter, while at the same time reducing the times of orders on the shop floor.

A sensitivity analysis finally studied the impact of parameter settings and environmental variables for the best performing method, which optimizes the load balance

for aggregate loads. It showed that the best performing method can successfully deal with high levels of utilization and a broad range of processing time variability levels, while due date parameters should be carefully chosen.

5.1 Theoretical implications

The findings have important implications for existing theory. The use of explicit balancing by optimization, which has limited impact in job shops (Yan et al. 2016), is found to be the most impactful factor for the design of an ORR method in the flow shop environment. Use of optimized balancing was also suggested by Fernandes et al. (2020). However, our findings show that their focus on the minimization of underloads should be replaced by the approach suggested by Portioli-Staudacher and Tantardini (2012) which allows loads exceeding the target. This study shows that workload balancing can best be based on traditional aggregate workload measures, as already suggested by Oosterman et al. (2000) for a workload limiting approach in flow shops. This deviates from the suggestion in (Portioli-Staudacher and Tantardini 2012) that workload balancing in flow shops should only consider the workload that has been released recently (released workload measure) for each workstation. When focusing on a released workload measure, the choice of an appropriate definition for the workload target was shown to be a critical element. Extending schedule visibility, which would consider future release decisions in the optimization approach was incorporated as an element of the ORR method of Portioli-Staudacher and Tantardini (2012). It was not included in the presented full experimental design, as preliminary simulations already indicated that the contribution of this element is minimal. This allows for a significant simplification of the optimization algorithm applied in Portioli-Staudacher and Tantardini (2012). A final contribution to theory follows from the findings on the role of dispatching. The results confirm the finding of Thürer et al. (2015a) that the simple SPT rule is highly effective in flow shops. The advantages compared to FCFS dispatching as done by Portioli-Staudacher and Tantardini (2012) are substantial in all aspects of performance. The more advanced MODD rule applied by Fernandes et al. (2020) considers order urgency as part of the dispatching decision, but this does not lead to significant advantages compared to the simple SPT rule. This study shows that it is more effective to consider order urgency as part of the ORR decision.

5.2 Managerial implications

Besides theoretical consequences, the study has important implications for practice. Supported by smart industry influences, more and more companies are moving towards flow shop configurations, with variable processing times resulting from high levels of customization. Flow shops reveal a performance behaviour that is different from job shops, but clear guidelines for ORR decisions have mainly been developed for job shops. Most of the findings from this study are easy to implement in companies. For example, the information needed for aggregate workload measures can simply be retrieved from information systems, while smart industry

developments further improve the quality of processing time data and feedback data on order progress. The integer programming approach used for explicit workload balancing may add to complexity in comparison with a simpler workload limiting method. However, the models could be highly simplified in this study by avoiding extended schedule visibility and using straightforward workload target levels. The remaining complexity will be less of an issue with the computing facilities available in the emerging smart manufacturing environments. The use in practice of an ORR method that combines the best elements, as pointed out by this study, does not only improve due date performance, but also helps in maintaining lower work-in-process levels and shorter shop floor times. Orders can spend on average around 30% less time on the shop floor with an effective ORR method, while at the same time the percentage tardy is strongly reduced. This allows companies to be more flexible in coping with last-minute changes in order specifications by customers.

5.3 Further research

This study provides an important step by establishing the core elements that build up the best performing ORR method for periodic release of orders in flow shops. However, several interesting questions remain for future research. The number of factors that could be included in this study was limited and had to be limited in order to focus on core elements. For the same reason the ORR method has been studied in isolation, while interactions with other control decisions such as decisions on capacity adjustments certainly deserve attention in future flow shop studies. We should also investigate the impact of having continuous rather than periodic release opportunities. Thürer et al. (2015a) already showed the advantages of continuity for workload limiting methods. In our current study, the optimization algorithms select a combined set of on average 15 orders that provide a balanced package of work to be released at the beginning of each day. As continuous methods would only release one or just a few orders at once, it is a challenge to improve balance effectively by an optimization algorithm. However, according to this study, optimization of balance is the key to performance improvement in flow shops.

References

Baker KR (1984) Sequencing rules and due-date assignments in a job shop. Manag Sci 30(9):1093–1104

Baykasoğlu A, Göçken M (2011) A simulation based approach to analyse the effects of job release on the performance of a multi-stage job-shop with processing flexibility. Int J Prod Res 49(2):585–610

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

Caprihan R, Kumar A, Stecke KE (2006) A fuzzy dispatching strategy for due-date scheduling of FMSs with information delays. Int J Flex Manuf Syst 18:29–53

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

Cigolini R, Perona M, Portioli A (1998) Comparison of order review and release techniques in a dynamic and uncertain job shop environment. Int J Prod Res 36(11):2931–2951

Das SR, Khumawala BM (1991) An efficient heuristic for scheduling batches of parts in a flexible flow system. Int J Flex Manuf Syst 3:121–147

Fernandes NO, Land MJ, Carmo-Silva S (2016) Aligning workload control theory and practice: lot splitting and operation overlapping issues. Int J Prod Res 54(10):2965–2975
Fernandes NO, Thürer M, Pinho TM, Torres P, Carmo-Silva S (2020) Workload control and optimised order release: an assessment by simulation. Int J Prod Res 58(10):3180–3193
Fullerton RR, Kennedy FA, Widener SK (2014) Lean manufacturing and firm performance: the incremental contribution of lean management accounting practices. J Oper Manag 32(7–8):414–428
Germs R, Riezebos J (2010) Workload balancing capability of pull systems in MTO production. Int J Prod Res 48(8):2345–2360
Haeussler S, Netzer P (2019) Comparison between rule- and optimization-based workload control concepts: a simulation optimization approach. Int J Prod Res. https://doi.org/10.1080/00207543.2019.1634297 **(to appear)**
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. https://doi.org/10.1016/j.ijpe.2020.107682 **(to appear)**
Henrich P, Land M, Gaalman G (2004) Exploring applicability of the workload control concept. Int J Prod Econ 90:187–198
Irastorza JC, Deane RH (1974) Loading and balancing methodology for job shop control. AIIE Trans 6(4):302–305
Krishnamurthy A, Suri R (2009) Planning and implementing POLCA: a card-based control system for high variety or custom engineered products. Prod Plan Contr 20(7):596–610
Land MJ (2006) Parameters and sensitivity in workload control. Int J Prod Econ 40:196–209
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) 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 (2014) Integrating load-based order release and priority dispatching. Int J of Prod Res 52(4):1059–1073
Lin JT, Chen CM (2015) Simulation optimization approach for hybrid flow shop scheduling problem in semiconductor back-end manufacturing. Simul Model Pract Theory 51:100–114
Lin JT, Chiu CC, Chang YH (2019) Simulation-based optimization approach for simultaneous scheduling of vehicles and machines with processing time uncertainty in FMS. Flex Serv Manuf J 31(1):104–141
Missbauer H (2020) Order release planning by iterative simulation and linear programming: theoretical foundation and analysis of its shortcomings. Eur J Oper Res 280(2):495–507
Onur L, Fabrycky WJ (1987) Input/output control system for the dynamic job shop. IIE Trans 19(1):88–97
Oosterman B, Land M, Gaalman G (2000) The influence of shop characteristics on workload control. Int J Prod Econ 68:107–119
Perona M, Portioli A (1998) The impact of parameters 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
Portioli-Staudacher A, Costa F, Thürer M (2020) The use of labour flexibility for output control in workload controlled flow shops: a simulation analysis. Int J Ind Eng Comput 11(3):429–442
Sabuncuoglu I, Karapinar HY (1999) Analysis of order review/release problems in production systems. Int J Prod Econ 62:259–279
Shimoyashiro S, Isoda K, Awane H (1984) Input scheduling and load balance control for a job shop. Int J Prod Res 22(4):597–605
Theorin A, Bengtsson K, Provost J, Lieder M, Johnsson C, Lundholm T, Lennartson B (2017) An event-driven manufacturing information system architecture for Industry 4.0. Int J Prod Res 55(5):1297–1311
Thürer M, Stevenson M (2016) Card-based delivery date promising in pure flow shops with order release control. Int J Prod Res 54(22):6798–6811
Thürer M, Silva C, Stevenson M (2011) Optimising workload norms: the influence of shop floor characteristics on setting workload norms for the workload control concept. Int J Prod Res 49(4):1151–1171

Thürer M, Stevenson M, Silva C, Land MJ, Fredendall LD, Melnyk SA (2014) Lean control for make-to-order companies: integrating customer enquiry management and order release. Prod Oper Manag 23(3):463–476

Thürer M, Stevenson M, Protzman CW (2015a) COBACABANA (control of balance by card based navigation): an alternative to kanban in the pure flow shop? Int J Prod Econ 166:143–151

Thürer M, Land MJ, Stevenson M, Fredendall LD, Godinho Filho M (2015b) Concerning workload control and order release: the pre-shop pool sequencing decision. Prod Oper Manag 24(7):1179–1192

Thürer M, Stevenson M, Silva C, Qu T (2017) Drum-buffer-rope and workload control in high-variety flow and job shops with bottlenecks: an assessment by simulation. Int J Prod Econ 188:116–127

Thürer M, Stevenson M, Renna P (2019a) Workload control in dual-resource constrained high-variety shops: an assessment by simulation. Int J Prod Res 57(3):931–947

Thürer M, Zhang H, Stevenson M, Costa F, Ma L (2019b) Worker assignment in dual resource constrained assembly job shops with worker heterogeneity: an assessment by simulation. Int J Prod Res. https://doi.org/10.1080/00207543.2019.1677963 **(to appear)**

Thürer M, Stevenson M, Land MJ, Fredendall LD (2019c) On the combined effect of due date setting, order release, and output control: an assessment by simulation. Int J Prod Res 57(6):1741–1755

Thürer M, Ma L, Stevenson M (2020) Workload control order release in general and pure flow shops with limited buffer size induced blocking: an assessment by simulation. Int J Prod Res. https://doi.org/10.1080/00207543.2020.1735667 **(to appear)**

Van Ooijen HPG (1998) Delivery performance improvement by controlled work-order release and work-center load balancing. Int J Prod Econ 56–57:661–675

Yan H, Stevenson M, Hendry LC, Land MJ (2016) Load-oriented order release (LOOR) revisited: bringing it back to the state of the art. Prod Plan Contr 27(13):1078–1091

Zorzini M, Corti D, Pozzetti A (2008) Due date (DD) quotation and capacity planning in make-to-order companies: results from an empirical analysis. Int J Prod Econ 112:919–933

Kaustav Kundu works as a Research Fellow at the Centre for Next Generation Logistics, National University of Singapore, Singapore. He holds a PhD with a concentration in operations management from Politecnico di Milano, Italy. His main areas of focus are production and logistics in the manufacturing sector using simulation and modelling.

Martin J. Land is Associate Professor in Operations Management at the Department of Operations of the University of Groningen in The Netherlands. His special research interest is in improving flow. This relates to the broad range of patient crowding problems in emergency departments, problems resulting from the intermittency of electricity supply from renewable energy sources and order flows in high variety manufacturing. He published in for example Production and Operations Management, European Journal of Operational Research, Omega, International Journal of Production Economics and the International Journal of Production Research.

Alberto Portioli-Staudacher is full professor of Operations Management at Politecnico di Milano. His research interest is mainly in Operational Excellence and Improvement, with a particular focus on performances, on the impact on employees' satisfaction and on the opportunities enabled by digitalization. He addressed many different industries, from healthcare, to banking, to high volume manufacturing, to customized manufacturing. He published 4 books and over 100 articles in main journals and conferences proceedings.

Jos A. C. Bokhorst is an assistant professor at the Faculty of Economics and Business of the University of Groningen. He obtained his Ph.D. in Operations Management at the University of Groningen in 2005. His research interests include performance analysis of planning concepts, dual resource constrained systems, lean and smart manufacturing, and work design.

Affiliations

Kaustav Kundu[1] · Martin J. Land[2] · Alberto Portioli-Staudacher[1] · Jos A. C. Bokhorst[2]

Kaustav Kundu
kaustav.kundu@polimi.it

Martin J. Land
m.j.land@rug.nl

Jos A. C. Bokhorst
j.a.c.bokhorst@rug.nl

[1] Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy

[2] Department of Operations, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands