Operations Management Research (2021) 14:420–433
https://doi.org/10.1007/s12063-021-00210-2

Labor flexibility integration in workload control in Industry 4.0 era

Federica Costa[1] · **Alberto Portioli-Staudacher**[1]

Received: 26 August 2020 / Revised: 23 July 2021 / Accepted: 15 August 2021 / Published online: 17 September 2021

Abstract

The paradigm shift toward Industry 4.0 is facilitating human capability, and at the center of the research are the workers—Operator 4.0—and their knowledge. For example, new advances in augmented reality and human–machine interfaces have facilitated the transfer of knowledge, creating an increasing need for labor flexibility. Such flexibility represents a managerial tool for achieving volume and mix flexibility and a strategic means of facing the uncertainty of markets and growing global competition. To cope with these phenomena, which are even more challenging in high-variety, low-volume contexts, production planning and control help companies set reliable due dates and shorten lead times. However, integrating labor flexibility into the most consolidated production planning and control mechanism for a high-variety, low-volume context—workload control—has been quite overlooked, even though the benefits have been largely demonstrated. This paper presents a mathematical model of workload control that integrates labor flexibility into the order review and release phase and simulates the impact on performance. The main results show that worker transfers occur when they are most needed and are minimized compared to when labor flexibility is at a lower level of control—shop-floor level—thus reducing lead time.

Keywords Workload control · Output control · Labor flexibility · Human factors · Production planning · Industry 4.0

1 Introduction

Growing global market competition and the increased diversity of customer demands have led to the rapid development of manufacturing (Tao et al. 2017). These phenomena make the demand more difficult to predict along, thus affecting the setting of reliable due dates. Companies that fail to establish long-lasting relationships with customers because of unreliable due dates are destined to disappear from the market (Kingsman and Hendry 2002). This is the reason why companies must shorten lead times and improve due-date estimation, which is especially challenging for companies that are producing on a to-order basis. To cope with this challenge, production planning and control (PPC) mechanisms (Małachowski and Korytkowski 2016; Reuter and Brambring 2016) play a fundamental role since PPC is intended to ensure a company's profitability by delivering goods to customers on time and in the right quantities. Workload control (WLC) is a consolidated PPC mechanism that was developed for high-variety, low-volume contexts, such as small- and medium-sized make-to-order (MTO) companies (Kundu et al. 2020; Stevenson et al. 2005), and it has been demonstrated to provide good margins of improvements (Kundu et al. 2020, 2018; Portioli-Staudacher et al. 2020). In addition to PPC, to cope with variability and uncertainties—for example, of demand—companies adopt different buffering strategies, such as inventory or capacity buffers (i.e., excess capacity). However, flexibility plays a fundamental role in mitigating buffering needs and costs (Hopp and Spearman 2011). Capacity buffers can be realized through outsourcing production, hiring new workers or cross-training them in a wider range of tasks to achieve higher labor flexibility. In particular, labor flexibility is a managerial tool that can be used to achieve volume and mix flexibility in production, which is a strategic means of facing uncertainty in markets and growing global competition (Goyal and Netessine 2011). This approach has recently become even more crucial with the advent of Industry 4.0 technologies since humans are still by far the most flexible production factor. In fact, one of the key issues in Industry

---

✉ Federica Costa
federica.costa@polimi.it

Alberto Portioli-Staudacher
alberto.portioli@polimi.it

[1] Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Via Lambruschini 4/B, 20156 Milan, Italy

4.0 is enabling human-centric capability, which leads to Operator 4.0 (Romero et al. 2016) and Operator–Workstation Interaction 4.0 (Golan et al. 2020).

The increasing need for labor flexibility leads workers to perform a wider range of tasks and to share more responsibilities, creating a need for more overall on-the-job knowledge that is available at the right time and in the right place (Hannola et al. 2018). Such knowledge is currently enabled by Industry 4.0 technologies, such as human–machine interfaces and augmented reality (AR).

The social environment for manufacturing has changed considerably in recent years (Jardim-Goncalves et al. 2016), leading to the rapid development of manufacturing (Tao et al. 2017). In line with this development, the skills, flexibility and efficiency of shop-floor workers are decisive factors in ensuring accurate product specifications, meeting deadlines and keeping machines running (Yew et al. 2016). Although often neglected, human factors, especially flexibility, are important elements in real production settings (Gong et al. 2017). Workers who are flexible and can perform a variety of tasks are likely to solve problems more efficiently and to generate new product ideas (Oke 2013). The wider a worker's variety of skills, the more flexible the worker is in terms of the variety of goods/services produced or the range of job assignments (Sawhney 2013).

In the last decade, an increasing number of novel digital technologies, such as AR and the Internet of Things (IoT), have shown their potential to empower human workers (Köffer 2015). Here, the term "empowerment" refers to empowering the users of digital technology, for example, by increasing their strengths, competences, performance and satisfaction. This can be done by providing them with action-relevant knowledge for their tasks. Hence, digital technologies are expected to vastly improve the knowledge management processes of workers and generate immediate benefits for their work (Hannola et al. 2018). In this sense, Industry 4.0 technologies currently represent strong enablers for labor flexibility, which is crucial in present-day markets.

Having introduced the synergism between PPC and labor flexibility enabled by Industry 4.0 technologies, as both target the delivery of goods on time, making companies more profitable in the market, this study aims to present and test through computer-based simulation a new PPC model for high-variety, low-volume companies that integrates labor flexibility.

The remainder of this paper is structured as follows: In Sect. 2, the authors review the relevant literature on WLC, the consolidated PPC mechanism for high-variety, low-volume companies, and labor flexibility and the associated Industry 4.0 enabling technologies. Additionally, the research questions that motivated the study are presented. Section 3 presents the mathematical WLC model that integrates labor flexibility along with the simulation model. Then, the results are presented and discussed in Sect. 4 before the conclusions are presented in Sect. 5.

2 Literature review

2.1 Workload control

WLC is a PPC mechanism specifically developed for a high-variety, low-volume context. The concept has been shown to significantly improve production system performance through simulation (Kundu et al. 2020, 2018; Portioli-Staudacher et al. 2020) and, on occasion, in practice.

While there are several different approaches to WLC (Bergamaschi et al. 1997), a major unifying principle is input/output control (I/OC); i.e., a shop's input rate should be equal to its output rate (e.g., Wight 1970; Plossl and Wight 1971). Consequently, there are two control mechanisms within the WLC concept (Land and Gaalman 1996; Kingsman 2000): (i) input control (I/C), which regulates the work that enters the shop and/or shop floor, and (ii) output control (O/C), which uses capacity adjustments to regulate the outflow of work.

As in previous WLC studies, it is assumed that all jobs are accepted, materials are available, and all necessary information, e.g., regarding shop-floor routing and processing times, is known. If order release is applied, jobs are not immediately released to the shop floor but are retained in a so-called preshop pool (PSP) from which they are released to meet certain performance targets. There are many order release methods in the WLC literature; for example, see the reviews by Wisner (1995), Land and Gaalman (1996) and Bergamaschi et al. (1997). Workload limiting (WL) is recognized by practitioners and academics as the seminal order review and release (ORR) method for workload implementation (see Bechte 1988), and it has been proven to lead to significant performance improvement (Oosterman et al. 2000; Yan et al. 2016). It keeps the workload that has been released from the PSP but not yet completed at each station within certain limits or norms.

While I/C has received much attention in the WLC literature (Fredendall et al. 2010; Land et al. 2015; Melnyk and Ragatz 1989; Philipoom et al. 1993; Sabuncuoglu and Karapinar 1999; Kundu et al. 2020; Thürer et al. 2015), the effective realization of O/C has been largely neglected (Thürer et al. 2016a, b). Only recently has more research emerged that uses WLC theory to guide O/C decisions—in particular, regarding when to *increase* capacity (Land et al. 2015; Thürer et al. 2014, 2015). Few studies have investigated the effect of O/C within WLC theory (most of them have used a job-shop configuration), and most have not focused on a specific O/C strategy except Thürer et al. (2014, 2015), who investigated how the subcontracting decision improves performance. Land et al. (2015) used simulation to investigate a short-term increase in capacity in response to the violation of a given level of the workload and showed that small but timely capacity adjustments targeted at high-load periods significantly improve delivery performance. More recently, Thürer et al. (2016a, b, 2018) combined this O/C mechanism with WLC order release and due-date setting, and Shoaib-ul-Hasana et al. (2018) in two case studies proposed and tested a routine-based framework implementing WLC in which, at the execution stage, capacity is adjusted by means of dynamic allocation of workers among different assembly cells when the capacity requirement exceeds the limit.

However, major assumptions of all existing studies on O/C in the WLC literature are that capacity is increased (e.g., Land et al. 2015; Shoaib-ul-Hasana et al. 2018), work subcontracted (e.g., Thürer et al. 2014), or work rejected (e.g., Kingsman and Hendry 2002). This perspective neglects an important factor often used in practice to deal with temporary overloads: worker allocation from underloaded stations to overloaded stations. This approach was considered by Portioli-Staudacher et al. (2020), who showed the impact on performance of allocating dynamically idle workers from underloaded stations to overloaded stations with a workload-controlled release of orders, with the labor flexibility decision made at the execution stage on the shop floor. Additionally, Costa et al. (2019) presented a preliminary study on how to consider labor flexibility decisions at a higher level of control, that is, the order release phase of WLC.

2.2 Labor flexibility

The term labor flexibility is used to indicate the relative ease with which workers can be shifted among organizational units (Frye 1974). An important characteristic of the workforce is the development of multiple skills (Hopp et al. 2004); the larger a worker's range of skills, the more flexible the worker is, either in terms of the variety of goods and/or services he/she can produce or in terms of the range of job assignments he/she can undertake (Sawhney 2013). This flexibility in turn can be used as a buffer to protect throughput from variability, specifically in production contexts with high variability (Treleven 1989; Kher and Fredendall 2004). Labor flexibility is typically modeled through cross-training or flexibility matrices (Park and Bobrowski 1989; Park 1991; Brusco and Johns 1998) and through worker assignment rules since workers are relocated among stations.

Worker assignment is typically driven by three questions: When should a worker be transferred? Where should the worker be transferred? Who (i.e., which worker) should be transferred? Two main types of *When* rules are used in the literature: a *centralized* rule under which a worker is transferred each time an order is completed and a *decentralized* rule under which a worker is transferred after completing all orders (in process and queueing) at his/her current station. In contrast, a broad set of different *Where* rules exists, but the literature typically argues that the *Where* rule has less impact than the *When* rule (Xu et al. 2011). Meanwhile, *Who* rules are dominated by workers' efficiency considerations (Bobrowski and Park 1993); they are relevant only if workers are not perfectly interchangeable. In fact, labor efficiency is service rate at which a worker can work at a station (Bobrowski and Park, 1993). If a worker is assigned to a new station, then a productivity loss is likely to occur. The concept of flexibility does not consider this effect; rather, workers are assumed to either have the maximum level of efficiency or to be unable to work at a station.

2.3 Industry 4.0 enabling technologies

With the paradigm shift toward Industry 4.0, resources should become "smart" by providing real-time awareness and interaction capacity for the manufacturing environment and personnel. This shift is completed not only by installing smart machines in a factory but also by facilitating human capability (Wang et al. 2020), with new types of interactions between humans and machines that improve the nature of work and increase the flexibility of production (Oztemel and Gursev 2020). According to Orio (2015), integrating context awareness and data mining techniques with traditional and control solutions will reduce production line downtimes, maintenance problems, and operational costs of manufacturing and at the same time guarantee more efficient management of human resources in manufacturing environments.

AR is an innovative human–machine interaction (HMI) that overlays virtual components on a real-world environment (Dini and Della Mura 2015). It has many potential manufacturing applications, including assembly, maintenance, product design, layout planning, robotics (Backhaus and Reinhart 2017) and machining (Yew et al. 2016), with several correlated benefits.

For example, in the assembly domain, Tang et al. (2002) evaluated the effectiveness of spatially overlaid instructions using AR in an assembly task compared with other traditional media and found that the error rate was reduced by 82%. In the field of automobile production, Reiners et al. (1999) introduced AR for assembling car doors by creating a real-time fully three-dimensional HMD-based training application showing how to assemble the door lock in the door. In maintenance engineering systems, Lipson et al. (1998) presented a new online product maintenance approach based on AR. Takata et al. (2001) proposed an operation support system that provided information on disassembly sequences of copying machines.

To accelerate operators' cognition process, Seki (2003) invented a production cell called "Digital Yatai" that monitors assembly progress and presents information about the next step in the process. Using a semitransparent head-mounted display, Reinhart and Patron (2003) developed an AR system to supply information to an operator about the task he/she is performing. Sugi et al. (2005) used a projector to provide assembly information to an operator, while Longo et al. (2017) introduced a Sophos-MS that could integrate AR content and intelligent tutoring systems to support operators in complex HMIs.

According to Reinhart and Patron (2003), using AR in production offers advantages wherever there is a large proportion of search time, workers have to frequently change work content, or the assembly task is very complex and requires a large amount of information. Currently, the increasing need for the flexibility of production workers has resulted in their performing a wider range of tasks and sharing more responsibilities. This change has created a need for more overall on-the-job knowledge that is available at the right time and in the right place (Hannola et al. 2018).

2.4 Discussion of the literature

WLC studies on O/C are limited. Moreover, the existing literature is not concerned with the type of capacity adjustment but simply assumes an increase in capacity (Land et al. 2015; Thürer et al. 2016a, b), with the exception of Portioli-Staudacher et al. (2020), who considered labor flexibility at the shop-floor level, and Costa et al. (2019), who examined the order release stage. Thus, the possible impact of labor flexibility that is currently enabled by on-the-job knowledge technologies, such as AR, on achieving volume flexibility, which is crucial in facing growing global competition, has been neglected. Moreover, small capacity adjustments targeted at handling high-load periods improve performance (Land et al. 2015), suggesting that high- and low-load periods must be taken into account in releasing customers' order release. Additionally, Fredendall et al. (1996) incorporated labor availability information into order release, disregarding, however, the main advances in order release in the WLC literature.

Based on this premise, this study aims to (i) present a WLC model that integrates labor flexibility at the order release phase of WLC and (ii) assess, through simulation, the model's impact on performance.

2.5 Research methodology

This section is divided into two subsections. Section 3.1 presents the WLC model that integrates labor flexibility—with workers dynamically allocated from underloaded stations to overloaded stations—at the order release phase of WLC (WLWorker +). Section 3.2 presents the simulation model that aims to assess the impact on performance of WLWorker+.

For this study, we consider companies operating in a high-variety, low-volume context with production that follows a dominant flow sequence, as in previous studies in the WLC domain (Kundu et al. 2018, 2019; Portioli-Staudacher et al. 2020; Costa et al. 2019). Such companies can be found, for example, in the ceramics industry and in furniture manufacturing (Portioli-Staudacher and Tantardini 2012). We consider pure flow lines with multiple manned workstations that produce large products, as in the automotive or CNC machining industry (Dimitriadis 2006). The product size is sufficient to allow two workers to perform together on the same order, avoiding any blocking situation. The pure flow shop in this study has a number of workers equal to the number of stations—five.

2.6 WLWorker+

One of the early implementations of WL was Bertrand's workload concept presented in Land and Gaalman (1996). It has received much attention in the literature, with researchers testing different methodologies to account for job workload (Oosterman et al. 2000), investigating the integration of due-date information (Thürer et al. 2017), and trying to integrate O/C (Thürer et al. 2016a, b). In all cases, the most commonly implemented version is as follows:

1. All jobs in the PSP are prioritized according to the dispatching rule (e.g., first come first served).
2. Job i with the highest priority is considered first for release.
3. Job i is released if its contribution to station load together with the current station load does not exceed the workload norm for all stations at the same time. However, if the job's contribution load exceeds the workload norm, it is retained in the PSP.

We extend the WL model, proposing WLWorker+, which integrates labor flexibility decisions into ORR. It is built on the consideration that after the release of orders from the PSP, the workload at each station is known. Indeed, it happens that some stations (there is at least one that violates the workload norm and stops the release of orders) are fully loaded, meaning their workload is equal to the workload norm, and some others are not fully loaded, meaning their workload is lower than the workload norm.

Based on this premise, a second review of orders remaining in the PSP—if there are any left—may occur. During this second order review, orders could eventually be released if their workload contribution to stations that are fully loaded is absorbed by stations whose workload is lower than the workload norm. Figure 1 presents this situation: station 1 and station 4 reached the workload norm and limited the release of orders from the PSP, while stations 2, 3 and 5 did not reach the workload norm.

A new order in the PSP is evaluated, and its contribution to the workload at each station is considered (area with wide downward diagonal lines). The evaluation of the new order causes some stations to require extra capacity (station 1 and station 4) that can be absorbed by stations 2 and 3.

The extra load that stations 2 and 3 could undertake and the extra capacity that stations 1 and 4 could receive is stored in the matrix of capacity adjustments (CAdj) (Table 1).

Table 1 Matrix of CAdj(i,j), capacity adjustments, of worker *i* to station *j*

CAdj(i, j)		TO (j)				
		1	2	3	4	5
FROM (i)	1	/	0	0	0	0
	2	0	/	0	20	0
	3	35	0	/	0	0
	4	0	0	0	/	0
	5	0	0	0	0	/

All the values of CAdj(i, j) in the matrix are greater than or equal to 0, and they indicate the free capacity in time units that could be provided by worker *i*—because the workload of station *i* is lower than the norm—to station *j*—which has reached the workload norm. In Table 1, the value of CAdj(3, 1) is equal to 35 time units, which means that worker 3 could provide an extra load of 35 time units to station 1. This results in an extra load for worker 3 and a temporary free capacity for station 1.

After the evaluation of the matrix of capacity adjustments, the following steps are performed (Costa et al. 2019):

1. The expected workload for worker i is computed as follows:

$$ExpectedWorkload(i) = StationWorkload(i) + ExtraLoad(i) - FreeCapacity(j) \quad (1)$$

where Station Workload(i) is computed as for station workload in WL (Oosterman et al. 2000). ExtraLoad(i) and Free Capacity(j) are computed from the matrix of capacity adjustments as follows:

$$ExtraLoad(i) = \sum_{J=0}^{N} CAdj(i, j) \quad (2)$$

ExtraLoad(i) is defined as the workload that worker *i* performs at station *j*. It is obtained through a horizontal sum of the matrix of capacity adjustments.

$$FreeCapacity(j) = \sum_{J=0}^{N} CAdj(j, i) \quad (3)$$

FreeCapacity(j) is the workload of station *i*, which needs free capacity, performed by worker *j*. It is obtained through a vertical sum of the matrix of capacity adjustments.

2. Two sets of stations are defined: a set φ of stations with a workload lower than the norm and a set Ω of stations with a workload that is above the norm. Set φ is com-

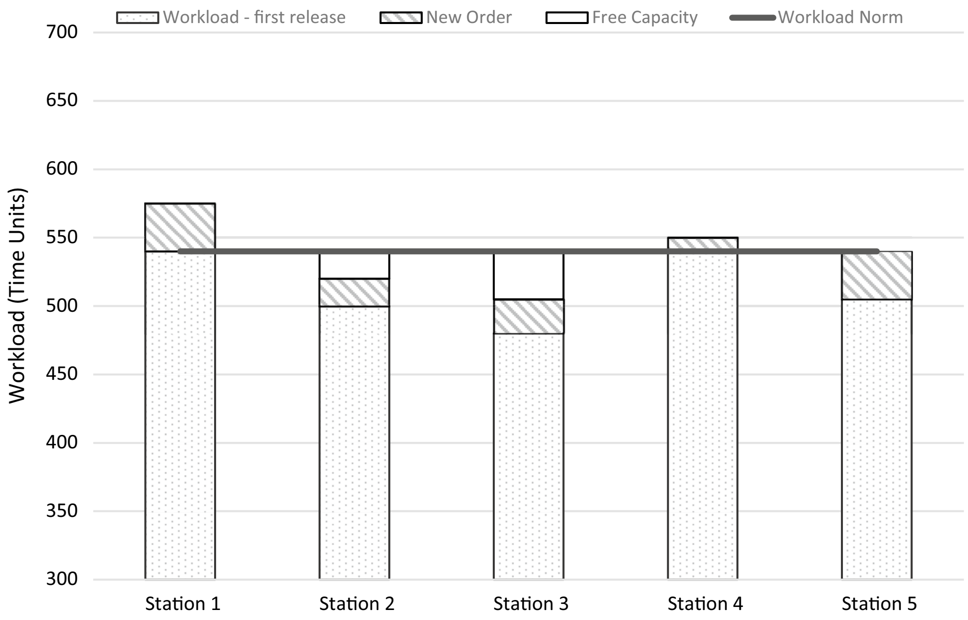


Fig. 1 Workload computation for WLWorker+

posed of stations with Expected Workload(i) lower than the average expected workload (average computed across all stations, considering $ExpectedWorkload(i)$). They can provide their free capacity to stations in Ω, which is composed of stations with Expected Workload(i) higher than the average expected workload. The extent to which stations in Ω need temporarily extra capacity is equal to the workload contribution of the order in the PSP under evaluation. The extent to which stations in φ can provide free capacity is defined according to the difference between their expected workload and the workload norm.

3. All possible permutations (Pe) of the stations in φ to stations in Ω are calculated. Every permutation of the resulting set represents a possible allocation from a station in φ to a station in Ω.
4. All possible worker allocations that combine stations in φ with stations in Ω that are not feasible (i.e., the free capacity of stations in φ is not enough to fulfill the need for extra load of stations in Ω) are eliminated.
5. A score is assigned to each permutation as follows:

$$Score(Pe) = \sum_j ExtraTime(j) \tag{4}$$

where ExtraTime(j) is the amount of extra load in time units.

The permutation that minimizes the score is selected, and in this way, the allocation of worker i to station j that minimizes the time units that worker i spends at station j is determined.

6. Release of the job from the PSP.

At the end of this second phase of WLWorker+, all the jobs that have not been released are retained in the PSP, and they are reevaluated for release during the next release period.

The decisions of *Who* is selected for allocation and *Where* they will be allocated are intrinsically defined by WLWorker+, as shown above.

2.7 Simulation assessment

Discrete event simulation is implemented as one of the most commonly used techniques to analyze and test manufacturing systems (Negahban and Smith 2014; Thomas et al. 2018). It is used to assess WLWorker+ (presented in Sect. 3.1) against WL with labor flexibility at the execution stage (WLWorker), as in Portioli-Staudacher et al. 2020. The nonflexible scenario (WL static)—workers who do not move among stations—is also presented as the baseline. A stylized simulation model was implemented in the Python programming language using the SimPy simulation module to prevent interactions that could interfere with the understanding of the main experimental factors. While every shop in practice could differ from the stylized model, it captures the job and shop characteristics of MTO companies, i.e., high processing time variability and high arrival rate variability (Rossini et al. 2019). We have kept our flow shop relatively small since this allows causal factors to be identified more easily. The shop is a line composed of 5 stations, as in previous studies (Kundu et al. 2018, 2020; Portioli-Staudacher et al. 2020; Costa et al. 2019). There is one worker per station, so the shop is fully staffed, and no dual resource constraint exists. Each station can allow the simultaneous operation of more than one worker at the same station, and a maximum of 2 workers can work simultaneously at the same station. Operation processing time follows a log-normal distribution truncated at 360 time units with a mean of 30 time units and variance equal to 900 time units. The interarrival time of jobs follows an exponential distribution with a λ that ensures that workers are, on average, occupied for 93.75% of their time. Due dates are set exogenously by adding a random allowance factor, uniformly distributed, to the job entry time. The minimum is set considering the maximum processing time and the routing length of jobs. Finally, the first come first served rule is used for priority dispatching.

The labor flexibility parameters are summarized in Table 2.

For WLWorker+, the *Where* and *Who* rules are determined by the matrix of capacity adjustments, as presented in Sect. 3.1.

The input parameters to the simulation model (operation processing times, interarrival times and due dates) are generated from the probability distribution of random variables. The probability distribution functions are those most commonly used in the operations literature (Thürer et al. 2016a, b; Kundu et al. 2018, 2020; Portioli-Staudacher and Tantardini 2012) to recreate the behavior of a theoretical flow shop in the high-variety, low-volume context. To increase the accuracy of the model, common random numbers and antithetic variates techniques have been used as variance reduction techniques (Kohlas 1982). The simulation model was verified and validated by means of face and statistical validation (Chung 2003).

Table 2 Labor flexibility parameters

Factors	Level
When Rule	Decentralized
Where rule	MaxEff rule for WLWorker
Who rule	Random rule
Labor flexibility	5—all workers are able to work at all five stations
Labor efficiency	100% at their main station and α decrement efficiency at the remaining four stations

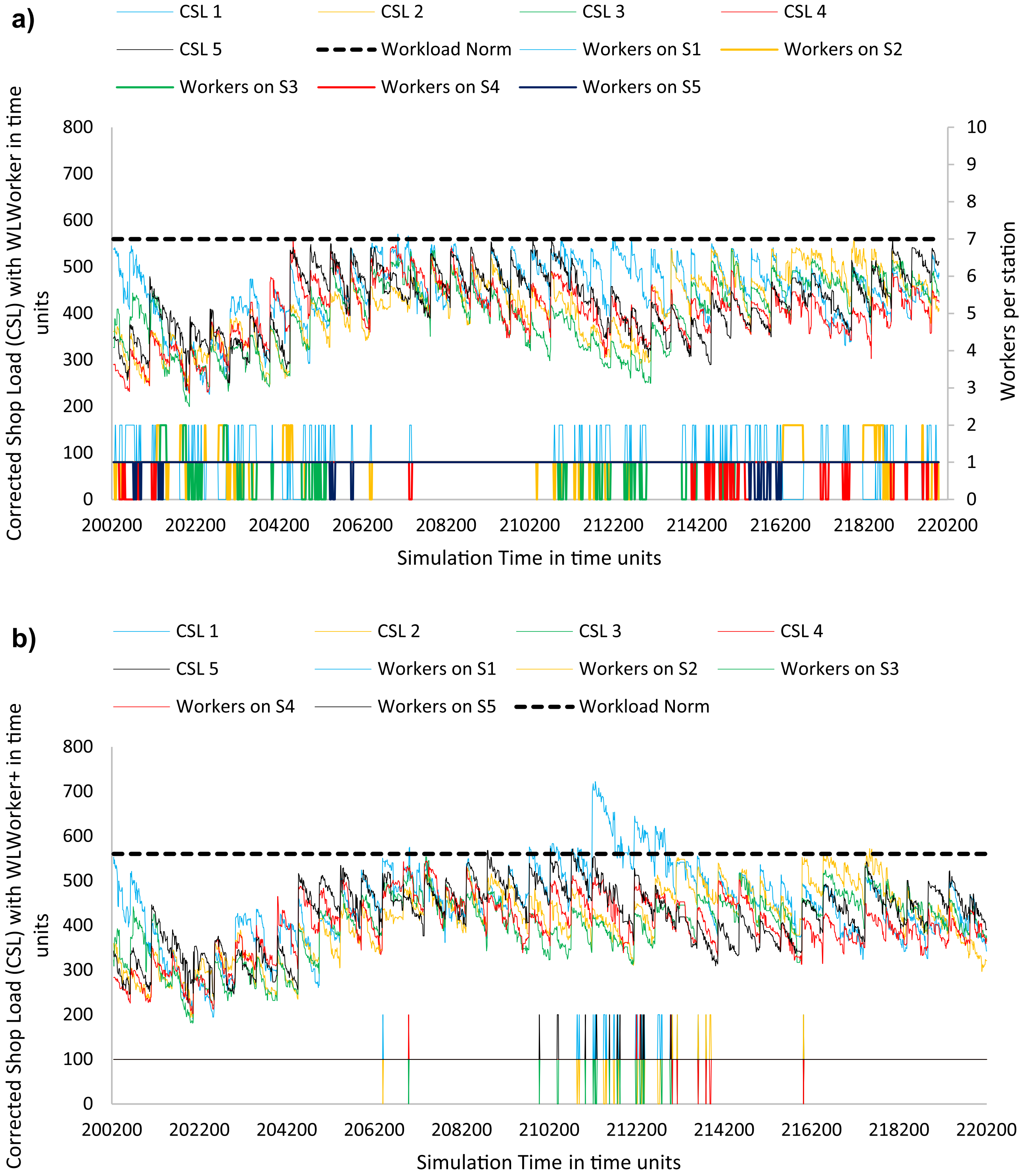


Fig. 2 Corrected Shop Load (CSL) in time units for WLWorker (**a**) and WLWorker+ (**b**) and over the simulation time for the five Station (S) along with number of worker on each Station (S)

3 Results

To assess WLWorker+ against WLWorker, we show for both models the workload at each station—measured as corrected shop load (CSL), as in Oosterman et al. (2000)—and the number of workers at each station. CSL measures in time units the number of orders queueing in front of the station plus all the upstream orders that are arriving in that queue. On the Y-axis, CSL in time units is presented together with the number of workers at each station (secondary Y-axis), while the X-axis shows a simulation period of 20,000 time units, starting after the warm-up period. Both graphs of Fig. 2 show 5 series of data, one for each station. Figure 2a presents CSL for WLWorker, while Fig. 2b presents CSL for WLWorker+. In both graphs, the number of workers per station is shown by the bottom line, which can assume the following values: 0, zero workers at the station; 1, one worker at the station; and 2, two workers at the station.

Graphs 2a and 2b show that CSL released is higher than the workload norm in 2b, while it is always lower than or equal to the workload norm in 2a. Figure 2b also shows that when CSL is higher than the workload norm, workers are transferred from stations where CSL is lower than the norm to stations where CSL is higher than the norm. For example, this situation is visible for station 1 (blue line) at approximately 200,000 time units, station 5 at approximately 206,000, and station 3 (yellow line) at approximately 212,000. This is confirmed by the line showing the number of workers in Fig. 2b. Every time CSL is higher than the norm at a station because additional jobs are released from the PSP, 2 workers are working at that station. This extra workload is released, surpassing the norm, because WLWorker+ ensures that it is absorbed by workers whose workload is lower than the norm.

From the line showing number of workers on the secondary vertical axis of graphs 2a and 2b, we observe the number of workers at a station throughout the simulation period. In 2a, the number of workers at stations changes more frequently, ranging from 0 to 1 and 2, than in 2b. This results in a lower number of workers transfers with WLWorker+ than with WLWorker. Moreover, the number of workers at stations with WLWorker changes without regard for CSL at the stations, in contrast to 2b, in which workers are transferred from stations where CSL is lower than the norm to stations where CSL is higher than the norm.

To better explain how worker transfers are decided at the order release with WLWorker+, Fig. 3 is presented. ExtraLoad (2) and FreeCapacity (3) are shown on the Y-axis, while the same simulation period as in Fig. 2 is shown on the X-axis.

FreeCapacity is presented in the upper part of the graph with positive Y-values, while ExtraLoad is presented in the lower part of the graph with negative Y-values. As discussed in Sect. 3.1, ExtraLoad and FreeCapacity are computed from the matrix of capacity adjustments. The former computes the workload of the order under review at PSP that causes an extra load at stations that have reached the norm, while the latter represents the available capacity at stations that have not reached the norm. Figure 3 shows that the X-axis mirrors the areas below the FreeCapacity and ExtraLoad curves or that the areas below FreeCapacity are larger than those

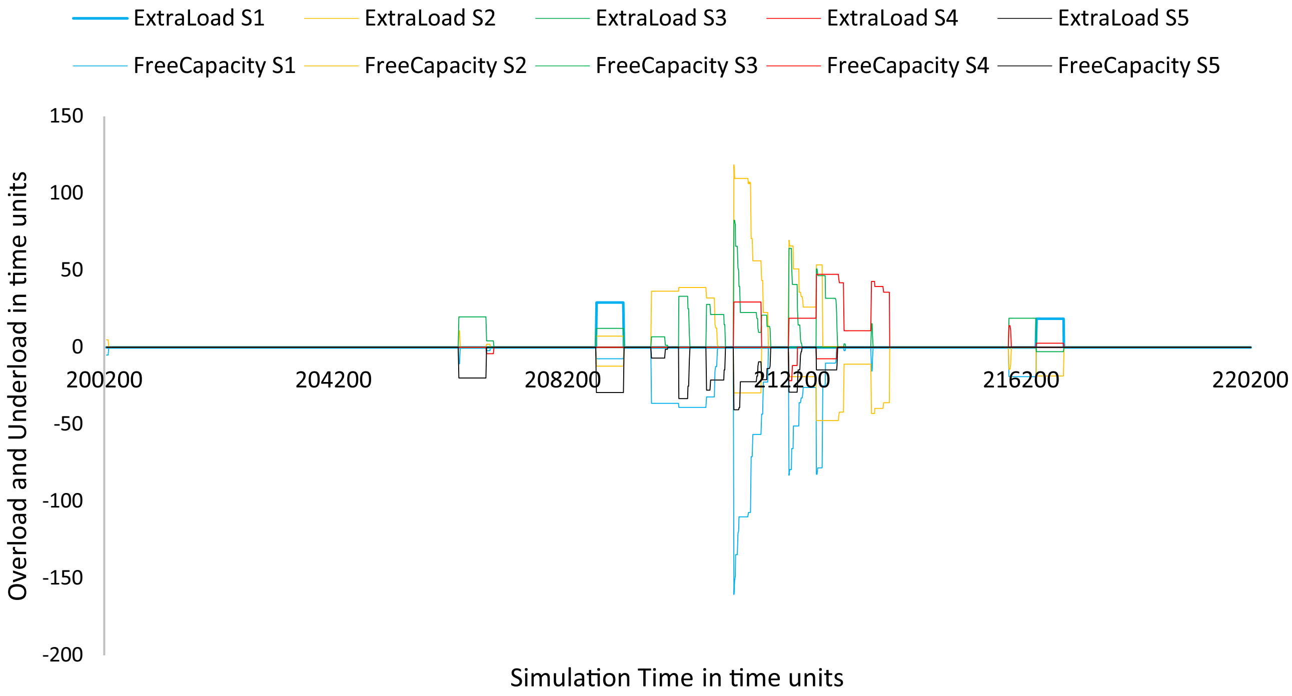


Fig. 3 ExtraLoad and FreeCapacity in time units for each Station over the same simulation time used in Fig. 2. ExtraLoad is represented by values whose y-values are lower than zero, while FreeCapacity is represented by values whose y-values are greater than zero

Table 3 Experimental setting

Factors	Levels
Release model	WLWorker–WLWorker+
Transfer time	10–20
Decrement efficiency α	0–50%

below ExtraLoad. This is evidence of how the matrix of capacity adjustments works and how WLWorker+ assesses which stations could devote extra capacity and which stations could receive it. When the ExtraLoad of an additional order can be absorbed by FreeCapacity, the additional order under evaluation is released, with a consequent increase in CSL (as shown in Fig. 2b), and the worker at the station that has FreeCapacity is transferred to the station that needs ExtraLoad—as shown by the bottom line of Fig. 2b. Stations that have ExtraLoads receive an extra worker—2 workers per station. Figure 2 shows how worker allocation is decided considering worker availability—ExtraLoad and FreeCapacity—and that allocations are realized when needed, in contrast to WLWorker. In fact, allocations occur so that workers whose workload is lower than the norm can perform the ExtraLoad released from the PSP.

It is worth noting from Fig. 3 that even if the ExtraLoad and FreeCapacity areas match, worker transfers do not occur and CSL does not increase—no additional order is released with WLWorker+. The reason is that worker transfer conditions—idleness of workers, for example—must exist before worker transfers can take places.

3.1 Performance analysis

In the previous section, we graphically showed how WLWorker+ limits the number of worker transfers compared to WLWorker, an aspect that is highly relevant in practice since transfers negatively affect performance.

In this section, we show the impact of the two different release models on four performance measures, considering the experimental setting presented in Table 3. The experimental setting includes (i) two different release methods, WLWorker and WLWorker+; (ii) worker transfer time among stations; and (iii) a decrement of efficiency that affects workers when they are transferred.

The following four main performance measures are used:

- *Gross throughput time* (GTT), the time between order entry and completion*;*
- *Shop-floor throughput time* (SFTT), the time between order release from the pool and completion;
- *Percentage of tardy orders* (%Tardy), the percentage of orders with positive lateness (given by the completion date minus the due date); and
- *Mean tardiness* (Tardiness), given by max(0;lateness).

Statistical analysis of our results was first conducted using an analysis of variance (ANOVA) that shows that the factors in the experimental setting (Table 4) are statistically relevant.

The results were collected over 500,000 time units following a warm-up period of 200,000 determined through the Welch method, as presented in Mahajan and Ingalls (2004), for all performance measures. To determine the number of runs, the mean square pure error (MSPE) was implemented for the performance measures used. The number of runs chosen—100—is equal to the maximum that allows the convergence of the MSPE for all performance measures.

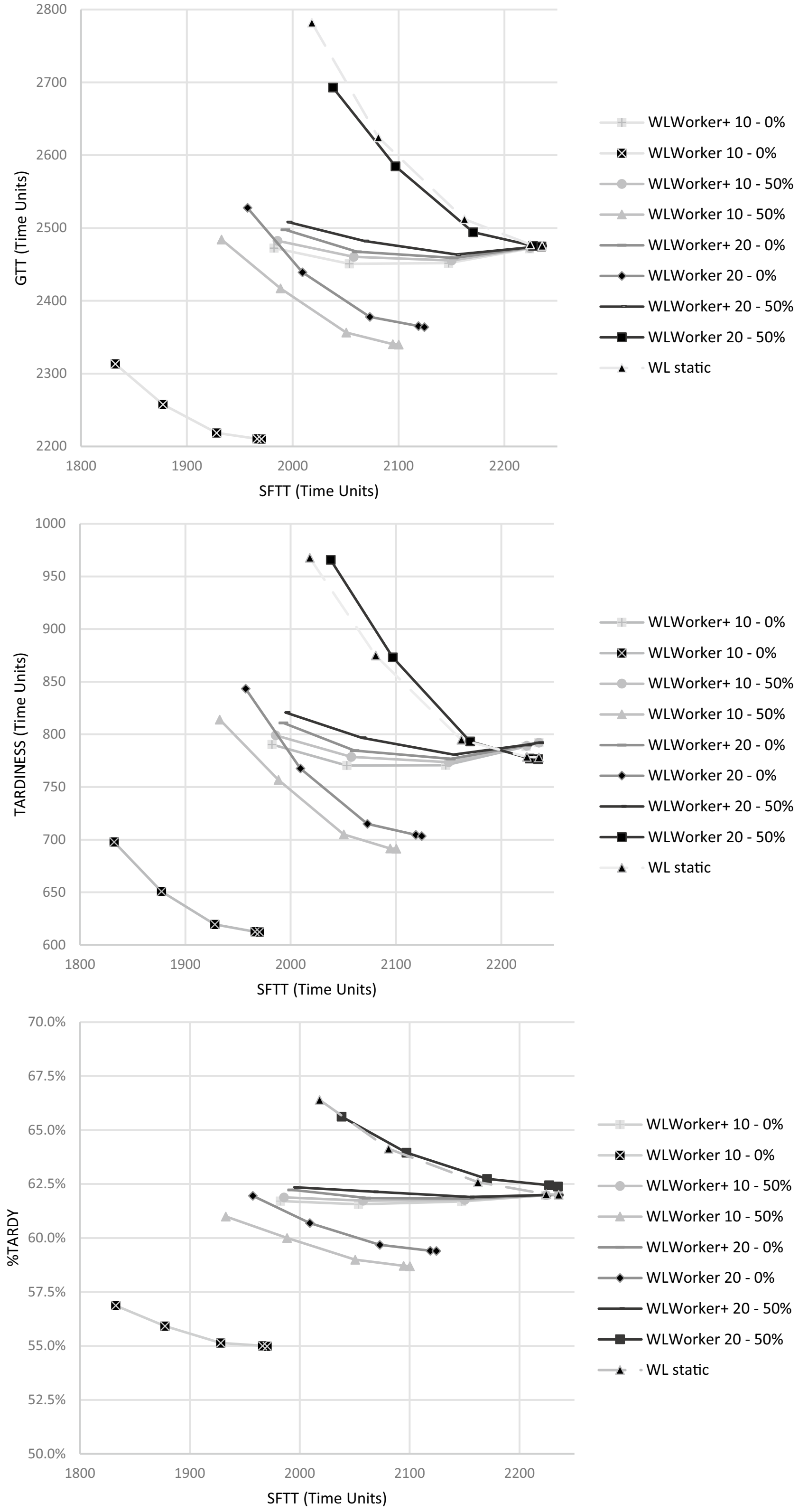
To aid the interpretation of the results, they are presented in the form of performance curves. The left-hand starting point of the curves represents the lowest workload norm. The workload norm increases from left to right in each graph, with each data point representing one norm level. Increasing the norm increases the level of work in process and, as a result, lengthens the SFTT. In Fig. 4, GTT, %Tardy, and Tardiness are shown against SFTT for WLWorker+ and WLWorker and for different levels of transfer time and efficiency decrement.

Figure 4 shows that with WLWorker delivery, related performances and lead-time performances are strongly impacted by transfer time and efficiency loss. In fact, WLWorker outperforms WLWorker+ when the efficiency loss is 0% and the

Table 4 Anova results for SFTT

Factors	DoF	Adj SS	Adj MS	F-value	P-value
Release model	1	76,487	76,487	31.48	0.000
Transfer time	2	77,084	38,542	15.86	0.000
Decrement efficiency	1	80,476	80,476	33.13	0.000
Release model × transfer time	2	59,779	29,890	12.30	0.000
Release model × decrement efficiency	1	61,783	61,783	25.43	0.000
Transfer time × decrement efficiency	2	222	111	0.05	0.955
Error	50	121,472	2429	0.06	
Total	59	477,303			

Fig. 4 Gross Throughput time (GTT), Tardiness, %Tardy Jobs are presented against Shop Floor Throughput time (SFTT) for WLWorker+ and WLWorker with different transfer time 10–20 and efficiency decrement 0%-50%. The WL static is inserted as baseline scenario



transfer time is lowest. When the efficiency loss is highest, WLWorker is outperformed by WLWorker+ except at transfer time 10. This result is confirmed for all the performance measures considered.

With WLWorker, the results are strongly dependent on transfer time and efficiency loss values, but this does not hold true for WLWorker+. When efficiency loss and transfer time increase, the performance of WLWorker worsens dramatically, and it is even outperformed in the nonflexible scenario that is represented by the dashed line in each graph. This means that with WLWorker, it is preferable not to rely on labor flexibility under certain conditions and to keep workers static at their main task. This is not the case for WLWorker+, which always improves performance with respect to the nonflexible scenario, disregarding the level of efficiency loss and transfer time.

For WLWorker+, we observe that all performance measures are quite stable when transfer time and efficiency loss change, since all four curves of WLWorker+ remain stable around values that are close.

This finding is explained by the highest number of transfers that WLWorker determines, as shown in Fig. 2. Transfer time and efficiency loss, when higher than zero, represent at each transfer a temporary loss of capacity—for transfer time, workers are not using their capacity to process an order, and for efficiency loss, they are slower to process an order—that negatively affects performances, increasing delivery-related and lead-time performances. This results in a worsening of performance that is much higher for WLWorker than for WLWorker+.

4 Conclusions

To cope with uncertainty and variability in current markets, PPC plays a fundamental role in ensuring the profitability of companies. In addition to PPC, flexibility in terms of capacity and volume is recognized as a fundamental strategy for facing the aforementioned challenges. Many studies have recently considered labor flexibility in terms of capacity, which is currently possible since with the spread of Industry 4.0 technologies, AR and HMI enable the transfer of knowledge among workers and the allocation of shop-floor workers to different tasks/stations.

Among PPC mechanisms, WLC is the most consolidated in high-variety, low-volume contexts; however, labor flexibility is generally overlooked. Few studies have considered labor flexibility in WLC, showing that performance is improved when it is decided at the execution stage—shop-floor level (Portioli-Staudacher et al. 2020). To the best of our knowledge, research is scarce that considers and tests labor flexibility at the higher level of control of the order release phase of WLC through simulation. Based on this premise, this paper aimed to present a WLC that integrates labor flexibility at the ORR phase of WLC (i) and to assess, through simulation, its impact on performance (ii).

Extending the most investigated ORR limiting approach, the authors presented a mathematical model called WLWorker+ that integrates labor flexibility decisions into the ORR phase. It considers, in a second ORR phase, the workload at each station, deciding to release additional orders if their workload contribution to stations that have reached the workload norm can be absorbed by stations where the workload is lower than the norm. From the evaluation of the workload contribution of the additional order, together with the workload at each station, WLWorker+ decides whether to release the additional order and at the same time determines the allocation of workers—which workers are transferred to perform an extra load at which stations (*Who* and *Where* rules). We showed how to consider worker availability information during the ORR phase—through the matrix of capacity adjustments—and how to exploit this information to release additional orders.

Then, through simulation, we assessed the performance of WLWorker+, showing that it optimizes worker transfers since it transfers workers when required most: extra capacity is provided by a worker who has a workload below the workload norm to stations where the workload surpasses the workload norm. In this way, WLWorker+ minimizes the number of transfers compared to WLWorker, which does not consider labor flexibility at the ORR phase and disregards the overall workload at stations in deciding worker transfers. Moreover, we showed that WLWorker does not consider the overall workload at stations when deciding worker transfers but greedily moves workers if they are idle at that moment. To show how the lower number of transfers with WLWorker+ represents an advantage for performance, we showed how delivery-related performance and lead time are impacted by the introduction of transfer time between stations and efficiency loss of workers when they are transferred. The results revealed that performance is worsened by efficiency losses and transfer times when adopting WLWorker, which was outperformed even in the nonflexible scenario. The results also show that this does not happen with WLWorker+, where performances remain quite stable independent of transfer time and efficiency loss and are always improved with respect to the nonflexible scenario. This is explained by the low number of transfers, which is extremely relevant for practitioners that, with the advent of AR and HMI technologies, are increasingly benefiting from labor flexibility adoption. Considering labor flexibility at the order release rather than at the shop-floor level also has the advantage of reducing the complexity of managing worker transfers since who is going to move and where is known in advance before the release of orders. This is relevant for practice.

5 Limitations and future research

A first important limitation of our study is our focus on a pure flow shop. While this is justified by the practical relevance of this shop type in the context of multiple manned work stations, future research is needed to extend our findings to more complex shops. Moreover, we limited the experimental setting to transfer time and efficiency loss to keep the focus of the study on the presentation of the new WLC model that integrates labor flexibility into the order release phase and on a simple comparison between labor flexibility decisions at order release and at the shop-floor level. Future studies will focus on expanding the experimental setting, for example, with learning and forgetting factors, heterogeneous efficiency and flexibility among workers, and different *When/Where/Who* rules. Then, we decided to first present the integration of labor flexibility in the most consolidated limiting ORR; however, future research could consider other ORRs in the literature, for example, the balancing approach (Portioli-Staudacher and Tantardini 2012) or the latest advancements in continuous order releases.

Acknowledgements This work was supported by The Lean Excellence Centre, observatory of Politecnico di Milano, www.lean.polimi.it.

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

References

Backhaus J, Reinhart G (2017) Digital description of products, processes and resources for task-oriented programming of assembly systems. J Intell Manuf 28(8):1787–1800

Bechte W (1988) Theory and practice of load-oriented manufacturing control. Int J Prod Res 26:375–395. https://doi.org/10.1080/00207548808947871

Bergamaschi D, Cigolini R, Perona M, Portioli-Staudacher A (1997) Order Review and release strategies ina job shop environment: a review and a classification. Int J Prod Res 35(2):399–420. https://doi.org/10.1080/002075497195821

Bobrowski PM, Park PS (1993) An evaluation of labor assignment rules when workers are not perfectly interchangeable. J Oper Manage 11(3):230–249. https://doi.org/10.1016/0272-6963(93)90003-8

Brusco MJ, Johns TJ (1998) Staffing a multiskilled workforce with varying levels of productivity: an analysis of cross-training policies. Decis Sci 29(2):499–515. https://doi.org/10.1111/j.1540-5915.1998.tb01586.x

Chung CA (ed) (2003) Simulation modeling handbook: a practical approach, 1st edn. CRC Press, Boca Raton. https://doi.org/10.1201/9780203496466

Costa, F., Portioli-Staudacher, A., Nisi, D. & Rossini, M. (2019). Integration of order release and output control with worker's allocation in a pure flow shop. *9th IFAC Conference MIM 2019, Berlin.*

Dimitriadis SG (2006) Assembly Line balancing and group working: a heuristic procedure for workers' groups operating on the same product and workstation. Comput Oper Res 33:2757–2774. https://doi.org/10.1016/j.cor.2005.02.027

Dini G, Della Mura M (2015) Application of augmented reality techniques in through-life engineering services. Procedia CIRP. 38:14–23. https://doi.org/10.1016/j.procir.2015.07.044

Fredendall LD, Melnyk SA, Ragatz G (1996) Information and scheduling in a dual resource constrained job shop. Int J Prod Res 34(10):2783–2802

Fredendall LD, Ojha D, Patterson JW (2010) Concerning the theory of workload control. Eur J Oper Res 201(1):99–111. https://doi.org/10.1016/j.ejor.2009.02.003

Frye JS (1974) Labor flexibility in multiechelon dual-constraint job shops. Manage Sci 20(7):1073–1081

Golan M, Cohen Y, Singer G (2020) A framework for operator-workstation interaction in Industry 4.0. Int J Prod Res 58(8):2421–2432

Gong X, Deng O, Gong G, Liu W, Ren Q (2017) A memetic algorithm for multi-objective flexible job-shop problem with worker flexibility. Int J Prod Res. https://doi.org/10.1080/00207543.2017.1388933

Goyal M, Netessine S (2011) Volume flexibility, product flexibility, or both: the role of demand correlation and product substitution. Manuf Serv Oper Manag 13(2):145–280. https://doi.org/10.1287/msom.1100.0311

Hannola L, Richter A, Richter S, Stocker A (2018) Empowering production workers with digitally facilitated knowledge processes—a conceptual framework. Int J Prod Res 56:4729–4743. https://doi.org/10.1080/00207543.2018.1445877

Hopp WJ, Spearman ML (2011) Factory physics. Waveland Press Inc, New York

Hopp WJ, Tekin E, Van Oyen MP (2004) Benefits of skill chaining in serial production lines with cross-trained workers. Manage Sci 50(1):83–98. https://doi.org/10.1287/mnsc.1030.0166

Jardim-Goncalves R, Grilo A, Popplewell K (2016) Novel strategies for global manufacturing systems interoperability. J Intell Manuf 27(1):1–9

Kher HV, Fredendall LD (2004) Comparing variance reduction to managing system variance in a job shop. Comput Ind Eng 46(1):101–120. https://doi.org/10.1016/j.cie.2003.11.002

Kingsman BG (2000) Modelling input-output workload control for dynamic capacity planning in production planning systems. Int J Prod Econ 68(1):73–93. https://doi.org/10.1016/S0925-5273(00)00037-2

Kingsman B, Hendry L (2002) The relative contributions of input and output controls on the performance of a workload control system in make-to-order companies. Prod Planning Control 13(7):579–590. https://doi.org/10.1080/0953728021000026285

Köffer S (2015) Designing the digital workplace of the future—what scholars recommend to practitioners. In: Proceedings of the International Conference of Information Systems (ICIS). Fort Worth, TX, USA

Kohlas J (1982) Stochastic methods of operations research. Cambridge University Press, Cambridge

Kundu K, Rossini M, Portioli-Staudacher A (2018) Analysing the impact of uncertainty reduction on WLC methods in MTO flow shops. Prod Manuf Res 6(1):328–344. https://doi.org/10.1080/21693277.2018.1509745

Kundu K, Rossini M, Portioli-Staudacher A (2019) A study of a kanban based assembly line feeding system through integration of simulation and particle swarm optimization. Int J Ind Eng Comput 10(3):421–442

Kundu K, Land MJ, Portioli-Staudacher A, Bokhorst JAC (2020) Order review and release in make-to-order flow shops: analysis and design of new methods. Flex Serv Manuf J. https://doi.org/10.1007/s10696-020-09392-6

Land MJ, Gaalman GJC (1996) Workload control concepts in job shops a critical assessment. Int J Prod Econ 46–47:535–548. https://doi.org/10.1016/S0925-5273(96)00088-6

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. https://doi.org/10.1016/j.ijpe.2015.07.007

Lipson H, Shpitalni M, Kimura F, Goncharenko I (1998) Online product maintenance by web-based augmented reality. In: Proceedings of CIRP Design Seminar on New Tools and Workflows for product Development, pp 131–143.

Longo F, Nicoletti L, Padovano A (2017) Smart operators in industry 4.0: a human-centered approach to enhance operators' capabilities and competencies within the new smart factory context. Comput Ind Eng 113:144–159. https://doi.org/10.1016/j.cie.2017.09.016

Mahajan, S., & Ingalls, R.G. (2004). Evaluation of methods used to detect warm-up period in steady state simulation. In *Proceedings of the 2004 Winter Simulation Conference*.

Małachowski B, Korytkowski P (2016) Competence-based performance model of multi-skilled workers. Comput Ind Eng 91:165–177. https://doi.org/10.1016/j.cie.2015.11.018

Melnyk SA, Ragatz GL (1989) Order Review_release, Research Issues and Perspectives.Pdf. Int J Prod Res 27(7):1081–1096. https://doi.org/10.1080/00207548908942609

Negahban A, Smith JS (2014) Simulation for manufacturing system design and operation: literature review and analysis. J Manuf Syst 33(2):241–261. https://doi.org/10.1016/j.jmsy.2013.12.007

Oke A (2013) Linking manufacturing flexibility to innovation performance in manufacturing plants. Int J Prod Econ 143(2):242–247. https://doi.org/10.1016/j.ijpe.2011.09.014

Oosterman B, Land M, Gaalman G (2000) Influence of shop characteristics on workload control. Int J Prod Econ 68(1):107–119. https://doi.org/10.1016/S0925-5273(99)00141-3

Orio GD (2015) The adapter module: a building block for self-learning production systems. Robot Comput Integr Manuf 36:25–35

Oztemel, E., & Gursev, S. (2020). Literature review of Industry 4.0 and related technologies. *Journal of Intelligent Manufacturing*, 31(1), 127–182.

Park PS (1991) The examination of worker cross-training in a dual resource constrained job shop. Eur J Oper Res 52(3):291–299. https://doi.org/10.1016/0377-2217(91)90164-Q

Park PS, Bobrowski PM (1989) Job release and labor flexibility in a dual resource constrained job shop. J Oper Manage 8(3):230–249

Philipoom PR, Malhotra MK, Jensen JB (1993) An evaluation of capacity sensitive order review and release procedures in job shops. Decis Sci 24(6):1109–1134. https://doi.org/10.1111/j.1540-5915.1993.tb00506.x

Portioli-Staudacher A, Tantardini M (2012) A lean-based ORR system for non-repetitive manufacturing. Int J Prod Res 50(12):3257–3273. https://doi.org/10.1080/00207543.2011.564664

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(429–442):4

Reiners D, Stricker D, Klinker G, Müller S (1999) Augmented reality for construction tasks: doorlock assembly. In: Proceeding of the international workshop on Augmented Reallty: placing artificial objects in real scenes, pp 31–46.

Reinhart G, Patron C (2003) Integrating augmented reality in the assembly domain -fundamentals, benefits and applications. CIRP Ann 52(1):5–8. https://doi.org/10.1016/S0007-8506(07)60517-4

Reuter C, Brambring F (2016) Improving data consistency in production control. Procedia CIRP 41:51–56. https://doi.org/10.1016/j.procir.2015.12.116

Romero D, Bernus P, Noran O, Stahre J, Fast-Berglund Å (2016) The operator 4.0: human cyber-physical systems & adaptive automation towards human-automation symbiosis work systems. In Proceedings of the IFIP international conference on advances in production management systems (APMS.), Springer, Cham, pp 677–686

Rossini M, Audino F, Costa F, Cifone F, Kundu K, Portioli-Staudacher A (2019) Extending lean frontiers: a kaizen case study in an Italian MTO manufacturing company. The International Journal of Advanced Manufacturing Technology. https://doi.org/10.1007/s00170-019-03990-x

Sabuncuoglu I, Karapinar HY (1999) Analysis of order review / release problems in production systems. Int J Prod Econom 62:259–279

Sawhney R (2013) Implementing Labor flexibility: a missing link between acquired labor flexibility and plant performance. J Oper Manage 31(1–2):98–108. https://doi.org/10.1016/j.jom.2012.11.003

Seki S (2003) One by one production in the "Digital Yatai": practical use of 3D-CAD data in the fabrication (digital engineering). J Jpn Soc Mech Eng 106(1013):32–36

Shoaib-ul-Hasana S, Macchi M, Pozzetti A, Carrasco Gallego R (2018) A routine-based framework implementing workload control to address recurring disturbances. Prod Planning Control 29(11):943–957

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. https://doi.org/10.1080/0020754042000298520

Sugi M, Nikaido M, Tamura Y, Ota J, Arai T, Kotani K, Takamasu K, Shin S, Suzuki H, Sato Y (2005) Motion control of self-moving trays for human supporting production cell "attentive workbench".&nbsp;In: ICRA. Proc. IEEE Int. Conf. Robot. Autom., Barcelona, Spain, pp. 4080–4085.

Takata S, Isobe H, Fujii H (2001) Disassembly operation support system with motion monitoring of a human operator. Ann CIRP 50(1):305–308

Tang A, Owen C, Biocca F, Weimin M (2002) Experimental evaluation of augmented reality in object assembly task. In: Proceedings. International Symposium on Mixed and Augmented Reality, Darmstadt, Germany, pp 265–266. https://doi.org/10.1109/ISMAR.2002.1115105

Tao F, Cheng Y, Zhang L, Nee AYC (2017) Advanced manufacturing systems: socialization characteristics and trends. J Intell Manuf 28(5):1079–1094. https://doi.org/10.1007/s10845-015-1042-8

Thomas T, Sherman SR, Sawhney RS (2018) Application of lean manufacturing principles to improve a conceptual 238Pu supply process. J Manuf Syst 46:1–12. https://doi.org/10.1016/j.jmsy.2017.10.007

Thürer M, Stevenson M, Qu T, Godinho Filho M (2014) The design of simple subcontracting rules for make-to-order shops : an assessment by simulation. Eur J Oper Res 239:854–864. https://doi.org/10.1016/j.ejor.2014.06.018

Thürer M, Stevenson M, Qu T (2015) Simple subcontracting rules for make-to-order shops with limited subcontractor capacity: an assessment by simulation. Production Planning and Control

26(13):1145–1161. https://doi.org/10.1080/09537287.2015.1019590

Thürer M, Stevenson M, Land MJ (2016a) On the integration of input and output control: workload control order release. Int J Prod Econ 174:43–53. https://doi.org/10.1016/j.ijpe.2016.01.005

Thürer M, Stevenson M, Land MJ (2016b) On the integration of input and output control: workload control order release. Int J Prod Econ 174:43–53. https://doi.org/10.1016/j.ijpe.2016.01.005

Thürer M, Land MJ, Stevenson M, Fredendall LD (2017) On the integration of due date setting and order release control. Prod Planning Control 28(5):420–430. https://doi.org/10.1080/09537287.2017.1302102

Thürer M, Stevenson M, Land MJ, Fredendall LD (2018) On the combined effect of due date setting, order release, and output control: an assessment by simulation. Int J Prod Res. https://doi.org/10.1080/00207543.2018.1504250

Treleven M (1989) A review of the dual resource constrained system research. IIE Trans 21(3):279–287

Wang K, Rizqi DA, Nguyen H (2020) Skill transfer support model based on deep learning. J Intell Manuf. https://doi.org/10.1007/s10845-020-01606-w

Wight O (1970) Input/output control a real handle on lead time. Prod Invent Manage J 11(3):9–31

Wisner JD (1995) A review of the order release policy research. Int J Oper Prod Manage 15(6):25–40. https://doi.org/10.1108/01443579510090318

Xu J, Xu., X., & Xie, S.Q., (2011) Recent developments in Dual Resource Constrained (DRC) system research. Eur J Oper Res 215:309–318

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 Planning Control 27(13):1078–1091. https://doi.org/10.1080/09537287.2016.1183831

Yew AWW, Ong SK, Nee AYC (2016) Towards a griddable distributed manufacturing system with augmented reality interfaces. Robot Comput Integr Manuf 39:43–55. https://doi.org/10.1016/j.rcim.2015.12.002