Direct Workload Control:
Simplifying Continuous Order Release

Abstract

When Workload Control is applied, orders are withheld from the shop floor in a backlog from which they are released to meet certain performance metrics. This release decision precedes the execution of orders at shop floor stations. For each station there are consequently three types of workload: (i) indirect, i.e. released work that is still upstream of the station; (ii) direct, i.e. work that is currently at the station; and, (iii) completed, i.e. work that is still on the shop floor but is downstream of the station. Most Workload Control release methods control an aggregate workload made up of some representation of at least two of these three workload types. Yet the core objective of Workload Control release methods relates to only one of the three types – that is, to create a small, stable *direct* load in front of each station. Clearly, order release would be greatly simplified if only the direct load had to be considered. Using discrete event simulation, we show that Direct Workload Control leads to performance levels that match those of more complex and sophisticated approaches to Workload Control. Further, it greatly simplifies continuous Workload Control order release, decentralising the release decision by allowing it to be executed at each gateway station. This has important implications for research and practice.

Keywords: *Workload Control; Order Release; Job Shop; Simulation.*

1. Introduction

This study assesses the performance of a new continuous Workload Control order release method that significantly simplifies order release compared to existing release methods whilst maintaining the performance benefits of more sophisticated approaches in balanced shops that produce a high-variety of orders on a make-to-order basis. Workload Control order release was specifically developed for this type of high-variety make-to-order shops (Zäpfel & Missbauer, 1993; Stevenson *et al.*, 2005), where products typically have a specific routing (i.e. require more than one station to be completed) and order release only occurs after demand is known. A key challenge that these shops face is in striking a balance between the input rate of orders and their capacity (i.e. the output rate) to ensure that the shop and each station remains busy while simultaneously delivering confirmed orders in a timely fashion (Kingsman *et al*., 1989). If Workload Control order release is applied, jobs are not directly released to the shop floor but rather they are withheld in a so-called pre-shop pool (Melnyk & Ragatz, 1989) or backlog (Spearman *et al*., 1990) from where jobs are released to meet certain performance metrics, such as due date adherence, whilst keeping the workload within limits or norms. Thus, the release decision precedes the actual execution of the production process at downstream stations, and the further downstream a station is positioned in the routing of a job, the longer the time between release and production (Oosterman *et al*. 2000).

The time lag between the order release decision taking place and the actual materialisation of the workload at a given station prompted researchers to develop alternative workload accounting approaches (Land & Gaalman, 1996; Bergamaschi *et al*., 1997). These alternative approaches were based on the knowledge that all jobs released to the shop floor with a given station in their routing will materialise at this station at some point in time. This means that for each station there are three types of workload (Land & Gaalman, 1996): (i) the indirect load released to the shop floor but still upstream of the station; (ii) the direct load actually queuing (or being processed) at the station; and, (iii) the completed load still on the shop floor but downstream of the station. Thus, most order release methods presented in the Workload Control literature control some aggregated representation of at least two of these three load types. Yet, the objective of Workload Control order release is the creation of a small and stable direct load in front of each station (Thürer *et al*. 2012). In other words, the direct load should be small and should not fluctuate. This prompts us to ask: Why not simply control the direct load at each station instead of some form of an aggregated load?

There are three potential benefits of controlling the direct load only. First, it aligns more explicitly with Workload Control's main objective of creating a small and stable direct load in

front of each station. Second, it automatically incorporates the starvation avoidance mechanism proposed by Thürer *et al*. (2012) to overcome premature station idleness, which refers to the phenomenon whereby one station starves or runs idle because release to this station is blocked by the workload limit imposed at another station (Kanet, 1988; Land & Gaalman, 1998). Third, it greatly simplifies the release procedure (Bergamaschi *et al*., 1997) since direct load only occurs at the stations. This means workers at each gateway station can pull work from the pool directly whenever their workload allows for it. Stevenson *et al*. (2011) stated that, in most small and medium sized enterprises, the speed of information feedback is not quick enough to enable the effective implementation of a continuous order release method. Thus, periodic release methods dominate the literature on Workload implementation (e.g. Bechte, 1994; Wiendahl *et al*., 1992; Hutter *et al*., 2018; Hendry *et al*., 2013). However, only considering the direct load of the gateway station overcomes this practical limitation of continuous order release since all the information needed to support the release decision is available at the station. Decentralising the release decision in this way also overcomes the problem of a centralised planner tending to make release decisions only once a shift or day (e.g., Sabuncuoglu & Karapınar 1999; Stevenson *et al*. 2011; Thürer *et al*. 2012), thereby enabling more timely release decisions.

Despite of all the aforementioned advantages, the performance of a release method that only controls the direct load at each station has been rarely assessed in the Workload Control literature. A main exception is the station workload trigger presented by Melnyk & Ragatz (1989), which releases work whenever the direct workload falls below a certain triggering threshold. But this method has been outperformed by alternative order release methods that use an upper workload bound and consider an aggregated workload (e.g. Thürer *et al*., 2014). In this study, we will present a new continuous release method that uses an upper bound but in combination with the control of the direct load only. This new method significantly simplifies Workload Control order release. Discrete event simulation is then used to compare its performance against existing methods from the literature in balanced high-variety make-to-order shops. It is hoped that the results will facilitate more applications of Workload Control in practice given that the inherent complexities of Workload Control have been one of the major obstacles to its implementation (Stevenson *et al*., 2011).

The remainder of this paper is structured as follows. Section 2 provides the theoretical background to our study. It reviews the literature to identify Workload Control order release methods to be included in our study and it outlines the Workload Control order release method that is newly proposed in this study. The simulation model used to assess the performance of the alternative order release methods is then presented in Section 3 before the results are

presented and discussed in Section 4. Finally, conclusions are presented in Section 5, where managerial implications, limitations and future research directions are also presented.

2. Theoretical Background

Order release controls the release of work to the shop floor. A main objective is the stabilisation of the direct workload actually queuing at a station, i.e. the alignment of the input and output of work at each station (Wight, 1970). Meanwhile, this stabilisation should occur at low workload levels to reduce the level of work-in-process on the shop floor. A simple means of controlling (i.e. limiting and stabilising) station workloads is via the use of a workload limit or norm. This means that jobs are only released to the shop floor if their workload contribution does not violate the norm. So, order release controls the workload *released* to the shop floor in order to control the workload *queuing* at each station. If there is more than one station in the routing of jobs, then the workload released to a station *s* can be divided into three different parts, depending on the current progress of the set of jobs released onto the shop floor: L_s^U– the indirect load, i.e. work released but still upstream of station *s*; L_s^D– the direct load, i.e. work actually queueing or being processed at station *s*; and, L_s^C– the completed load, i.e. work completed and downstream of station *s*. A main difference between the various Workload Control release methods presented in the literature concerns the part of the workload that is subject to norms and thus controlled, i.e. whether the method controls the direct load only or some aggregated representation of the direct, indirect, and/or completed load.

2.1 Release Methods Controlling the Direct Load Only

These release methods use the direct load at stations to define when the pool should be inspected to see whether new jobs should be released (the direct load may hereby include or not include the load currently being processed at a station in addition to the queue). The station workload trigger activates the release procedure if L_s^D at a station *s* falls below the norm. Jobs in the pool for which the triggering station is the first station in their routing are considered for release according to the pool sequencing rule (e.g. with orders sequenced according to their planned release dates or based on the earliest due date). An example is the Work Centre workload trigger Earliest Due Date (WCEDD) selection method presented by Melnyk & Ragatz (1989). A station workload trigger was also used, for example, in Hendry & Wong (1994) and Sabuncuoglu & Karapinar (1999). Meanwhile, a version of the station workload trigger method that only controlled the bottleneck station was used, for example, in Glassey & Resende (1988) and Enns & Prongue-Costa (2002).

2.2 Release Methods Controlling the Indirect and Direct Loads

Two types of release methods have been presented in the literature that control the aggregate of L_s^U and L_s^D. The first type of release method controls the sum of all of the load released to a station *s* and not yet completed at a station *s*; i.e. the uncompleted station load L_s^A. The load of a job at station *s* is added to L_s^A at release and subtracted once the job is completed at station *s*. The second controls the sum of all load released and not yet completed across all stations, i.e. the uncompleted shop load.

The release procedure typically executed in the literature for methods that control the station load (e.g. Thürer *et al*. 2012; Fernandes *et al*. 2017) can be summarised as follows:

(1) All jobs in the set of jobs *J* in the pre-shop pool are sorted according to the priority determined by a pool sequencing rule (e.g. planned release dates). The job $j \in J$ with the highest priority is considered for release first.

(2) Take R_j to be the ordered set of operations in the routing of job *j*. If job *j*'s workload w_{ij} at the i^{th} operation in its routing together with the workload L_s^A released to station *s* (corresponding to operation *i*) and yet to be completed fits within the workload norm N_s at this station, that is $w_{ij} + L_s^A \leq N_s$ for all operations in the routing of the job, then the job is selected for release. That means it is removed from *J* and its load contribution w_{ij} is added to L_s^A for all operations in the routing of the job. Otherwise, the job remains in the pool and its processing time does not contribute to the station load.

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

Note that variants of this type of release method in the literature typically differ according to the way in which a job contributes to L_s^A. There are three key approaches: the classical aggregate load approach, which simply aggregates L_s^U and L_s^D using the full processing times p_{ij}, i.e. $w_{ij} = p_{ij}$ (see e.g. Bertrand & Wortmann, 1981; Hendry, 1989); the probabilistic approach, which converts L_s^U using a depreciation factor based on historical (probabilistic) data (see, e.g. Bechte, 1988 and 1994); and, the corrected aggregate load approach, which converts L_s^U by dividing the full job workload by the position of a station in the routing of a job, i.e. $w_{ij} = \frac{p_{ij}}{i}$ (Oosterman *et al.*, 2000).

Finally, methods that control the shop load activate the release procedure if the shop load falls below a predetermined load limit. Jobs are released onto the shop floor in accordance with

the pool sequencing rule applied (e.g. planned release dates). Examples of this type of order release method are the Aggregate workload trigger Work-in-Next-Queue (AGGWNQ) method presented by Melnyk & Ragatz (1989) and the WIPLoad control method applied by Qi *et al*. (2009).

2.3 Methods Controlling the Indirect, Direct and Completed Loads

There are two types of methods presented for controlling the aggregate of L_s^U, L_s^D and L_s^C. In an attempt to reduce the feedback requirements for the class of methods described in Section 2.2 above, Tatsiopoulos (1993) suggested only feeding back information after the completion of all operations of a job. This results in the so-called extended aggregate load method (Land & Gaalman, 1996). Meanwhile, authors such as Hendry & Wong (1994) and Sabuncuoglu & Karapinar (1999) adapted Melnyk & Ragatz's (1989) aggregated workload trigger to control the total shop load L^T, i.e. the sum of the total work content of all jobs on the shop floor. Note that this method is similar to Constant Work-In-Process (ConWIP) but controls the total shop load measured in processing time units instead of the number of jobs (Thürer *et al*., 2019)

2.4 Designing a New Method for Controlling the Direct Load Only

Order release is a main function of production control. Consequently, a broad set of different order release methods have emerged in the literature, specifically the Workload Control literature. A main objective of order release is the control of the workload actually queuing or being processed at a station, i.e. L_s^D. Only this direct load can act as an inventory buffer protecting the throughput of the station from variability. To realise lean production, L_s^D should be at a small and stable level (Thürer *et al*. 2012). However, most of the literature on order release does not directly control L_s^D. Rather, it uses some aggregate of L_s^U, L_s^D and L_s^C. Meanwhile, existing methods that focus on L_s^D do not limit the workload at a station. This restricts their workload balancing capabilities, and they were consequently outperformed by alternative release methods using an upper bound (e.g. Thürer *et al*. 2014). In response, this study outlines a new order release method – 'Direct Workload Control' – that controls the direct load using an upper bound in accordance with the following release procedure.

2.4.1 Direct Workload Control

Whenever a new job arrives in the pre-shop pool or a job's operation at a station on the shop floor is completed then:

(1) All jobs in the set of jobs J in the pre-shop pool are sorted according to the priority determined by a pool sequencing rule (e.g. planned release dates). The job $j \in J$ with the highest priority is considered for release first.
(2) If job j's processing time p_{1j} at the first operation in its routing together with the workload L_s^D queuing at station s (corresponding to operation i) fits within the workload norm N_s at this station, that is $p_{1j} + L_s^D \leq N_s$, then the job is selected for release. That means it is removed from J and its load contribution p_{1j} is added to the station load L_s^D. Otherwise, the job remains in the pool and its processing time does not contribute to the station load.
(3) If the set of jobs J in the pool contains any jobs that have not yet been considered for release, then return to Step 2 and consider the job with the next highest priority. Otherwise, the release procedure is complete, and the selected jobs are released to the shop floor.

An apparent drawback of the above method is that it only controls a proportion of the total workload released to the shop floor (i.e. the job workload at the first or gateway station). At the same time, the tightness of workload norms is restricted by the maximum processing times of jobs. That is, if workload norms are set too tight, jobs with a processing time larger than the load norm will never be released to the shop floor. To overcome this quandary, the following mechanism is introduced: whenever an operation is completed at a station and there is no job in its queue, the job $j \in J$ with the highest priority and with that station as the first in its routing is released regardless of whether or not it violates the workload norm at the station.

This new *Direct Workload Control* method provides a significant simplification when compared to alternative release methods that are focused on an aggregate load. It controls the release of work to gateway stations only based on the workload actually queuing and being processed at these stations. Thus, it can be implemented as a centralised release method or as a decentralised release method since workers at each station have the required information to make an informed decision to pull work from the pool. At the same time, the new method should yield similar benefits to more complex release methods since an upper workload bound is enforced. In order to prove this conjecture, we ask:

What is the performance impact of Direct Workload Control compared to alternative order release methods from the Workload Control literature?

Discrete event simulation will next be used to answer this research question.

3. Simulation model

In general, the simulation model can be described as follows. Jobs are created in accordance with the inter-arrival time distribution. The due date, the routing and the processing times are then assigned to the job as specified in detail in Section 3.1 below. Jobs then enter the backlog and wait for the release condition to become true. The release condition depends on the release method applied, as described in Section 3.2. Once released, the station loads are updated, and the release time is assigned to the jobs. Jobs will then request processing from all stations in their routing in accordance with the routing sequence. The dispatching rule applied to decide which job to process in a queue is described in Section 3.3. After processing has been performed at all stations, statistics concerning tardiness, the percentage of tardy jobs, the total throughput time and the shop floor throughput time are collected. These performance measures, together with the experimental setting, are described in Section 3.4.

3.1 Shop and Job Characteristics

We focus on shops with varying routing lengths, as is typical of many make-to-order shops in practice. The flow in these shops may be undirected or directed. Consequently, discrete event simulation models of two shops – a *pure job shop* and a *general flow shop* (Oosterman *et al*. 2000) – have been implemented using ARENA software. These simulation models are stochastic, whereby job inter-arrival times, routings, operation processing times and due dates are stochastic random variables. Common random number streams were used to reduce variability across experiments. Each shop contains six stations, where each station is modelled as a single constant capacity resource. We model a balanced shop to avoid distracting our focus away from our core research question to the problems created by bottlenecks.

To enable comparison with prior Workload Control literature, the parameters chosen for job and shop characteristics are similar to, for example, Oosterman *et al*. (2000), Land & Gaalman (1998) and Thürer *at el*. (2020). The routing length varies uniformly from one to six operations. All stations have an equal probability of being visited and a station is required at most once in the routing of a job. In the general flow shop, the resulting routing vector is sorted so there are typical upstream and downstream stations. Operation processing times follow a truncated Erlang-2 distribution with a maximum of 4 time units and a mean of 1 time unit before truncation. The maximum is based on the workload norms applied. Set-up times are assumed to be sequence independent, and hence part of the operation processing times. The inter-arrival time of jobs to the production system follows an exponential distribution with a mean that, based on the number of operations in the routing of a job, deliberately results in a 90%

utilisation level across experiments. To ensure comparability, utilization levels were measured for all experiments. Finally, due dates are set exogenously by adding an allowance to the job entry time. This allowance is uniformly distributed between 35 and 55 time units. These values were set arbitrarily to result in a percentage tardy that is neither too high nor too low. The percentage tardy should not be too high to avoid certain adverse effects, since rules that reduce the variance of lateness across jobs may even lead to an increase in the percentage tardy when due date allowances are too tight on average. The percentage tardy should not be too low to avoid our results being affected by incidental effects, as very few jobs would be responsible for the performance of the shop.

3.2 Workload Control Order Release

Four Workload Control order release methods are used: Station Workload Trigger, Corrected Aggregate Load, Total Shop Load, and Direct Workload Control. All four use a continuous timing convention, i.e. release may occur at any moment in time, triggered by a certain event (Bergamaschi *et al.*, 1997). In our case, this is whenever a new job arrives in the pre-shop pool or a job's operation at a station on the shop floor is completed. Each method is briefly described in Table 1 together with the workload norm settings applied. For each method, six levels of the norm are used. Different settings for the workload norm are considered since we cannot predict in advance which setting will lead to the best performance. As in previous simulation studies assessing the performance of Workload Control order release (e.g. Thürer *et al*., 2012), the spectrum for the workload limit was chosen such that we capture the best performance across all performance measures considered in this study.

[Take in Table 1]

Two alternative rules are applied in order to prioritise orders in the pool: the Planned Release Date (PRD) rule, which is a standard rule commonly applied in the Workload Control literature, and the Modified Capacity Slack (ModCS) rule, which was identified as best-performing by Thürer *et al*. (2015). The planned release date τ_j of a job j is calculated by $\tau_j = \delta_j - \sum_{s \epsilon R_j} b_s$ where δ_j is the due date of job j and b_s is the planned operation throughput time at station s. Planned operation throughput times are given by the cumulative moving average, i.e. the average of all operation throughput times realised until the current simulation time.

The ModCS rule divides the set of jobs in the pool into two classes: urgent, i.e. jobs with a Planned Release Date (PRD) that has already passed, and non-urgent jobs. Urgent jobs always receive priority over non-urgent jobs and are sequenced according to the lowest capacity slack

ratio S_j, (see e.g. Philipoom *et al*. 1993), which is calculated as $\sum_{i \in R_j} \left(\frac{p_{ij}}{N_s - L_s^D} \right)$ divided by the routing length n_j. This capacity slack ratio integrates three elements into one priority measure: the processing time p_{ij}, the load gap $N_s - L_s^D$, and the routing length n_j, i.e. he number of stations in the routing of job j. Non-urgent jobs are sequenced according to the earliest PRD.

Finally, the capacity slack ratio could become negative, which could result in the sequencing rule prioritising a job that contributes to the workload of an already overloaded station. Therefore, if the workload of a station is equal to or exceeds the workload norm, that is $N_s - L_s^D \leq 0$, then the job is positioned at the back of the queue by replacing the component (p_{ij}), related to this station in the priority value S_j by $(p_{ij} \cdot M)$, where M is a sufficiently large number.

3.3 Shop Floor Dispatching

Only one dispatching rule is applied to keep our study focused on order release control. Jobs waiting in a queue are prioritised according to operation due dates since this was shown to perform well in shops with high-variety routings (Kanet & Haya, 1982). The operation due date for the last operation in the routing of a job is equal to the due date, while the operation due date of each preceding operation is determined by successively subtracting the planned operation throughput time from the operation due date of the next operation. In this study, the planned operation throughput time is given by the cumulative moving average, i.e. the average of all operation throughput times realised until the current simulation time.

3.4 Experimental Design and Performance Measures

The experimental factors considered in the study are: (i) the four different order release methods; (ii) the two different pool sequencing rules (PRD and ModCS); (iii) the six different norm levels; and, (iv) the two shop configurations (i.e. the *pure job shop* and *general flow shop*). A full factorial design with 96 scenarios (4x2x6x2) was used. Each experimental scenario was replicated 100 times. All results were collected over 10,000 time-units following a warm-up period of 3,000 time-units. We used a commercial software package with an integrated random number generator, where each replication used a different random number stream. The random number stream is kept identical across the control strategies, i.e. the common random number stream technique is used. This is ensured by using the same simulation model for all strategies, i.e. only parameterizing the single model for a given strategy.

Finally, since we focus on a make-to-order context, our main performance criterion is delivery performance. In this study delivery performance will be measured by three main performance measures as follows: (i) the mean total throughput time, i.e. the mean difference between the arrival time and completion time of a job; (ii) the percentage of tardy jobs, i.e. the percentage of jobs completed after the due date; and, (iii) the mean tardiness of jobs. The percentage tardy provides the most general indication of delivery performance while the total throughput time indicates the mean lateness. Meanwhile, both the mean tardiness and the standard deviation of lateness can be used to measure the dispersion of lateness across jobs. We decided to measure the mean tardiness since the standard deviation of lateness is more sensitive to extreme values than the mean tardiness. In addition to these performance indicators, we also measure the mean shop floor throughput time, i.e. the mean difference between the release and completion time of a job. While the total throughput time includes the time that a job waits before being released into production, the shop floor throughput time only measures the time after the job has been released to the shop floor.

4. Results

Statistical analysis of our results was conducted using an ANOVA (Analysis of Variance). The results are presented in the appendix. All main effects and the majority of the two-way interactions were shown to be statistically significant at α=0.05, while there were also significant three-way interactions.

The Scheffé multiple comparison procedure was applied to obtain a first indication of the direction and size of the performance differences for our four release methods and our two pool sequencing rules. The Scheffé multiple comparison procedure was chosen since it is more conservative than, for example, the Tukey multiple comparison procedure. The results – as presented in Table 2 for the pure job shop and in Table 3 for the general flow shop – indicate significant differences for at least two performance measures for each pair. In general, the results indicate that Direct Workload Control and the Corrected Aggregate Load method perform the best, with the Corrected Aggregate Load method performing statistically better in terms of the percentage tardy and Directed Workload Control performing statistically better in terms of the mean tardiness. Similarly, ModCS pool sequencing performs statistically better in terms of the percentage tardy compared to PRD pool sequencing, but the former is outperformed by the latter in terms of the mean tardiness. To further assess these performance differences, detailed performance results will be presented next in Section 4.1 for the pure job shop. Section 4.2 then presents the results for the general flow shop to assess the robustness of

our results to changes in routing direction. This is followed by a more in-depth performance analysis in Section 4.3.

[Take in Table 2 & Table 3]

4.1 Performance Assessment in the Pure Job Shop

The simulation results are presented in the form of performance curves to aid interpretation. Data points correspond to the six workload norm levels, with the left-hand starting point of the curves representing the lowest workload norm. The workload norm increases stepwise by moving from left to right in each graph. Loosening the norm increases the level of work-in-process and, as a result, lengthens the shop floor throughput time. In addition, the results for immediate release are given by a single data point. These results are located to the right in each graph since they lead to the highest level of work-in-process. Figures 1a and 1b show the total throughput time, percentage tardy, and mean tardiness over the shop floor throughput time results in the pure job shop for PRD and ModCS pool sequencing, respectively. The performance of the release methods in the general flow shop will be assessed in Section 4.3, i.e. in our robustness analysis.

[Take in Figure 1]

The following can be observed from the results:

- *Release Method Performance*: As somewhat expected, based on previous literature, the Corrected Aggregate Workload method leads to better performance across all three main performance measures compared to the Station Workload Trigger, which in turn performs better than the Total Shop Load method. Meanwhile, Direct Workload Control matches the performance of the Corrected Aggregate Load method – the best-performing method – across all main performance measures considered in this study. A tighter workload norm restricts the work-in-process and thus leads to shorter shop floor throughput times, as can be observed from Figure 1. Once the total throughput time is equal to the shop floor throughput time plus the pool time, a tighter limit also leads to a shorter total throughput time. However, if the norm is set excessively tight, waiting times in the pre-shop pool are not compensated for by the shorter throughput times on the shop floor, and thus the total throughput time increases. Both, whether the initial gain in total throughput time can be realised and the specific norm level at which performance starts to deteriorate, will be dependent on the workload balancing capabilities of the order release method. Meanwhile, the gain in total throughput time (and thus mean lateness) leads to a reduction in the

percentage tardy until it is offset by an increase in the dispersion of lateness, which leads to an increase in the percentage tardy. This increase in the dispersion of lateness is reflected in the mean tardiness performance.

- *Impact of Pool Sequencing Rule*: ModCS allows for a reduction in terms of the percentage tardy, but this is at the expense of an increase in the mean tardiness. PRD sequencing is therefore considered to be a better option in the context of our study where only continuous order release methods are applied.

4.2 Robustness Analysis: Results for the General Flow Shop

Similar conclusions in terms of the ranking of the release methods to those for the pure job shop can be obtained for the general flow shop. This can be observed from Figures 2a and 2b, which show the total throughput time, percentage tardy, and mean tardiness over the shop floor throughput time results in the general flow shop for PRD and ModCS pool sequencing, respectively. However, Direct Workload Control appears to perform worse compared to the pure job shop in terms of the percentage tardy. This is because Direct Workload Control controls the gateway station. While in the pure job shop all stations have an equal probability of being the gateway station, in the general flow shop upstream stations have a higher probability of being the gateway, and consequently are controlled more tightly. This in turn further favours jobs with long routings, which are more likely to have an upstream station as the first in their routing (Thürer *et al.*, 2012). This results in an increase in the percentage tardy whilst maintaining good mean tardiness performance, as can be observed from Figure 2.

[Take in Figure 2]

4.3 Performance Analysis

The objective of Workload Control order release is the creation of a small and stable direct load in front of each station. To assess whether this is realized by our release methods, we measured the Coefficient of Variation (CV) of the direct load. The results are given in Table 4. Only results for PRD pool sequencing are given, since the results for ModCS pool sequencing were similar.

[Take in Table 4]

The results in Table 4 show that the Corrected Aggregate Load realizes the lowest CV values in the pure job shop whilst Direct Workload Control realizes the lowest CV values in the general flow shop. This was somehow expected given that the majority of orders enter the shop

floor at Station 1 in the general flow shop, whereas in the pure job shop the probability is equal across stations. Direct Workload Control controls the gateway station in the general flow shop more efficiently.

Meanwhile, while the Corrected Aggregate Load method and Direct Workload Control appear to perform similar on average, there may be significant performance differences across job classes. To assess potential differences, we collected results separately for each possible routing length n_j. The results are given in Table 5 for the Corrected Aggregate Load method and in Table 6 for Direct Workload Control. The best-performing norm level across performance measures is highlighted in bold in the tables. Only results for the pure job shop and PRD pool sequencing are given in the tables.

[Take in Table 5 & Table 6]

For the Corrected Aggregate Workload method, all operations in the routing of a job need to fit within the respective workload norms. This hinders the release of jobs with long routings as they have to fit within more norms to be released compared to jobs with short routings. As a consequence, jobs with short routings realise better percentage tardy and mean tardiness performance in Table 5. In contrast, for Direct Workload Control, jobs only need to fit within one workload norm – the one at the first station in the routing of a given job. Thus, the workload norm does not introduce differences across jobs with different routing lengths. However, the PRD tends to be earlier for jobs with long routings compared to jobs with short routings given that the PRD calculation does consider the routing length of jobs. As a consequence, Direct Workload Control tends to favour jobs with long routings thereby resulting in better percentage tardy and mean tardiness performance for these jobs in Table 6.

There is one further important effect for Direct Workload Control. The release of a job with a processing time larger than the norm is postponed until the queue at the job's gateway station becomes empty. To further investigate this effect, Table 7 summarises the performance of jobs for which $p_{1j} > N_s$, $p_{1j} \leq N_s$ and gives the average across all jobs for Direct Workload Control and PRD pool sequencing.

[Take in Table 7]

We can observe that Direct Workload Control favours jobs for which $p_{1j} \leq N_s$. Jobs for which $p_{1j} > N_s$ have long total throughput times and short shop floor throughput times. The total throughput time consists of the pool wating time and the shop floor throughput time. The

results consequently indicate that jobs for which $p_{1j} > N_s$ tend to be first delayed at release and then expedited on the shop floor by the dispatching rule, resulting in shorter shop floor throughput times. However, this speeding up behaviour does not outweigh the delay at release, which results in overall longer total throughput times and, as a result, an increase in both the percentage tardy and mean tardiness. Previous research would suggest that much stronger performance improvements can be obtained by delaying jobs with long processing times (Land *et al.*, 2010). But for Direct Workload Control, only p_{1j} is considered and not the workload contribution at downstream stations.

5. Conclusions

Workload Control is a production control concept specifically developed for high-variety make-to-order shops. If Workload Control order release is applied, jobs are not directly released to the shop floor but withheld in a pool or backlog from where they are released to meet certain performance metrics. This means that the release decision precedes the actual execution of the production process at downstream stations and, as a consequence, for each station there are three types of workload: the indirect load released to the shop floor but still upstream of the station, the direct load actually queuing (or being processed) at the station, and the completed load downstream of the station. Most of the order release methods presented in the Workload Control literature control some aggregate form of at least two of these three load types. Yet the objective of Workload Control order release is the creation of a small and stable direct load in front of each station. Therefore, we have designed a new order release method – Direct Workload Control – which directly controls the direct load at each station and, based on insights from prior Workload Control literature, uses an upper bound on this direct load. Using simulation, we have shown that controlling the direct load at each station only can match the performance of more sophisticated approaches to Workload Control with the results being robust to shop type (i.e. both the pure job shop and general flow shop). This has important implications for practice and research, which will be discussed next.

5.1 Managerial Implications

Direct Workload Control can realise performance results that match those of more sophisticated order release methods, but with much lower solution complexity. Direct Workload Control significantly simplifies workload calculations since only the load currently at a station needs to be tracked. In addition, there is another advantage: Direct Workload Control only controls releases to gateway stations. This means that release can not only be

executed as a centralised decision, but it can also be decentralised. It can be executed independently by each gateway station since each gateway station has all of the information it requires to make the decision (i.e. the only information required is regarding the workload currently at this station). The worker can simply pull new work in whenever the workload at his/her station allows for it.

Finally, release methods for which jobs have to fit the workload norm at all stations in their routing favour jobs with short routings while Direct Workload Control does not introduce any performance difference. In practice, the choice of release method consequently depends on the proportion of jobs with long routings in the company's current job mix and the way in which the performance of the company is measured (i.e. is one on-time job with a short routing evaluated in the same way as one on-time job with a long routing, or is the total work content of a job considered when determining shop performance?). A small performance loss for jobs with short routings may be acceptable if the performance of jobs with long routings is clearly improved.

5.2 Limitations and Future Research Directions

A main limitation of our study is the restricted environmental setting. For example, we have only considered one level of processing time variability and one level of due date tightness. While we consider this to be justified by the need to keep the study focused, future research could extend our study by considering a broader set of environmental factors. Future research could also explore different approaches to controlling the direct load. Finally, a main task for future research is the implementation of our new release method in practice. While Workload Control has been widely advocated in the literature as a good solution for high-variety make-to-order shops, reports on its successful implementation in practice are few and far between. One major challenge to implementation has been the complexity of Workload Control order release in terms of its workload calculations and its requirements for information feedback from the shop floor on changes to the workload (or the progress of jobs). Our new release method significantly simplifies Workload Control order release while maintaining its performance. Thus, it is hoped that this will trigger more implementations of this important production control concept.

References

Bechte, W., 1988, Theory and practise of load-oriented manufacturing control, *International Journal of Production Research*, 26, 3, 375 – 395.

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

Bergamaschi, D., Cigolini, R., Perona, M., and Portioli, A., 1997, Order review and release strategies in a job shop environment: A review and a classification, *International Journal of Production Research*, 35, 2, 399-420.

Bertrand, J.W.M., and Wortmann, J.C., 1981, *Production control and information systems for component-manufacturing shops*, Elsevier Scientific Publishing Company, Amsterdam.

Enns, S.T., and Prongue Costa, M., 2002, The effectiveness of input control based on aggregate versus bottleneck workloads, *Production Planning and Control*, 13, 7, 614 - 624.

Fernandes N.O., Thurer, M., Silva, C., Carmo-Silva, S., 2017, Improving Workload Control Order Release: Incorporating a Starvation Avoidance Trigger into Continuous Release, *International Journal of Production Economics*, 194, 181-189.

Glassey, C.R., and Resende, M.G., 1988, Closed-loop job release control for VLSI circuit manufacturing, *IEEE Transactions on Semiconductor Manufacturing*, 1, 36 – 46.

Hendry, L.C., 1989, *A decision support system to manage delivery and manufacturing lead times in make to order companies*, PhD Thesis, Lancaster University, Lancaster, UK.

Hendry, L.C., and Wong, S.K., 1994, Alternative order release mechanisms: A comparison by simulation, *International Journal of Production Research*, 32, 12, 2827 – 2842.

Hendry, L.C., Huang, Y., and Stevenson, M., 2013, Workload control: Successful implementation taking a contingency-based view of production planning & control, *International Journal of Operations & Production Management*, 33, 1, 69-103.

Hutter, T., Haeussler, S., and Missbauer, H., 2018, Successful implementation of an order release mechanism based on workload control: a case study of a make-to-stock manufacturer, *International Journal of Production Research*, 56, 4, 1565-1580.

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

Kanet, J.J., 1988, Load-limited order release in job shop scheduling systems, *Journal of Operations Management*, 7, 3, 44-58.

Kingsman, B.G., Tatsiopoulos, I.P., and Hendry, L.C., 1989, A structural methodology for managing manufacturing lead times in make-to-order companies, *European Journal of Operational Research*, 40, 196 – 209.

Land, M.J., and Gaalman, G.J.C., 1998, The performance of workload control concepts in job shops: Improving the release method, *International Journal of Production Economics*, 56-57, 347-364.

Land, M.J., and Gaalman, G.J.C, 1996, Workload control concepts in job shops: A critical assessment, *International Journal of Production Economics*, 46–47, 535–538.

Land, M.J., Su, N.P.B. and Gaalman, G..J.C, 2010, In search of the key to delivery improvement, *16th International Working Seminar on Production Economics, 1st – 5th March, Innsbruck, Austria, Conference Proceedings, 2,* 297-308.

Melnyk, S.A., and Ragatz, G.L., 1989, Order review/release: research issues and perspectives, *International Journal of Production Research*, 27, 7, 1081 – 1096.

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

Qi, C., Sivakumar, A.I., and Gershwin S.B., 2009, An efficient new job release control methodology, *International Journal of Production Research*, 47, 3, 703 – 731.

Sabuncuoglu, I., and Karapinar, H.Y., 1999, Analysis of order review/release problems in production systems, *International Journal of Production Economics*, 62, 259 - 279.

Spearman, M.L., Woodruff, D.L., Hopp, W.J., 1990, CONWIP: a pull alternative to Kanban, *International Journal of Production Research*, 28, 5, 879-894.

Stevenson, M., Huang, Y., Hendry, L.C. and Soepenberg, E., 2011, The theory and practice of workload control: A research agenda and implementation strategy, *International Journal of Production Economics*, 131, 2, 689-700.

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

Tatsiopoulos, I.P., 1993, Simplified production management software for the small manufacturing firm, *Production Planning & Control*, 4, 1, 17-26.

Tatsiopoulos, I.P., and Kingsman, B.G., 1983, Lead time management, *European Journal of Operational Research*, 14, 351 – 358.

Thürer, M., Fernandes, N.O., and Stevenson, M., 2020, Material Flow Control in High-Variety Make-to-Order Shops: Combining COBACABANA and POLCA, *Production & Operations Management*, (in print)

Thürer, M., Fernandes, N.O., Ziengs, N., and Stevenson, M., 2019, On the Meaning of ConWIP Cards: An Assessment by Simulation, *Journal of Industrial & Production Engineering*, 36, 1, 49-58.

Thürer, M., Land, M.J., Stevenson, M., Fredendall, L.D., and Godinho Filho, M., 2015, Concerning Workload Control and Order Release: The Pre-Shop Pool Sequencing Decision, *Production & Operations Management*, 24, 7, 1179-1192.

Thürer, M., Qu, T., Stevenson, M., Maschek, T., and Godinho Filho, M., 2014, Continuous Workload Control Order Release Revisited: An Assessment by Simulation, *International Journal of Production Research*, 52, 22, 6664-6680.

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

Wiendahl, H.P., Gläßner, J., and Petermann, D., 1992, Application of load-oriented manufacturing control in industry, *Production Planning & Control*, 3, 2, 118 – 129.

Wight, O., 1970, Input/Output control a real handle on lead time, *Production and Inventory Management Journal*, 11, 3, 9-31.

Zäpfel, G. and Missbauer, H., 1993, New concepts for production planning and control, *European Journal of Operational Research*, 67, 297-320.

Table 1: Summary of Order Release Methods Applied

Release Method	Load Controlled?	Norm Setting
Station Workload Trigger	Jobs are released until $L_s^D > N_s$	N_s= 2, 4, 6, 8, 10 and 12
Corrected Aggregate Load	A job is only released if $w_{ij} + L_s^A \leq N_s$ $\forall i \in R_j$	N_s= 4, 6, 8, 10, 12 and 14
Total Shop Load	Jobs are released until $L^T > N$	N_s= 120, 140, 160, 180, 200 and 220
Direct Workload Control	A job is only released if $w_{1j} + L_s^D \leq N_s$, but large jobs may violate the norm if the queue is empty after an operation has been completed.	N_s= 1, 2, 3, 4, 5, and 6

Table 2: Results for Scheffé Multiple Comparison Procedure: Pure Job Shop

	Rule (x)	Rule (y)	Total Throughput Time		Percent Tardy Jobs		Mean Tardiness	
			lower[1)]	upper	lower	upper	lower	upper
Relese Methods	SWT[2)]	DWLC	1.83	2.32	1.00	1.55	0.16	0.36
	TSL	DWLC	3.38	3.86	3.22	3.77	-0.09*	0.12
	CAL	DWLC	0.64	1.13	-0.46*	0.09	0.53	0.73
	TSL	SWT	1.30	1.78	1.94	2.49	-0.34	-0.14
	CAL	SWT	-1.44	-0.95	-1.74	-1.19	0.27	0.47
	CAL	TSL	-2.98	-2.49	-3.96	-3.41	0.51	0.72
Pool Sequencing Rule	ModCS	PRD	-0.35	-0.11	-1.76	-1.49	0.21	0.31

[1)] 95% confidence interval; * not significant at 0.05
[2)] Station Workload Trigger (SWT); Corrected Aggregate Load (CAL); Total Shop Load (TSL); Direct Workload Control (DWLC)

Table 3: Results for Scheffé Multiple Comparison Procedure: General Flow Shop

	Rule (x)	Rule (y)	Total Throughput Time		Percent Tardy Jobs		Mean Tardiness	
			lower[1)]	upper	lower	upper	lower	upper
Relese Methods	SWT[2)]	DWLC	1.25	1.77	0.11	0.81	0.30	0.51
	TSL	DWLC	3.03	3.55	3.15	3.85	0.16	0.36
	CAL	DWLC	-0.43*	0.09	-1.36	-0.67	0.08	0.28
	TSL	SWT	1.52	2.05	2.69	3.39	-0.25	-0.04
	CAL	SWT	-1.94	-1.42	-1.82	-1.13	-0.33	-0.12
	CAL	TSL	-3.72	-3.20	-4.86	-4.17	-0.18*	0.02
Pool Sequencing Rule	ModCS	PRD	-0.28	-0.02	-1.32	-0.98	0.08	0.18

[1)] 95% confidence interval; * not significant at 0.05
[2)] Station Workload Trigger (SWT); Corrected Aggregate Load (CAL); Total Shop Load (TSL); Direct Workload Control (DWLC)

Table 4: Coefficient of Variation of Direct Load: PRD Pool Sequencing

Release Method	Parameter	Pure Job Shop	General Flow Shop					
			Station 1	Station 2	Station 3	Station 4	Station 5	Station 6
IMM	None	1.11	1.30	1.40	1.40	1.40	1.41	1.39
Station Workload Trigger	2	1.35	1.09	1.27	1.30	1.30	1.31	1.29
	4	1.35	1.11	1.29	1.29	1.28	1.29	1.28
	6	1.36	1.12	1.29	1.29	1.29	1.30	1.29
	8	1.36	1.14	1.30	1.30	1.30	1.32	1.31
	10	1.37	1.16	1.31	1.31	1.32	1.33	1.32
	12	1.67	1.18	1.32	1.33	1.33	1.35	1.34
Direct Workload Control	1	1.32	0.87	1.36	1.41	1.40	1.38	1.36
	2	1.35	1.01	1.41	1.40	1.38	1.38	1.36
	3	1.37	1.28	1.39	1.37	1.37	1.38	1.37
	4	1.38	1.44	1.35	1.35	1.36	1.38	1.37
	5	1.39	1.45	1.33	1.35	1.37	1.39	1.38
	6	1.69	1.41	1.33	1.36	1.37	1.39	1.38
Corrected Aggregate Load	4	0.83	1.50	1.46	1.41	1.38	1.35	1.31
	6	0.85	1.47	1.41	1.35	1.31	1.29	1.26
	8	0.89	1.39	1.36	1.32	1.30	1.30	1.28
	10	0.94	1.33	1.34	1.32	1.31	1.32	1.31
	12	0.99	1.29	1.33	1.33	1.34	1.35	1.34
	14	1.23	1.28	1.34	1.35	1.36	1.37	1.36
Total Shop Load	120	1.51	1.55	1.54	1.51	1.49	1.49	1.47
	140	1.46	1.46	1.48	1.45	1.44	1.44	1.42
	160	1.44	1.40	1.43	1.42	1.41	1.42	1.40
	180	1.44	1.36	1.42	1.40	1.40	1.41	1.39
	200	1.44	1.34	1.40	1.39	1.39	1.40	1.39
	220	1.73	1.33	1.40	1.39	1.39	1.40	1.39

Table 5: Performance Across Routing Length: Corrected Aggregate Load and PRD Pool Sequencing

	Routing Length (RL)	N_s=4	**N_s=6**	N_s=8	N_s=10	N_s=12	N_s=14
Total Throughput Time	RL 1	13.93	**12.46**	13.57	14.16	14.43	15.54
	RL 2	19.86	**18.31**	19.55	20.25	20.61	20.81
	RL 3	23.95	**21.07**	22.14	22.84	23.26	23.49
	RL 4	26.68	**22.47**	23.26	23.95	24.41	24.65
	RL 5	28.58	**23.31**	23.90	24.54	25.01	25.30
	RL 6	30.28	**24.02**	24.42	24.97	25.41	25.68
PercentTardy Jobs	RL 1	3.28	**1.64**	1.08	0.85	0.91	1.98
	RL 2	4.39	**2.17**	1.43	1.19	1.41	2.02
	RL 3	5.24	**2.55**	1.65	1.34	1.72	2.51
	RL 4	6.20	**2.90**	1.84	1.52	1.93	2.78
	RL 5	7.01	**3.20**	1.93	1.61	2.04	2.96
	RL 6	7.86	**3.52**	2.13	1.72	2.10	3.08
Mean Tardiness	RL 1	4.32	**0.88**	0.55	0.36	0.25	0.19
	RL 2	5.00	**1.09**	0.68	0.45	0.30	0.24
	RL 3	5.99	**1.27**	0.78	0.50	0.35	0.26
	RL 4	6.75	**1.44**	0.82	0.53	0.37	0.28
	RL 5	7.30	**1.55**	0.88	0.54	0.37	0.30
	RL 6	7.91	**1.69**	0.97	0.58	0.38	0.29

Table 6: Performance Across Routing Length: Direct Workload Control and PRD Pool Sequencing

	Routing Length (RL)	N_s=1	**N_s=2**	N_s=3	N_s=4	N_s=5	N_s=6
Total Throughput Time	RL 1	14.85	**12.71**	13.07	13.72	14.16	14.37
	RL 2	18.80	**18.10**	19.04	19.88	20.37	20.65
	RL 3	21.05	**20.70**	21.66	22.49	23.02	23.35
	RL 4	22.30	**21.93**	22.80	23.62	24.19	24.54
	RL 5	23.10	**22.62**	23.44	24.24	24.81	25.19
	RL 6	23.66	**23.16**	23.92	24.68	25.26	25.62
PercentTardy Jobs	RL 1	6.71	**2.88**	1.83	1.40	1.37	1.49
	RL 2	6.38	**2.85**	1.97	1.81	2.08	2.48
	RL 3	6.13	**2.89**	2.10	2.07	2.54	3.08
	RL 4	5.89	**2.88**	2.14	2.23	2.79	3.48
	RL 5	5.67	**2.86**	2.20	2.30	2.95	3.66
	RL 6	5.45	**2.89**	2.24	2.42	3.12	3.88
Mean Tardiness	RL 1	2.10	**1.27**	0.76	0.48	0.32	0.23
	RL 2	1.99	**1.27**	0.79	0.52	0.36	0.29
	RL 3	1.86	**1.25**	0.80	0.52	0.38	0.31
	RL 4	1.77	**1.23**	0.77	0.52	0.39	0.34
	RL 5	1.69	**1.18**	0.76	0.50	0.37	0.34
	RL 6	1.58	**1.17**	0.77	0.52	0.39	0.35

Table 7: Performance Across Regular and Large Jobs: Pure Job Shop, Direct Workload Control and PRD Pool Sequencing

		N_s=1	N_s=2	N_s=3	N_s=4	N_s=5	N_s=6
Shop Floor Throughput Time	All	11.90	14.59	17.09	18.96	20.23	21.07
	$P_{1j} > N_s$	10.19	11.25	12.26	None		
	$P_{1j} \leq N_s$	13.07	14.92	17.15	18.96	20.23	21.07
Total Throughput Time	All	20.63	19.87	20.66	21.44	21.97	22.29
	$P_{1j} > N_s$	27.75	37.05	47.64	None		
	$P_{1j} \leq N_s$	15.77	18.18	20.27	21.44	21.97	22.29
Percent Tardy Jobs	All	6.04	2.87	2.08	2.04	2.48	3.01
	$P_{1j} > N_s$	14.58	21.07	26.32	None		
	$P_{1j} \leq N_s$	0.21	1.08	1.73	2.04	2.48	3.01
Mean Tardiness	All	1.83	1.23	0.78	0.51	0.37	0.31
	$P_{1j} > N_s$	4.49	11.43	20.07	None		
	$P_{1j} \leq N_s$	0.02	0.23	0.50	0.51	0.37	0.31

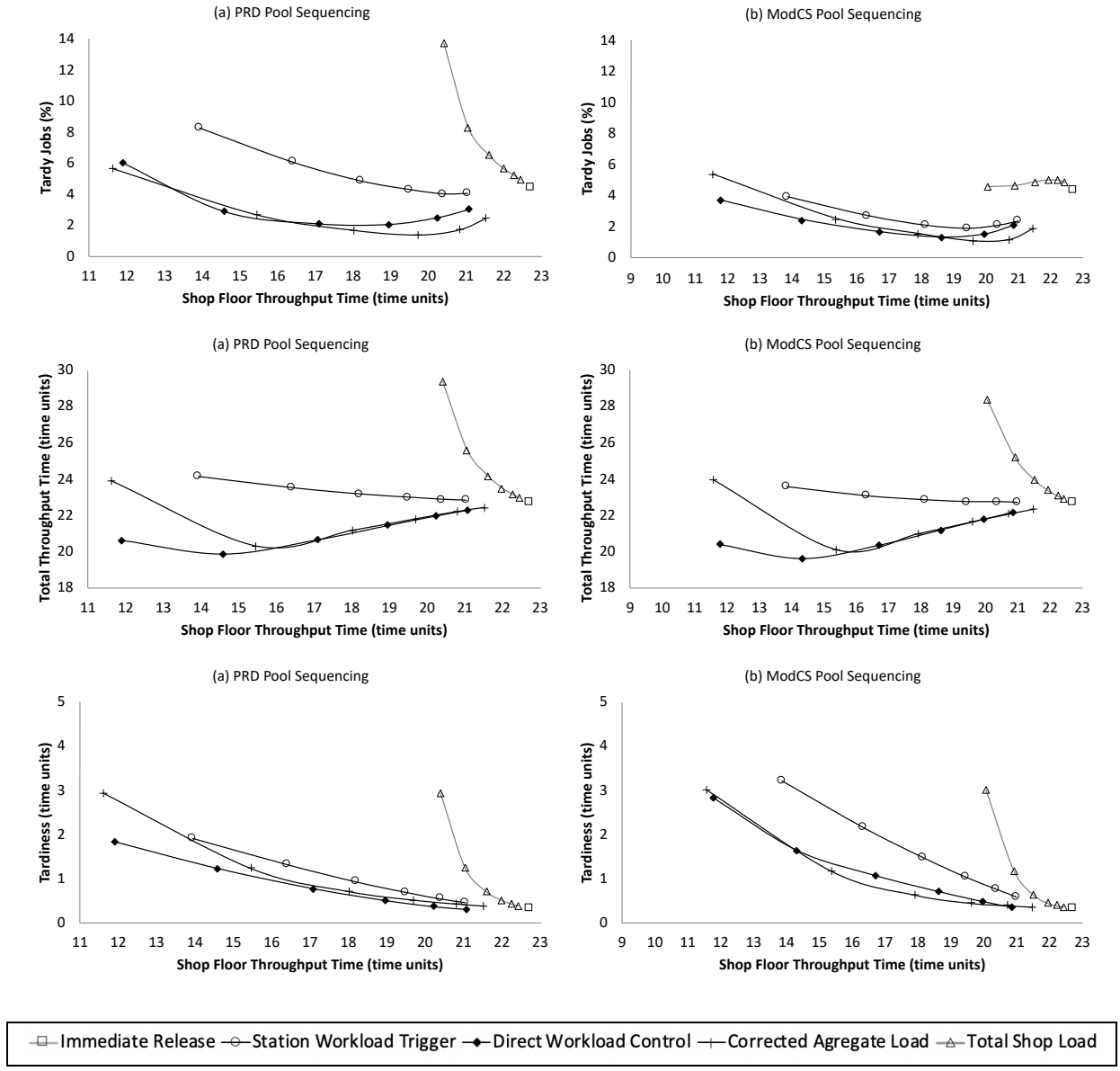

Figure 1: Performance Assessment in the Pure Job Shop

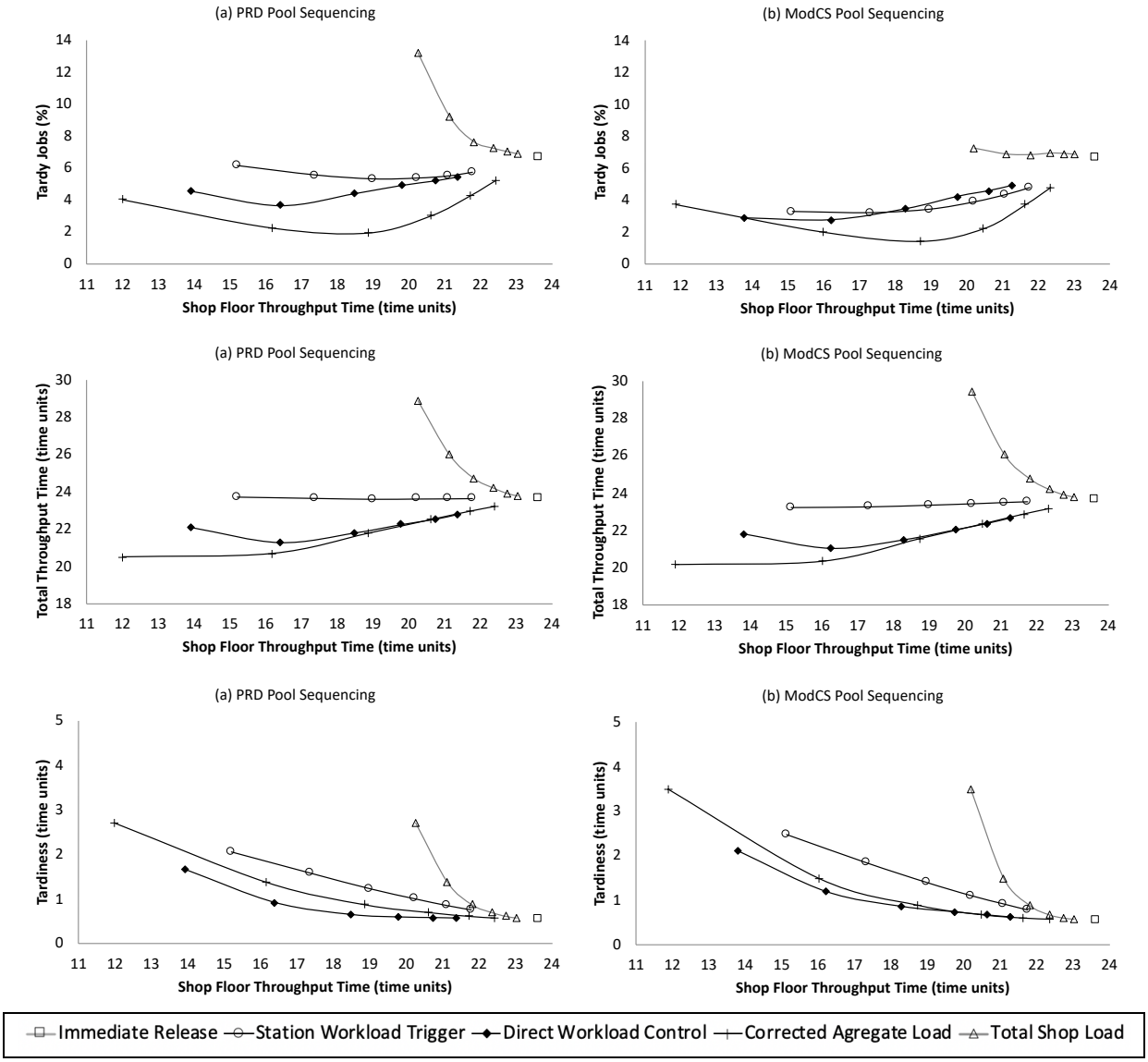

Figure 2: Performance Assessment in the General Flow Shop