University of Groningen

Aligning workload control theory and practice

Fernandes, Nuno O.; Land, Martin J.; Carmo-Silva, S.

Published in:
International Journal of Production Research

DOI:
[10.1080/00207543.2016.1143134](https://doi.org/10.1080/00207543.2016.1143134)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2016

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Fernandes, N. O., Land, M. J., & Carmo-Silva, S. (2016). Aligning workload control theory and practice: lot splitting and operation overlapping issues. *International Journal of Production Research*, 54(10), 2965-2975. <https://doi.org/10.1080/00207543.2016.1143134>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Aligning workload control theory and practice: lot splitting and operation overlapping issues

Nuno O. Fernandes^{a,c,*} , Martin J. Land^b and S. Carmo-Silva^c

^a*Escola Superior de Tecnologia, Instituto Politécnico de Castelo Branco, Castelo Branco, Portugal;* ^b*Faculty of Economics and Business, University of Groningen, Groningen, The Netherlands;* ^c*ALGORITMI research unit, University of Minho, Braga, Portugal*

(Received 3 July 2015; accepted 6 January 2016)

This paper addresses the problem of lot splitting in the context of workload control (WLC). Past studies on WLC assumed that jobs released to the shop floor proceed through the different stages of processing without being split. However, in practice, large jobs are often split into smaller transfer sublots so that they can move more quickly and independently through the production process and allow operations overlapping relating to the same job. This paper assesses the performance of different lot splitting policies for job release and dispatching strategies under lot splitting. A new dispatching rule was designed to specifically take advantage of lot splitting and operations overlapping in the context of WLC. Discrete-event simulation is used to assess system performance in relation to the ability to provide shorter delivery times and on time deliveries. Results highlight the importance of releasing the sublots of the same job together and demonstrate that combining an effective lot splitting policy with an appropriate dispatching rule can enhance the performance of production systems.

Keywords: workload control; lot splitting; release; dispatching

1. Introduction

Workload control (WLC) is a production planning and control system designed for the needs of the Make-to-order (MTO) industry (Stevenson, Hendry, and Kingsman 2005; Stevenson et al. 2011). WLC aims at firmly controlling job throughput times through the production system by means of input and output control decisions for meeting the promised delivery dates. Several WLC methods, varying in the degree of sophistication, have been described in the literature (see e.g. Land and Gaalman 1996). The most comprehensive methods comprise four stages, namely: customer enquiry, job entry, job release and priority dispatching. The methods have in common, that they use a pre-shop pool where the jobs wait for a release decision, and a release mechanism that regulates access to the shop floor (Hendry, Huang, and Stevenson 2013). This makes it possible to control the length of the queues on the shop floor, reduce shop congestion and increase load balancing among workstations, thus shortening and stabilising throughput times (Land and Gaalman 1996; Portioli-Staudacher and Tantardini 2011). WLC is specific in that it sets workload norms for all workstations on the shop floor, thus allowing for control under shifting bottlenecks with both balanced and unbalanced capacities (Fernandes, Land, and Carmo-Silva 2014).

Several simulation studies have demonstrated the effectiveness of WLC in the MTO production (e.g. Oosterman, Land, and Gaalman 2000; Land 2006; Thürer, Qu et al. 2014). The majority of the research in the field of WLC has been conducted on the assumption that released job lots proceed through the different stages of processing without being split. However, in practice, due to the large size of some customer orders, jobs are often split into smaller transfer sublots so they can move independently through the production process, increasing the likelihood of consecutive operation overlapping and smoothing the production flow. Current manufacturing trends also emphasise small lot sizes. This allows job progress to be accelerated and due date performance to be improved, as pointed out by Jacobs and Bragg (1988) and Wanger and Ragatz (1994), among other authors. It also reduces the amount of storage space as well as the capacity of material handling equipment required. Chang and Chiu (2005) and Cheng, Mukherjee, and Sarin (2013) are recommended for their comprehensive literature reviews on lot streaming, i.e. combined lot splitting with operations overlapping.

Prior research showed that the effects of lot splitting on performance depends on a variety of factors, including sublot sizes and set-up times. Smunt, Buss, and Kropp (1996) investigate various lot splitting policies in both stochastic job shop and flow shop environments. They found that the number of lot splits is more important than the exact form of

*Corresponding author. Email: nogf@ipcb.pt

splitting. Previously, Kropp and Smunt (1990) concluded that using equal-sized transfer sublots works well in many situations. However, large amounts of lot splitting may result in increased set-up frequency due to the reduction in transfer subplot size. This can offset gains in performance. Karmarkar et al. (1985) demonstrated that as subplot sizes begin to diminish, throughput times are reduced due to the overlapping of successive operations at machines. However, as sublots become smaller the number of set-ups also increases and at some point the additional set-up time overtakes the savings from overlapping, so that throughput times begin to increase. Ruben and Mahmoodi (1998) looked at the effects of lot splitting in unbalanced production systems. They found that when bottlenecks exist, their longer queues limit the benefits of operations overlapping. Several heuristic procedures have been employed to determine the subplot sizes; see e.g. Azzi, Maurizio, and Persona (2012), Buscher and Shen (2009), Jeong, Park, and Leachman (1999), Li (1997) and Kim, Kang, and Lee (1997), among others.

Although, much research has been done on lot splitting, nearly all studies assume that jobs are immediately released to the shop floor without considering the shop floor situation. An exception is a study by Russell and Fry (1997), who included the impact of lot splitting in the specific context of drum-buffer-rope (DBR). They found that significant improvement in shop performance could be realised if lot splitting is used. The DBR approach controls release of work to a shop floor with a specific bottleneck. Also, the use of lot splitting has been suggested in the broader concept of WLC. Hendry and Wong (1994) suggest the use of lot splitting or overlapping as a remedial action in situations where orders have a negative slack. Here, slack refers to the difference between the due date, the current date, and the time required to complete the job including expected queuing time. Wiendahl (1995) and Missbauer (2002) also suggest that the jobs with the largest work content should be split. However, until now these suggestions for the application of lot splitting in the context of WLC have never been tested and, as a consequence, their impact is not known.

Despite the fact that several WLC methods have been presented and evaluated in literature, relatively little is known about the performance impact of lot splitting in this context. Authors such as Fredendall, Divesh, and Patterson (2010), Thürer, Land et al. (2014) and Thürer, Qu et al. (2014) contributed to the WLC literature by highlighting the potential of pool sequencing and shop floor dispatching to improve performance, but without considering the additional possibility of only releasing or dispatching sublots of prioritised jobs. As such, these studies neglected lot splitting as an additional important source of performance improvement. In the context of WLC, the following research questions arise: (1) How should the job lots be released to the shop floor? Should they be split into sublots in the pre-shop pool and then be considered for releasing, or should they be first released and then split on the shop floor? (2) How should the sublots be dispatched on the shop floor? Will the traditional dispatching rules adopted in the WLC studies still apply, or should new rules be applied to take advantage of lot splitting?

This paper aims to reveal the performance advantages that can be gained when embedding lot splitting in the concept of WLC. A comprehensive simulation study is carried out to assess the performance of different lot splitting policies for job release and dispatching. In doing so, this research contributes to both practice and theory. For practitioners, this research provides insights into what to expect from implementing these splitting policies and dispatching strategies in practice. For academics, this research answers the *how* questions important to the embedding of lot splitting in the concept of WLC. In order to achieve its objectives, the paper is structured as follows. First, the expected implications of lot splitting in WLC is explained. The resulting focus of our study is elaborated in Section 2. The model and the experimental design are then described in Section 3. In Section 4, the simulation results are presented and discussed. Finally, conclusions and managerial guidelines are drawn and suggestions for further research are given in Section 5.

2. Implications of lot splitting for WLC

WLC has often been presented as the most suitable planning and control concept for high variety MTO companies (Stevenson, Hendry, and Kingsman 2005). However, a phenomenon indicated as premature idleness has been identified as one of the problems hindering the performance of WLC (Land and Gaalman 1998). Therefore, many studies have found exactly here the key to further improving the concept, e.g. Thürer et al. (2012). Kanet (1988) already concluded that throughput times deteriorate to the extent that the workload norms introduce idle time. Land and Gaalman (1998) labelled idleness as premature when a workstation becomes idle while a job with a first operation at that particular workstation is available for release. This may occur when the workload at a workstation exceeds the workload norm, preventing the release of jobs for immediately feeding idle workstations. This is a somewhat narrow view of premature idleness. Premature idleness may also occur when a workstation is starving while part of the next job is already processed at a preceding workstation. However, the successive operation may not be allowed to begin before the entire lot is completed on the previous workstation. In this case, operations overlapping may allow for increased throughput and throughput time reduction. Figure 1 schematically shows the logic of operations overlapping that is applied in this paper.

Figure 1. Operation overlapping.

In Figure 1 a job is divided into two sublots. When the first sublot is completed on operation A, it is transferred to operation B. While operation A continues with the second sublot, operation B starts on the first sublot. Throughput time may be reduced by the overlap time and by the elimination of queue time.

Splitting a large job into a number of smaller transfer sublots may have important benefits within WLC, namely:

- (1) It may facilitate the release of sublots of large jobs from the pool to the shop floor that as a whole do not fit within the workload norms.
- (2) Sublots can move independently and more quickly through the production process. The resulting overlap of consecutive operations will contribute to reducing the premature idleness of workstations.

Both release and dispatching decisions may thus be relevant when considering lot splitting. As such, two alternatives will be investigated, namely: splitting jobs before release to the shop floor and splitting jobs on the shop floor. The aim is to determine if the sublots of a job should best be considered for release together or independently.

Splitting jobs before release when they are in the pre-shop pool is expected to reduce the premature idleness mainly by facilitating the partial release of large jobs to the shop floor. However, when sublots are released separately, overlapping operations opportunities at the shop floor tend to be smaller than when sublots are released together. Moreover, when early delivery to the customer is not allowed a sublots' production synchronisation problem may also arise. This means that finished sublots need to wait for others still progressing in the shop floor before all the sublots of a job can be assembled and expedited to the customer.

Splitting jobs on the shop floor (i.e. after release) is expected to reduce the premature idleness by enabling overlapping of consecutive operations and thus shortening throughput times. However, when sublots are released together this may hinder a good timing of job release, particularly for large jobs which will have a smaller probability to fit the workload norms of workstations.

In what concerns the dispatching strategy, the earliest *planned operation start time* (PST) rule (see Section 3.3) is used as a baseline for experimentation, and a new rule was specifically designed to take advantage of lot splitting and operations overlapping under WLC.

The next section presents the details of the simulation study carried out to investigate the performance behaviour of the release policies in combination with the new dispatching rule. This is compared with the PST rule above referred.

3. Simulation study

3.1 Overview of the simulation model

A simulation model was developed using Arena[®] software. To improve the practical applicability of the findings, the model was kept simple in order to avoid interactions that might inhibit the full understanding of the effects of the experimental factors. Table 1 summarises the characteristics of the simulation model.

A six-workstation general flow shop is considered, which is seen as a more realistic model of the flow structure of job shops in practice than pure job shops (Enns 1995; Oosterman, Land, and Gaalman 2000). Each workstation consists of a single resource, i.e. machine, with constant capacity.

In the model, arriving job lots are split into smaller sublots of equal size which can be processed independently in the different stages of job progress. Thus, sublots may move from one workstation to the next for processing without waiting for the entire job to be processed at the previous workstation. This means that successive operations of a job are allowed to be executed simultaneously when sublots are involved. The number of possible sublots in each job is directly related with the job size, i.e. the order quantity. To reflect the environment where customers demand unique or small quantities of products, the order quantity is randomly generated according to a discrete uniform distribution between one and four units. Thus, also the number of sublots for each job can range from one to four.

Table 1. Characteristics of the simulation model.

Shop configuration	General flow shop
No. of workstations	6
Workstation capacities	All equal and constant over time
Workstation utilisation	90%
Inter-arrival times	Exponentially distributed; mean = 0.647 time units
No. of operations per job	Discrete uniformly distributed [1, 6] operations
Job size (quantity)	Discrete uniformly distributed [1, 4] units
Unit operation processing times	2-Erlang, truncated at 4*mean; mean = 0.4 time unit
Set-up time	Sequence independent, included in the processing times

The inter-arrival times of jobs entering the production system follow an exponential distribution. The average inter-arrival time of jobs is chosen so that, combined with the routing and processing time characteristics, it leads to an average utilisation of 90%.

Directly upon job arrival, its due date, routing and operation processing times are defined. The due date d_j of each job j is modelled to be determined exogenously by adding a random allowance factor α to the job entry time t_j^E , as given in Equation (1)

$$d_j = t_j^E + \alpha \quad (1)$$

The random allowance factor α was set to vary uniformly between 25 and 55 time units after pretests. These values were found to lead to percentages of tardy jobs that give a good impression of the relative performance of the compared control policies. The number of operations per job is drawn from a discrete uniform distribution, with a minimum of one and a maximum of six. Unit processing times at each operation are drawn from a truncated 2-Erlang distribution with a mean of 0.4 time units and a maximum of four times the mean. These are then multiplied by the job size to obtain the processing time of each operation of the job.

The following common assumptions are adopted:

- Each operation of a job is processed in a different workstation.
- Machines are continuously available.
- Distances and transfer times between workstations are negligible.
- Set-up times are assumed to be sequence independent and included in the processing times of operations.
- Early shipment is not allowed.

3.2 Job release

As in previous simulation studies of WLC release methods (Land 2006; Frendendall, Divesh, and Patterson 2010; Thürer, Qu et al. 2014; Fernandes, Silva, and Carmo-Silva 2015), it is assumed that all jobs are accepted and materials are available. Jobs flow into a pre-shop pool and wait for release according to the Periodic with intermediate Pull Release (PPR) method (Hendry and Kingsman 1991; Thürer et al. 2012). In this method, jobs are considered for release based on their urgency, i.e. according to their planned release time (see Section 3.3), and are released only if the workload at none of the required workstations becomes larger than a predefined limit or workload norm as a result of their release. After fixed intervals of time, the workload at stations is computed and the release decision is taken. However, PPR combines periodic (push release) and continuous (pull release). A pull release may take place between periodic releases when the workload of a workstation falls to zero. In this case, only those jobs within the pre-shop pool, that have the first operation at the starving workstation, are considered for release. The job with the earliest planned release date is then released irrespective of whether it fits within the workload norms or not. The workloads at stations are calculated by the corrected aggregate load approach introduced by Oosterman, Land, and Gaalman (2000). Under this approach the load contribution of job to a workstation at the moment of release is depreciated according to the position of the workstation in the routing of a job. The further downstream a workstation is located, the higher the depreciation.

WLC requires a number of other parameters to be specified for job release in addition to workload norms, which will be an experimental factor (see 3.3). These are the planned throughput times for each workstation, the release period length and the time limit. The planned throughput times were fixed at five time units based on the observed throughput

times in preliminary simulation runs. As suggested in Land (2006), performance is not very sensitive to this parameter. The release period length was fixed at one time unit. This parameter defines the time interval between push release activations, i.e. the release frequency. The time limit is a time window that determines the set of jobs in the pre-shop pool that can be selected for release each time job release is activated. In the study, this parameter is set to infinity. This avoids needlessly retaining a job in the pool and minimises the average system throughput time (Land 2006).

3.3 Experimental design

The experimental factors and the levels at which they were evaluated in this study are summarised in Table 2. The lot splitting policy is tested at three levels, the dispatching strategy at two levels and workload norms at ten levels.

Three alternative policies regarding lot splitting are analysed in this study to determine whether lot splitting should be considered before or after release:

- Policy P0: The job is never split.
- Policy P1: The job is split before release and sublots are considered independently for release to the shop floor.
- Policy P2: The job is split on the shop floor, i.e. sublots of the same job are released together.

After release, priority-dispatching rules are used to control the progress of jobs (in case of P0) or sublots (in case of P1 and P2) on the shop floor. Whenever a workstation becomes available, a dispatching rule selects the job/sublot with the highest priority of the jobs/sublots currently waiting to be processed on that workstation. In order to ascertain how the dispatching strategy can be redesigned to take advantage of lot splitting, the following dispatching strategies were tested:

- Strategy S1: applies the *earliest Planned operation Starting Time* (PST) rule. PST is a due date oriented rule that acts by giving priority to the jobs/sublots that become most urgent at each workstation. It is a commonly used rule within WLC and is focused on reducing the variation of the lateness across job/sublots. The PST of a job j and each of its sublots at a workstation v is determined as follows:

$$PST_{jv} = d_j - \sum_{w \in S_{jv}} T_w \quad (2)$$

where T_w is the planned throughput time at workstation w , S_{jv} is the set of workstations in the remaining routing of j starting from workstation v and d_j is the due-date of j .

- Strategy S2: applies the *earliest Planned operation Starting Time/Zero work-in-next-queue* of next operation (PST/ZWINQ) rule. This is a new rule that selects a job or sublot for processing if its next operation is on a starving workstation, i.e. if the number of jobs in queue of the workstation is zero. If two or more jobs are in this situation, the job with the earliest PST is processed first. If no workstation is starving, jobs are processed following the PST rule.

This new rule is expected to take advantage of lot splitting by reducing the premature idleness of workstations, as discussed in Section 2, as it speeds up sublots for processing at starving workstations. At the same time, the rule tries to keep jobs on time by applying PST.

The final experimental factor is the workload norm, which is tested at 10 levels, varying from 3.8 to 14 and infinity. It is common practice in WLC simulation studies to define the workload norms as an experimental factor (e.g. Oosterman, Land, and Gaalman 2000; Land 2006; Thürer, Qu et al. 2014; Fernandes, Silva, and Carmo-Silva 2015) to determine at which range of workload levels the tested policies are able to function well.

Table 2. Experimental factors and levels.

Experimental factor	Levels			
Lot splitting policy	P0	P1		P2
Priority dispatching strategy	S1		S2	
Workload norm levels	3.8, 4.5, 5.3, 6.2, 7.3, 8.6, 10.1, 11.9, 14.0, ∞ time units			

4. Simulation results

4.1 Performance assessment and analysis

This section examines the performance results of the simulation study. During the simulation experiments, data was collected under steady state. The length of each replication was 13,000 time units, each with a warm-up period of 3,000 time units. The average values of 100 independent replications are presented as the result of each experiment. Common random numbers were used as a variance reduction technique. Whenever a performance difference between experiments is discussed, the significance of the difference could be shown by a paired t -test at a 95% confidence level.

Two types of criteria primarily measure system performance: the ability to provide shorter delivery times, and the ability to deliver jobs on time. Performance measures used with regard to the former are the total throughput time (TTT), the shop floor throughput time (STT), the pre-shop pool time and the synchronisation time. Performance measures with regard to the latter are the percentage of tardy jobs (P_{tardy}) and the standard deviation of the lateness (SD_l).

TTT is the time a subplot spends waiting in the pre-shop pool (before release) plus on the shop floor (after release). STT refers to the time that elapses between job release and job completion. Note that a job is not completed until all the sublots that belong to the job are fully processed. The synchronisation time is the time that elapses between the completion of the first and last subplot of a job.

An overview of the performance results of lot splitting policies under PST dispatching is presented in Figures 2 and 3. A logistic performance curve was developed for each policy. A marker on a curve is the result of simulating release at a specific workload norm level.

The percentage of tardy jobs, the total throughput time and the standard deviation of the lateness are plotted as a function of the average shop floor throughput time, Figures 2(a)–(c), respectively. The rightmost mark on each curve refers to an experiment with a workload norm of infinity, which means unrestricted release. Tighter workload norms result in shorter shop floor throughput times. Thus, the horizontal axis on each figure reflects the norm tightness level.

Figure 2. Performance results for lot release policies P0, P1 and P2.

Figure 3. Pool time of jobs for different release policies.

Results can be summarised as follows:

- From Figure 2 it can be observed that splitting under unrestrictive release has a positive effect on all performance measures studied, i.e. P_{tardy} , TTT, SD_I and STT. Lot splitting allows sublots to move independently and more quickly through the production process, shortening the STT and thus TTT and P_{tardy} .
- It can also be observed from Figure 2 that workload-based release significantly reduces P_{tardy} relative to unrestricted release, in spite of increasing the SD_I , especially under tight workload norms. This is explained by the increasing delay of jobs and sublots in the pre-shop pool with workload norms tightening, which disturbs the release of jobs according to their planned release times.
- Comparing the performance of lot splitting policies, it can be observed that splitting jobs on the shop floor, policy P2, outperforms the other two policies for the P_{tardy} criteria. Policy P1, splitting jobs at release, results in the highest P_{tardy} of all policies at each reduced level of STT, in spite of achieving slightly lower SD_I values than P2. In line with this behaviour of P1 is the marked increase of its synchronisation time.

Table 3 shows the results of 95% confidence level paired-samples t -tests of lot splitting policies P0, P1 and P2 to indicate the significance of discussed differences. To avoid the problem of multiple comparisons (see e.g. Law and Kelton 2000) only the t -tests for the percentage of tardy orders at a shop floor throughput time of 16 time units, around which releasing policies perform best, are presented.

Table 3. Paired-samples t -test comparisons of means of lot splitting policies P0, P1 and P2.

Identifier	Estimated mean difference	Standard deviation	0.950 C.I. Half-Width	Number of observations
P1 – P0	1.91	1.61	0.320	100
P2 – P0	–2.31	0.838	0.166	100
P2 – P1	–4.21	2.14	0.425	100

Figures 2(d) and 3 explain the reasons for the weaker P_{tardy} performance of P1. Splitting jobs in the pool, Policy P1, facilitates the release of large jobs to the shop floor when loose workload norms are applied, thus contributing to a good timing of release. However, as workload norms are tightened some of the job sublots are retained in the pool, thus negatively affecting: (1) the synchronisation times of the sublots after completion, for sending the job to the customer, as can be seen in Figure 2(d) and (2) the timing of release of sublots, as can be concluded from Figure 3. Figure 3 shows the pool time of jobs in relation to the job size (order quantity) for the three lot splitting policies. From this figure it can also be observed that Policy P1 has a negative impact on the pool time of small and medium jobs, not only under tight norms but also under loose workload norms.

Next the impact of the dispatching strategies is analysed. Figure 4 presents a logistic performance curve for each combination of dispatching strategies S1 and S2 with the lot splitting policies P1 and P2. The results show that, under unrestricted release, strategy S2 that is based on the new PST/ZWINQ rule allows improving all performance measures, namely P_{tardy} , despite the larger synchronisation time (Figure 4(d)) of sublots. This rule also allows improving P_{tardy} , TTT and the SD_I under controlled release and tight workload norms. This is independent of the lot splitting policy, i.e. of splitting jobs in the pool or in the shop floor. Note that under tight workload norms the probability of workstation starvation becomes higher. In this situation the ZWINQ component of this rule takes clear advantage of lot splitting to avoid the premature idleness of workstations by promoting the overlapping of consecutive operations.

Table 4 shows the results of a 95% confidence level paired-sample t -test of dispatching strategies S1 and S2 for the percentage of tardy orders under unrestricted release.

4.2 Sensitivity analysis

The performance of lot splitting policies P0, P1 and P2 is expected to be affected by the release period length, i.e. the time interval between push release activations in the PPR method. We have therefore conducted a sensitivity analysis to the value of this parameter in order to determine how it affects performance. The release period length, T , was tested at different values. In Figure 5 we present the simulation results for 1 and 4 time units, when dispatching is based on strategy S1.

Figure 4. Performance results for combined lot release policies P1 and P2 and dispatching strategies S1 and S2.

Table 4. Paired-samples *t*-test comparisons of means of dispatching strategies S1 and S2 under unrestrictive release.

Identifier	Estimated mean difference	Standard deviation	0.950 C.I. Half-Width	Number of observations
S2 – S1	–3.27	1.75	0.347	100

Figure 5. Performance results under different release periods: (a) $T = 1$ time units; (b) $T = 4$ time units.

From this Figure it can be observed that the performance of all release policies deteriorates for a larger release period. This is because the pool delay of jobs increases. However, the performance of policies P0 and P2 deteriorate more than the performance of P1 that splits jobs in the pool, i.e. before being released to the shop. Thus, performance differences between these policies become weaker under large release periods. This confirms that the release period length has influence on the relative performance of lot splitting policies. However, policy P1 showed to be more robust to changes in the release period.

5. Conclusion and managerial implications

This paper presents the results of a simulation study into dispatching strategies and lot splitting policies for controlled job release in the context of a MTO production system. The main implications and managerial guidelines of the study are as follows.

The study confirms results of earlier studies which do not apply lot splitting, in that workload-based order release outperforms unrestrictive release regarding the main performance criteria considered in the study, the percentage of tardy jobs. As such, workload-based release is recommended over unrestricted release.

When workload-based release is in place, a decision must be made concerning how sublots should be released to the shop floor. Better results are obtained by releasing the sublots that belong to the same job together, as opposed to releasing them independently. In fact, allowing sublots of a single job to be processed independently on the shop floor, after releasing them together, allowed for substantial performance improvement of the percentage of tardy jobs.

Results also suggest that it is important to design appropriate rules for the purpose of dispatching when lot splitting is applied under WLC. Applying the new PST/ZWINQ rule – which takes advantage of small sublots by quickly supplying work to starving workstations – improves the performance for both tested lot splitting policies. This dispatching rule performs particularly well under tight workload norms, i.e. when workload on the shop floor is highly restricted. This proves once more that avoiding premature idleness is important in the context of WLC. However, where earlier studies focussed on avoiding premature idleness during the order release decision (Land and Gaalman 1998; Thürer et al. 2014), this study shows how premature idleness can be avoided at the dispatching level by using lot splitting. Even limited lot splitting opportunities with up to four sublots are shown to enable a strong performance impact. This provides new directions for future research which might address other approaches that can further reduce premature idleness, using lot splitting or even going beyond the scope of lot splitting.

This study has obvious managerial implications because in industrial practice jobs are often split into sublots so operations at workstations can overlap and progress be accelerated. By using a simple generic model, this study has revealed useful insights for general practice. However, we should notice that every real shop would have its own distinctive features, which may require further investigation. An extension of this research would be to investigate other

shop configurations and implementation issues, e.g. set-up times. It would be important to check if the main conclusions of this work, in avoiding premature idleness of workstations, still hold when set-up saving rules are used. Another direction would be to analyse how the customer enquiry and order entry decisions in the WLC concept can take advantage of lot splitting. This might further improve our understanding of the practical application of WLC and, at the same time, enhance WLC theory.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the FCT-Fundação para a Ciência e Tecnologia of Portugal [project number PEst2015-2020: UID/CEC/00319/2013].

ORCID

Nuno O. Fernandes <http://orcid.org/0000-0002-4682-1790>

Martin J. Land <http://orcid.org/0000-0003-1704-9874>

S. Carmo-Silva <http://orcid.org/0000-0002-5645-3127>

References

- Azzi, A., F. Maurizio, and A. Persona. 2012. "Lot Splitting Scheduling Procedure for Makespan Reduction and Machine Capacity Increase in a Hybrid Flow Shop with Batch Production." *The International Journal of Advanced Manufacturing Technology* 59 (5): 775–786.
- Buscher, U., and L. Shen. 2009. "An Integrated Tabu Search Algorithm for the Lot Streaming Problem in Job Shops." *European Journal of Operational Research* 199 (2): 385–399.
- Chang, J. H., and H. N. Chiu. 2005. "A Comprehensive Review of Lot Streaming." *International Journal of Production Research* 43 (8): 1515–1536.
- Cheng, M., N. J. Mukherjee, and S. C. Sarin. 2013. "A Review of Lot Streaming." *International Journal of Production Research* 51 (23–24): 7023–7046.
- Enns, S. T. 1995. "An Integrated System for Controlling Shop Loading and Work Flow." *International Journal of Production Research* 33 (10): 2801–2820.
- Fernandes, N. O., M. J. Land, and S. Carmo-Silva. 2014. "Workload Control in Unbalanced Job Shops." *International Journal of Production Research* 52 (3): 679–690.
- Fernandes, N. O., C. Silva, and S. Carmo-Silva. 2015. "Order Release in the Hybrid MTO–FTO Production." *International Journal of Production Economics* 170: 513–520.
- Fredendall, L. D., O. Divesh, and J. W. Patterson. 2010. "Concerning the Theory of Workload Control." *European Journal of Operational Research* 201 (1): 99–111.
- Hendry, L. C., and B. G. Kingsman. 1991. "A Decision Support System for Job Release in Make-to-Order Companies." *International Journal of Operations and Production Management* 11: 6–16.
- Hendry, L., and S. K. Wong. 1994. "Alternative Order Release Mechanisms: A Comparison by Simulation." *International Journal of Production Research* 32 (12): 2827–2842. 16 p.
- Hendry, L., Y. Huang, and M. Stevenson. 2013. "Workload Control: Successful Implementation Taking a Contingency-based View of Production Planning & Control." *International Journal of Operations and Production Management* 33 (1): 69–103.
- Jacobs, F. R., and D. J. Bragg. 1988. "Repetitive Lots: Flow-time Reductions Through Sequencing and Dynamic Batch Sizing." *Decision Sciences* 19: 281–294.
- Jeong, H., J. Park, and R. C. Leachman. 1999. "A Batch Splitting Method for a Job Shop Scheduling Problem in an MRP Environment." *International Journal of Production Research* 37 (15): 3583–3598.
- Kanet, J. J. 1988. "Load-limited Order Release in Job Shop Scheduling Systems." *Journal of Operations Management* 7 (3): 44–58.
- Karmarkar, U. S., S. Kekre, S. Kekre, and S. Freeman. 1985. "Lot-sizing and Lead-time Performance in a Manufacturing Cell." *Interfaces* 15 (2): 1–9.
- Kim, J. S., S. H. Kang, and S. M. Lee. 1997. "Transfer Batch Scheduling for a Two-stage Flowshop with Identical Parallel Machines at Each Stage." *Omega* 25 (5): 547–555.
- Kropp, D. H., and T. L. Smunt. 1990. "Optimal and Heuristic Models for Lot Splitting in a Flow Shop." *Decision Sciences* 21 (4): 691–709.

- Land, M. J. 2006. "Parameters and Sensitivity in Workload Control." *International Journal of Production Economics* 104 (2): 625–638.
- Land, M. J., and G. J. C. Gaalman. 1996. "Workload Control Concepts in Job Shops: A Critical Assessment." *International Journal of Production Economics* 46–47 (1): 535–548.
- Land, M. J., and G. J. C. Gaalman. 1998. "The Performance of Workload Control Concepts in Job Shops: Improving the Release Method." *International Journal of Production Economics* 56–57 (1): 347–364.
- Law, A. M., and W. D. Kelton. 2000. *Simulation Modeling and Analysis*. 3rd ed. Singapore: McGraw-Hill.
- Li, S. L. 1997. "A Hybrid Two-stage Flowshop with Part Family, Batch Production, Major and Minor Set-ups." *European Journal of Operational Research* 102 (1): 142–156.
- Missbauer, H. 2002. "Lot Sizing in Workload Control Systems." *Production Planning and Control* 13 (7): 649–664.
- Oosterman, B., M. J. Land, and G. J. C. Gaalman. 2000. "The Influence of Shop Characteristics on Workload Control." *International Journal of Production Economics* 68 (1): 107–119.
- Portioli-Staudacher, A., and M. Tantardini. 2011. "A Lean-based ORR System for Non-repetitive Manufacturing." *International Journal of Production Research* 50 (12): 3257–3273.
- Ruben, R. A., and F. Mahmoodi. 1998. "Lot Splitting in Unbalanced Production Systems." *Decision Sciences* 29 (4): 921–949.
- Russell, G. R., and T. D. Fry. 1997. "Order Review/Release and Lot Splitting in Drum-buffer-rope." *International Journal of Production Research* 35 (3): 827–845.
- Smunt, T. L., A. H. Buss, and D. H. Kropp. 1996. "Lot Splitting in Stochastic Flow Shop and Job Shop Environments." *Decision Sciences* 27 (2): 215–238.
- Stevenson, M., L. C. Hendry, and B. G. Kingsman. 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.
- Stevenson, M., Y. Huang, L. C. Hendry, and G. D. Soepenberg. 2011. "The Theory and Practice of Workload Control: A Research Agenda and Implementation Strategy." *International Journal of Production Economics* 131: 689–700.
- Thürer, M., M. Stevenson, C. Silva, M. J. Land, and L. D. Fredendall. 2012. "Workload Control and Order Release: A Lean Solution for Make-to-order Companies." *Production and Operations Management* 21 (5): 939–953.
- Thürer, M., M. J. Land, M. Stevenson, L. D. Fredendall, and M. Godinho Filho. 2014. "Concerning Workload Control and Order Release: The Pre-shop Pool Sequencing Decision." *Production & Operations Management* 24 (7): 1179–1192.
- Thürer, M., T. Qu, M. Stevenson, T. Maschek, and M. Godinho Filho. 2014. "Continuous Workload Control Order Release Revisited: An Assessment by Simulation." *International Journal of Production Research* 52 (22): 6664–6680.
- Wanger, B. J., and G. L. Ragatz. 1994. "The Impact of Lot Splitting on Due Date Performance." *Journal of Operations Management* 12 (1): 13–25.
- Wiendahl, H.-P. 1995. *Load-oriented Manufacturing Control*. Berlin: Springer.