Operations Management Research (2020) 13:70–84
https://doi.org/10.1007/s12063-020-00149-w

State-dependent service rates in make-to-order shops: an assessment by simulation

Matthias Thürer[1] · Mark Stevenson[2] · James Aitken[3] · Cristovao Silva[4]

Received: 12 June 2019 / Revised: 21 November 2019 / Accepted: 9 January 2020 / Published online: 4 February 2020

Abstract

Most literature on make-to-order shops assumes that service rates are independent of the system state. In practice however, the service rate is often dependent on the workload level experienced by the worker. While a body of knowledge on state-dependent service rates exists, the available literature has not given sufficient attention to make-to-order shops, which are often characterized by complex routings and defined due dates, which means delivery performance becomes a major concern. This study uses simulation to assess the performance impact of state-dependent service rates under different degrees of routing directedness. We show that including information on the load upstream of a station when making service rate adjustments has the potential to improve performance compared to considering the load directly queuing at a station only, as has been the case in previous research on state-dependent service rates. Moreover, using the same threshold to trigger service rate adjustments at each station in shops with directed routings leads to higher service rates at upstream stations. This service rate imbalance can be avoided by using different triggering thresholds for upstream and downstream stations. Further, and most importantly, we show that although speeding up behavior during high load periods significantly improves performance, if worker fatigue leads to a decrease in the service rate in response to the initial increase then performance may in fact deteriorate.

Keywords Behavioral operations management · Capacity planning · Make-to-order production · Workload control · State-dependent service rates

1 Introduction

This study assesses the impact of state-dependent processing or service times on the performance of make-to-order shops through simulation. It originated from a case in practice where a company considered substituting human operators on its production line with robots. Ever since the origins of humanity, humans have strongly impacted economic systems – given that most economic systems are created by humans to meet human needs (Weber 2014; Roser 2016). Economic systems today typically consist of humans and machines, both of which are different. While they both exhibit variability between the planned and realized service rates, human workers inherently react to the state of the system that surrounds them (Bendoly et al. 2010). Considering this interaction is central for companies that need to properly balance the complex trade-offs existing in any industry (Samson and Kalchschmidt 2019). For example, human workers may speed up processing times when the workload in their queue increases and slow down if they are at risk of becoming idle (Schultz et al. 1998). This means that the service rate becomes dependent on the workload of the queue, shop, or system, i.e.

✉ Matthias Thürer
matthiasthurer@workloadcontrol.com

Mark Stevenson
m.stevenson@lancaster.ac.uk

James Aitken
james.aitken@surrey.ac.uk

Cristovao Silva
cristovao.silva@dem.uc.pt

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

[2] Department of Management Science, Lancaster University Management School, Lancaster University, Bailrigg LA1 4YX, UK

[3] Department of Business Transformation, Surrey Business School,University of Surrey, Stag Hill, Guildford GU2 7XH, UK

[4] Mechanical Engineering Department, CEMUC - University of Coimbra, Pólo II Pinhal de Marrocos, 3030 Coimbra, Portugal

state-dependent (Powell and Schultz 2004). Note that state-dependency is different from *uncertainty* in, for example, processing times, machine availability or material availability (e.g. Kenne and Gharbi 2001; Mason et al. 2005; Galbreth et al. 2012; Altendorfer et al. 2014) – such factors are state-independent. In fact, state-independent variation in operator processing times should be avoided, since it may lead to blocking and starvation in tightly coupled systems (Folgado et al. 2015).

While there has been significant research attention on state-dependent service rates (e.g. Schultz et al. 1998; Van Ooijen and Bertrand 2003; Delasay et al. 2016; Shunko et al. 2018), this body of work is typically in the context of single/parallel servers and production lines, characterized by restricted routing complexity and focused on the evaluation of efficiencies. The available literature does not, to the best of our knowledge, consider more complex make-to-order contexts where due dates exist and tardiness performance is the major performance criterion. Most make-to-order shops in practice produce low volume, high variety products (Childerhouse et al. 2002). This results in high variability in terms of job arrivals, routings, and processing times; and, as a consequence, the workload queuing at a station varies (Mincsovics and Dellaert 2009). While in practice some reaction to the changing workload would be expected (Delasay et al. 2016), the literature typically considers a station's service rate to be state-independent, i.e. not dependent on the workload.

The main exceptions to the above are Bertrand and van Ooijen (2002), Land et al. (2015), and Thürer et al. (2016). Bertrand and van Ooijen (2002) introduced the notion of an ideal workload level in the system to explain the discrepancy between theoretical results on the performance of release methods that stabilize the workload in the system and results from practice when these systems had been implemented, which were typically much better than anticipated. Bertrand and van Ooijen (2002) assumed that any deviation from the ideal workload level would result in a decrease in the service rate, and therefore that stabilizing the workload close to the ideal level would improve performance. The literature on state-dependent service rates provides another explanation for discrepancies between theory and practice: a speeding-up phenomenon whereby workers adjust their behavior in response to the workload level, which is typically not a consideration in theory. This effect has since been assessed by Land et al. (2015) who explored the impact of output control, i.e. an adjustment to the service rate that is triggered based on a station's workload, in make-to-order job shops. Land et al. (2015) showed that small but timely capacity adjustments significantly improve tardiness performance. Thürer et al. (2016) then showed that the positive performance effect is also maintained when order release is applied to stabilize and reduce the workload in the system. Land et al. (2015) and Thürer et al. (2016) extended Bertrand and van Ooijen's (2002) work by highlighting the impact of state-dependent service rates. This provided an important contribution to the literature; however, the authors did not consider the literature on state-dependent service rates. This omission has resulted in three shortcomings: first, the use of a different workload measure than the prior literature; second, failing to consider the impact of routing direction, with the focus instead being on a randomly routed job shop only; and, third, the omission of fatigue, i.e. a decrease in the service rate if adjustments are prolonged.

There exists a broad literature on the impact of human behavior on the performance of production systems, for example, in the context of lean manufacturing (Sugimori et al. 1977, van Assen 2018). This literature highlights that it is the unique capability of humans to react to the system around them that leads to significant performance improvements if exploited and aligned with company goals, e.g. by a goal management system that captures the way an organization shapes employees' decisions and actions through incentives and rewards (Galeazzo et al. 2017). This study contributes to the literature by advancing theory and practice concerning one aspect of human behavior in production systems: state-dependent service rates. It contributes to the existing literature on state-dependent service rates in three ways. First, by introducing different measures for the workload that triggers behavioral change. These measures provide different levels of visibility, which refers to the availability of feedback information on the length of the queue at a station on the shop floor (Shunko et al. 2018). Second, by exploring the performance impact of routing direction, moving from serial lines towards undirected job shops. This includes the impact of different parameter settings – instead of using equal parameters, as in the prior literature – for upstream and downstream stations, as suggested in the broader operations management literature (Gstettner and Kuhn 1996; Thürer et al. 2015). Third, by extending Land et al. (2015) to consider the impact of decreased service rates during high load periods. Simulation is chosen as an appropriate tool for exploring the behavioral dynamics inherent in systems (Bendoly et al. 2010), with the findings extending the understanding of state-dependent service rates and providing guidance to managers on how best to take advantage of this important phenomenon in practice.

The remainder of this paper is organized as follows. In Section 2, the literature is reviewed, and the research questions developed. The simulation model used is then described in Section 3 before the results are presented, discussed, and analyzed in Section 4. Finally, conclusions are drawn in Section 5, where limitations and future research directions are also outlined.

2 Literature review

This section briefly reviews the general literature on the operational impact of state-dependent service rates in Section 2.1 to identify which aspects to consider in the simulation. The limited available literature specifically on state-dependent service rates in make-to-order shops is then reviewed in Section 2.2, where the research questions are also formulated and outlined.

2.1 State-dependent service rates

Different approaches towards the state-dependency of service rates are discussed or modelled in the literature. For example, Van Ooijen and Bertrand (2003) assumed an ideal workload level can be found and that any deviation from this ideal level results in a service rate loss. They argued that only by stabilizing the workload – at the ideal level – could the maximum throughput be realized. Meanwhile, Delasay et al. (2016) developed a queueing model in which the service rate increases with the workload but decreases in periods of prolonged overload, so-called overwork periods. The authors showed that the commonly used fixed-server-speed Erlang C capacity-planning model may lead to errors in predicting performance or prescribing capacity levels. The occurrence of fatigue in response to prolonged periods of high load has been further explored by Öner-Közen et al. (2017), who argued that the positive impact of state-dependent service rates has been overestimated in previous studies due to simplifying assumptions, such as a disregard for state-dependent worker fatigue. Worker fatigue was modelled by Öner-Közen et al. (2017) as the diminishing ability of the worker to increase the service rate.

But even if this study were to exclusively focus on an increase in the service rate in response to an increased workload, different conceptualizations would still exist in the literature. For example, Hopp et al. (2007) introduced the idea of differences between non-discretionary task completion criteria (i.e. when and how an operation is completed is determined by objective standards) and discretionary task completion criteria (i.e. when and how an operation is completed is determined by a worker's subjective standards). Hopp et al. (2007) then went on to argue that discretionary tasks allow for speeding up operations, a so-called quality buffer since service quality arguably suffers. In contrast, in this study, the focus is on the capacity buffer inherent within each worker. This refers to a worker's capacity to increase the service rate by working in the short-term at an above average speed. Similarly, Batt and Terwiesch (2017) summarized three load-adapting mechanisms from the literature – rushing, task reduction, and multitasking – and explored the impact of a fourth, early task initiation.

In this study, the focus is on simple speed-up effects, as observed, for example, in Schultz et al. (1998). Schultz et al. (1998) showed that, in systems with state-dependent service rates, there is less idle time and a higher output than would be predicted using assumptions of independence. This may, for example, explain why low-inventory systems do not exhibit the predicted productivity losses expected when applying operations research models (Schultz et al. 1999) and may challenge implicit assumptions on the negative effect of line length on throughput (Powell and Schultz 2004).

2.2 State-dependent service rates in make-to-order shops

Only three studies could be identified from the literature on state-dependent service rates in a make-to-order context: Bertrand and van Ooijen (2002), Land et al. (2015), and Thürer et al. (2016). Bertrand and van Ooijen (2002) introduced an ideal workload level and assumed that any deviation from this ideal level would reduce the service rate. In contrast, this study focusses on speeding up behavior and fatigue as in Delasay et al. (2016). The speeding up behavior is similar to Land et al. (2015), who introduced a new capacity adjustment mechanism that triggers adjustments based on a station's workload. Adjustments were implemented by reducing the realized processing times. In other words, Land et al. (2015) considered processing times to be state-dependent. Using a stylized job shop model, the authors then demonstrated that small capacity adjustments targeted at handling high load periods can improve the percentage tardy and other delivery-related performance measures. This finding was later confirmed by Thürer et al. (2016) in a job shop with order release control.

Land et al. (2015) in particular is important in two ways. First, it assessed the impact of a state-dependent service rate in a make-to-order shop. Second, it introduced the use of the corrected aggregate load as a new measure for triggering behavior change. The corrected aggregate load approach was first introduced by Oosterman et al. (2000). It measures the sum of all of the work for a given station that is on the shop floor but not yet completed, whereby the workload contribution is corrected by the position of the station in the routing of a job. In other words, the corrected aggregate load contribution of a job to the i^{th} station in its routing is determined by $\frac{p_{ij}}{i}$. A job contributes to the load of a station upon its entry to the shop and is excluded as soon as the operation at that particular station is complete. Dividing by the station position recognizes that a job's contribution to the direct load of a station is limited to the portion of the time that the job is actually queuing at the station. This measure gives the best representation of the future expected direct load of a station based on the mix

of routings actually present on the shop floor (Oosterman et al. 2000).

In contrast to Land et al. (2015), studies on state-dependent service rates trigger a change in worker behavior based on the load actually queuing at a station (e.g. Powell and Schultz 2004; Öner-Közen et al. 2017) – the so-called direct load. This neglects the load upstream of the station – the so-called indirect load – which is reflected in the corrected aggregate load. Shunko et al. (2018) recently explored the impact of the queue structure (single or pooled queues vs. dedicated queues for multiple servers), and queue-length visibility (full or blocked visibility) on performance. They found that a single queue structure and poor visibility of the queue length slows down the servers. However, in Shunko et al.'s (2018) study there is no routing, i.e. the workload at a station does not depend on job progress at an upstream station. In contrast, this study considers make-to-order shops with complex routings. The first research question therefore asks:

RQ1: Should the service rate be dependent on global information on the corrected aggregate load or only based on local information on the direct load queueing at a station?

Land et al. (2015) made a major contribution by assessing state-dependent service rates in job shops. However, the authors did not assess performance in more directed routing settings where typical upstream and downstream stations exist. Previous research from the wider operations management literature has highlighted that upstream and downstream stations are characterized by different load patterns that may influence performance and, consequently, require different parameter settings (Gstettner and Kuhn 1996; Thürer et al. 2015). It is therefore important to extend Land et al. (2015) to consider the impact of routing directedness on their results. The second research question therefore asks:

RQ2: How does routing direction influence performance in shops with state-dependent service rates?

Finally, Land et al. (2015) assumed that workers are capable of increasing the service rate for an infinite period of time. In practice, however, worker fatigue is likely to occur. This means there is a decrease in the service rate if high load periods are prolonged. The third and final research question therefore asks:

RQ3: What is the performance impact of worker fatigue that results from a prolonged increase of the service rate?

Controlled simulation experiments will next be used to answer above three research questions.

3 Simulation model

In this study a high variety make-to-order environment is considered. A powerful tool for analyzing this kind of complex, stochastic system is discrete event simulation, which has also been widely applied in previous behavioral research (Powell and Schultz 2004; Neumann and Medbo 2009; Dode et al. 2016; Öner-Közen et al. 2017). The model characteristics are first described in Section 3.1. How state-dependency of the service rate is modelled is then described in Section 3.2. Finally, the dispatching rule used to control the progress of orders on the shop floor is described in Section 3.3 before the experimental setting and the performance measures are summarized in Section 3.4.

3.1 Model characteristics

Three different shop models, representing different degrees of routing directedness, will be used. To improve the generalizability of the findings and to avoid interactions that might inhibit a full understanding of the effects of the experimental factors, a stylized model of a pure job shop, a general flow shop, and a pure flow shop is used. All three shop models have been implemented in the Python© programming language using the SimPy© simulation module. All three shops contain six stations, where each station is a single resource with constant capacity. For the *pure job shop*, the routing is undirected and the routing length of jobs varies uniformly from one to six operations. The routing length is first determined before the routing sequence is generated randomly without replacement. For the *general flow shop*, the resulting routing vector (i.e. the sequence in which stations are visited) is sorted such that the routing becomes directed and there are typical upstream and downstream stations. For the *pure flow shop*, all jobs visit all stations in increasing station number order. Operation processing times – before adjustment – follow a truncated 2-Erlang distribution with a mean of 1 time unit after truncation and a maximum of 4 time units. The inter-arrival time of jobs follows an exponential distribution with a mean of 0.648 time units for the pure job shop and general flow shop and a mean of 1.111 time units for the pure flow shop. Both settings deliberately result in a utilization level of 90% without adjustments. Due dates are set exogenously by adding a random allowance to the job entry time. This allowance is uniformly distributed between 28 and 36 time units for the pure job shop and general flow shop and between 40 and 55 time units for the pure flow shop.

Finally, Table 1 summarizes the shop and job characteristics. While in practice any individual high-variety shop will certainly differ from these stylized models, these models capture the high routing variability, processing time variability, and arrival variability that defines this context.

3.2 State-dependent service rate

As in previous simulation research on state-dependent service rates, the processing times are adjusted in response to a station's workload (Powell and Schultz 2004; Delasay et al. 2016; Öner-Közen et al. 2017). In other words, when a certain

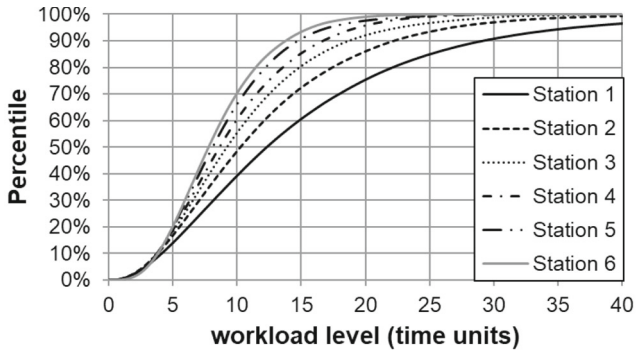
Table 1 Summary of the simulated shop and job characteristics

		Pure job shop	General flow shop	Pure flow shop
Shop characteristics	Routing Variability	Random routing	Random routing	Fixed sequence
	Routing Direction	Undirected routing	Directed routing	Directed routing
	No. of Stations	6	6	6
	Station Interchangeability	No interchange-ability	No interchange-ability	No interchange-ability
	Station Capacities	All equal	All equal	All equal
Job characteristics	No. of Operations	Discrete Uniform [1, 6]	Discrete Uniform [1, 6]	6
	Processing Times	Truncated 2–Erlang; (mean = 1; max = 4)	Truncated 2–Erlang; (mean = 1; max = 4)	Truncated 2–Erlang (mean = 1; max = 4)
	Due Date (DD)	DD = Entry Time + d; d U ~ [28, 36]	DD = Entry Time + d; d U ~ [28, 36]	DD = Entry Time + d; d U ~ [40, 55]
	Inter-Arrival Times	Exp. Distribution; mean = 0.648	Exp. Distribution; mean = 0.648	Exp. Distribution; mean = 1.111

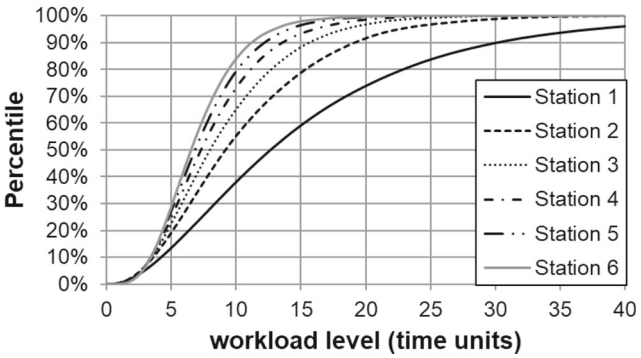
load threshold is violated, capacity adjustments are triggered. This models the worker's response to an increased workload. This study does not consider an increase in processing times in response to less work since this inclusion only led to a negligible effect in Powell and Schultz (2004). To avoid system nervousness, a second threshold is used to determine when to cease the capacity adjustment. Both thresholds – the commencing threshold and the stopping/ceasing threshold – are based on the cumulative frequency distribution of the workload obtained in preliminary simulations. As an example, the distributions for the General Flow Shop and the Pure Flow Shop are given in Fig. 1. The starting threshold is set to the 90th percentile and the stopping threshold is set to the 85th percentile; the use of only one level for these parameters is justified by the performance frontier observed in Land et al. (2015). As can be seen from Fig. 1, if the routing is directed then every station has a different distribution. Consequently, two scenarios are considered:

i. An equal load threshold for all stations based on the average shop load; and,
ii. A different load threshold for every station.

The strength of the positive worker response to an increased workload level is modelled by the adjustment size α. Schultz et al. (1998) observed an adjustment of approximately 20%. The same adjustment was also used in the simulations executed by Powell and Schultz (2004). Four different levels of the adjustment size are therefore considered to model the increase in the service rate: 0 (i.e. no capacity adjustment), 10, 20, and a 30% adjustment. As in Delasay et al. (2016) fatigue is modelled by decreasing the service rate stepwise after each job completion during high load periods until a certain threshold is reached. This means that the positive effect of a high workload is gradually reduced and transformed into a decrease in the service rate if it is sustained for too long. As in previous research, this overwork or 'fatigue' effect is modelled as increasing linearly over time since there is no other empirical data available (Jaber and Neumann 2010). Meanwhile, Delasay et al. (2016) showed that using the job count leads to similar performance as measuring the duration of the high load period. The final threshold was set to 110% of the processing time. It results in less than 100% utilization (99%) to avoid unstable simulations. Meanwhile, eight levels of the fatigue factor β are considered using a 2 logarithmic scale: 0, 0.00781, 0.00391, 0.00195, 0.00098, 0.00049, 0.00024, and 0.00012. This factor multiplied by the number of job completions during the current high load period gives



(a) The GFS



(b) The PFS

Fig. 1 Frequency distibution of the workload in: (**a**) the General Flow Shop (GFS); and, (**b**) the Pure Flow Shop (PFS)

the adjustment to the processing times (already adjusted by alpha).

3.3 Dispatching rule

Two dispatching rules will be considered: the Operation Due Date (ODD) rule and the Modified Operation Due Date (MODD) rule (see, e.g. Baker and Kanet 1983). The ODD rule prioritizes jobs with the earliest operation due date, where the calculation for the operation due date δ_{ij} of the i^{th} operation of a job j follows Eq. (1) below. The operation due date for the last operation in the routing of a job is equal to the due date δ_j, while the operation due date of each preceding operation is determined by successively subtracting an allowance c from the operation due date of the next operation. The allowance was set to 3 time units since this value resulted in the best overall performance in preliminary simulation experiments. Note that this study uses a constant allowance and not a dynamic allowance that is dependent on the workload. If the workload level increases then the operation throughput time for all jobs also increases. Thus, the priority ordering of jobs would still be the same as for a constant allowance.

$$\delta_{ij} = \delta_j - (n_j - i) \cdot c \quad i: 1 \ldots n_j \quad (1)$$

n_j – number of operations in the routing o job j

The MODD rule prioritizes jobs according to the lowest priority number, which is given by the maximum of the operation due date and the earliest finish time. In other words, $max(\delta_{ij},\ t + p_{ij})$ for an operation with processing time p_{ij}, where t refers to the time when the dispatching decision is made. The MODD rule shifts between a focus on ODDs to complete jobs on time and a focus on speeding up jobs – through shortest processing times – during periods of high load, i.e. when multiple jobs exceed their ODD (Land et al. 2015). Note that a dynamic allowance would prohibit the use of the same ODD's for ODD and MODD dispatching, since this switching behavior, which characterizes MODD, would not occur.

3.4 Experimental design and performance measures

The experimental factors are summarized in Table 2. A full factorial design was used for each shop type. Each scenario was replicated 100 times, while results were collected over 10,000 time units following a warm-up period of 3000 time units. These parameters allowed us to obtain stable results while keeping the simulation run time to a reasonable level.

The principal performance measures considered in this study are as follows: the *lead time* – the mean of the completion date minus the entry date across jobs; the *percentage tardy* – the percentage of jobs completed after the due date; and, the *mean tardiness* $T_j = \max(0, L_j)$,with L_j being the lateness of job j (i.e. the actual delivery date minus the due date of job j).

4 Results

To obtain a first indication of the relative impact of the experimental factors, statistical analysis has been conducted by applying an Analysis of Variance (ANOVA). ANOVA is here based on a block design, which is typically used to account for known sources of variation in an experiment. In the ANOVA, the dispatching rule is treated as the blocking factor. This allows the main effect of this factor and the main and interaction effects of the capacity related factors – the triggering parameter/threshold, the load measure, the adjustment size, and the fatigue factor – to be captured. The results are presented in Tables 3, 4, and 5 for the pure job shop, the general flow shop, and the pure flow shop, respectively. All main and interaction effects were shown to be statistically significant (0.05) in the pure job shop (Table 3). Most main effects in the general flow shop (Table 4) and the pure flow shop (Table 5) were also shown to be statistically significant. The main exception is the load measure. Meanwhile, significant two-way and three-way interactions exist for both shop types, whilst four-way interactions are only significant for the pure flow shop (Table 5).

The Scheffé multiple comparison procedure was applied to obtain a first indication of the direction and size of the performance differences across the dispatching rule, the triggering parameter/threshold, and the load measure. Since the level of the adjustment size and the fatigue factor represent a continuum, a multiple comparison procedure is arguably not meaningful for these factors. Table 6 gives the 95% confidence interval. If this interval includes zero, performance differences are not considered to be statistically significant. We can observe significant performance differences for most pairs for at least one performance measure, with the main exception being the triggering parameter/threshold in the general flow shop. To further explore these differences, detailed performance results will be presented next. Section 4.1 focusses on the pure job shop and an increase in the service rate while the impact of a more directed routing is assessed in Section 4.2 (i.e. the general and pure flow shops). Section 4.3 then assesses the impact of worker fatigue before a summary discussion is provided in Section 4.4.

4.1 Assessment of results: The pure job shop and an increase in the service rate

To answer the first research question – should the service rate be dependent on global information on the corrected aggregate load or only based on local information on the direct load queueing at a station? – the focus is first on the pure job shop and an increase in the service rate. The results for the pure job shop, i.e. with undirected random routings, are given in Table 7 together with the average utilization and the total number of adjustments per 1000

Table 2 Summary of the experimental design

	Pure job shop	General flow shop	Pure flow shop
Load measure	Direct and corrected load		
Trigger parameter	All equal and different across stations		All equal
Adjustment size	0 (no adjustment), 10%, 20% and 30%		
Fatigue factor	0.00781,0.00391,0.00195,0.00098,0.00049,0.00024,and 0.00012		
Dispatching rule	Operation Due Date (ODD) and Modified Operation Due Date (MODD)		

time units. The results confirm the positive performance impact of small but timely service rate adjustments, as reported in Land et al. (2015) for the pure job shop. If the results for the direct load and corrected load measures are compared for similar average utilization levels, then a better percentage tardy performance for the direct load can be observed if ODD dispatching is applied. However, the corrected load leads to a lower mean tardiness, whilst the lead time is similar for both load measures. The performance difference disappears under MODD dispatching, and the direct and corrected load measures perform similarly. As expected from previous literature considering both dispatching rules (e.g., Land et al. 2015), MODD improves performance compared to ODD dispatching. In general, the corrected load measure leads to fewer adjustments since it is less granular than the direct load. For example, if an operation of 1 time unit is completed at the last station in the routing of a job, the load measure is reduced by 1 time unit for the direct load but by only 1/6 time units for the corrected load approach. The latter is less

Table 3 ANOVA results for the pure job shop

Performance measure	Source of variance	Sum of squares	Degree of freedom	Mean squares	F-Ratio	*p* Value
Lead time	Dispatching	1,666,958.30	1	1,666,958.30	869.78	0.00
	Measure (M)	1,086,841.10	1	1,086,841.10	567.09	0.00
	Adjustment Size (A)	3,575,036.90	3	1,191,679.00	621.79	0.00
	Fatigue Factor (B)	3,475,095.90	7	496,442.28	259.03	0.00
	M x A	602,322.71	3	200,774.24	104.76	0.00
	M x B	846,673.21	7	120,953.32	63.11	0.00
	A x B	1,936,730.00	21	92,225.24	48.12	0.00
	M x A x B	498,212.88	21	23,724.42	12.38	0.00
	Error	24,407,032.00	12,735	1916.53		
Percentage tardy	Dispatching	80.82	1	80.82	3350.33	0.00
	Measure (M)	14.86	1	14.86	615.96	0.00
	Adjustment Size (A)	157.84	3	52.61	2181.01	0.00
	Fatigue Factor (B)	111.93	7	15.99	662.84	0.00
	M x A	4.02	3	1.34	55.59	0.00
	M x B	9.46	7	1.35	56.01	0.00
	A x B	55.35	21	2.64	109.26	0.00
	M x A x B	4.54	21	0.22	8.96	0.00
	Error	307.21	12,735	0.02		
Mean tardiness	Dispatching	1,543,230.40	1	1,543,230.40	867.37	0.00
	Measure (M)	974,389.70	1	974,389.70	547.66	0.00
	Adjustment Size (A)	2,877,026.90	3	959,008.97	539.01	0.00
	Fatigue Factor (B)	2,894,376.00	7	413,482.29	232.40	0.00
	M x A	566,463.27	3	188,821.09	106.13	0.00
	M x B	772,731.73	7	110,390.25	62.04	0.00
	A x B	1,631,818.60	21	77,705.65	43.67	0.00
	M x A x B	459,837.04	21	21,897.00	12.31	0.00
	Error	22,658,146.00	12,735	1779.20		

Table 4 ANOVA results for the general flow shop

Performance measure	Source of variance	Sum of squares	Degree of freedom	Mean squares	F-Ratio	p Value
Lead time	Dispatching	3,624,286.80	1	3,624,286.80	1993.10	0.00
	Parameter (P)	992,311.97	1	992,311.97	545.70	0.00
	Measure (M)	285.43	1	285.43	0.16	0.69
	Adjustment Size (A)	6,482,834.40	3	2,160,944.80	1188.37	0.00
	Fatigue Factor (B)	6,638,341.20	7	948,334.45	521.52	0.00
	M x P	31,881.96	1	31,881.96	17.53	0.00
	M x A	221,252.05	3	73,750.68	40.56	0.00
	M x B	611,716.27	7	87,388.04	48.06	0.00
	P x A	1644.08	3	548.03	0.30	0.82
	P x B	1146.57	7	163.80	0.09	1.00
	A x B	2,701,169.80	21	128,627.14	70.74	0.00
	M x P x A	23,567.40	3	7855.80	4.32	0.00
	M x P x B	24,102.58	7	3443.23	1.89	0.07
	M x A x B	212,316.96	21	10,110.33	5.56	0.00
	P x A x B	9441.69	21	449.60	0.25	1.00
	M x P x A x B	17,042.79	21	811.56	0.45	0.99
	Error	46,316,835.00	25,471	1818.41		
Percentage tardy	Dispatching	177.38	1	177.38	8288.97	0.00
	Parameter (P)	13.49	1	13.49	630.19	0.00
	Measure (M)	0.09	1	0.09	4.24	0.04
	Adjustment Size (A)	284.01	3	94.67	4423.87	0.00
	Fatigue Factor (B)	215.47	7	30.78	1438.44	0.00
	M x P	0.30	1	0.30	14.03	0.00
	M x A	0.21	3	0.07	3.32	0.02
	M x B	2.36	7	0.34	15.76	0.00
	P x A	0.28	3	0.09	4.42	0.00
	P x B	0.07	7	0.01	0.46	0.86
	A x B	81.75	21	3.89	181.92	0.00
	M x P x A	0.21	3	0.07	3.28	0.02
	M x P x B	0.15	7	0.02	0.98	0.44
	M x A x B	0.56	21	0.03	1.24	0.21
	P x A x B	0.18	21	0.01	0.41	0.99
	M x P x A x B	0.15	21	0.01	0.34	1.00
	Error	545.07	25,471	0.02		
Mean tardiness	Dispatching	3,395,236.30	1	3,395,236.30	2003.41	0.00
	Parameter (P)	886,295.03	1	886,295.03	522.97	0.00
	Measure (M)	110.67	1	110.67	0.07	0.80
	Adjustment Size (A)	5,191,507.20	3	1,730,502.40	1021.11	0.00
	Fatigue Factor (B)	5,512,697.10	7	787,528.16	464.69	0.00
	M x P	29,937.13	1	29,937.13	17.66	0.00
	M x A	232,874.09	3	77,624.70	45.80	0.00
	M x B	584,317.11	7	83,473.87	49.25	0.00
	P x A	1277.91	3	425.97	0.25	0.86
	P x B	1017.38	7	145.34	0.09	1.00
	A x B	2,260,130.80	21	107,625.27	63.51	0.00
	M x P x A	20,571.41	3	6857.14	4.05	0.01
	M x P x B	21,905.44	7	3129.35	1.85	0.07
	M x A x B	206,699.16	21	9842.82	5.81	0.00
	P x A x B	8488.26	21	404.20	0.24	1.00
	M x P x A x B	15,621.50	21	743.88	0.44	0.99
	Error	43,166,464.00	25,471	1694.73		

likely to directly trigger the stopping threshold, which means that the service rate increase continues for a longer period.

4.2 Assessment of results: The general flow shop and pure flow shop

To answer the second research question – how does routing direction influence performance in shops with state-dependent service rates? – attention now turns to the results in the general flow shop and pure flow shop with an increase in the service rate. The results for the general flow shop in Table 8, i.e. with directed random routings, largely confirm the findings from the pure job shop. However, using different parameters for every station has a positive performance impact if the corrected load measure is applied and a negligible effect if the direct load is used to trigger behavior change. As a result,

Table 5 ANOVA results for the pure flow shop

Performance measure	Source of variance	Sum of squares	Degree of freedom	Mean squares	F-Ratio	p Value
Lead time	Dispatching	16,149,185.00	1	16,149,185.00	1967.11	0.00
	Parameter (P)	10,845,096.00	1	10,845,096.00	1321.02	0.00
	Measure (M)	212,609.87	1	212,609.87	25.90	0.00
	Adjustment Size (A)	29,757,470.00	3	9,919,156.70	1208.24	0.00
	Fatigue Factor (B)	32,634,145.00	7	4,662,020.70	567.87	0.00
	M x P	259,595.40	1	259,595.40	31.62	0.00
	M x A	3,362,653.00	3	1,120,884.30	136.53	0.00
	M x B	5,282,553.10	7	754,650.45	91.92	0.00
	P x A	19,805.92	3	6601.97	0.80	0.49
	P x B	95,758.12	7	13,679.73	1.67	0.11
	A x B	11,416,293.00	21	543,632.99	66.22	0.00
	M x P x A	129,324.68	3	43,108.23	5.25	0.00
	M x P x B	133,924.32	7	19,132.05	2.33	0.02
	M x A x B	2,753,322.00	21	131,110.57	15.97	0.00
	P x A x B	277,392.47	21	13,209.17	1.61	0.04
	M x P x A x B	294,009.27	21	14,000.44	1.71	0.02
	Error	209,100,000.00	25,471	8209.61		
Percentage tardy	Dispatching	194.89	1	194.89	5502.17	0.00
	Parameter (P)	59.35	1	59.35	1675.64	0.00
	Measure (M)	0.00	1	0.00	0.06	0.81
	Adjustment Size (A)	530.04	3	176.68	4988.02	0.00
	Fatigue Factor (B)	490.06	7	70.01	1976.47	0.00
	M x P	0.02	1	0.02	0.46	0.50
	M x A	0.87	3	0.29	8.19	0.00
	M x B	16.01	7	2.29	64.58	0.00
	P x A	4.84	3	1.61	45.53	0.00
	P x B	3.53	7	0.50	14.25	0.00
	A x B	158.94	21	7.57	213.68	0.00
	M x P x A	0.32	3	0.11	2.98	0.03
	M x P x B	0.35	7	0.05	1.42	0.19
	M x A x B	15.91	21	0.76	21.39	0.00
	P x A x B	1.29	21	0.06	1.73	0.02
	M x P x A x B	2.16	21	0.10	2.90	0.00
	Error	902.21	25,471	0.04		
Mean tardiness	Dispatching	15,389,884.00	1	15,389,884.00	1963.15	0.00
	Parameter (P)	9,795,400.70	1	9,795,400.70	1249.51	0.00
	Measure (M)	209,909.77	1	209,909.77	26.78	0.00
	Adjustment Size (A)	24,273,869.00	3	8,091,289.70	1032.13	0.00
	Fatigue Factor (B)	27,241,955.00	7	3,891,707.90	496.43	0.00
	M x P	282,171.77	1	282,171.77	35.99	0.00
	M x A	3,463,246.70	3	1,154,415.60	147.26	0.00
	M x B	5,079,155.80	7	725,593.69	92.56	0.00
	P x A	22,644.44	3	7548.15	0.96	0.41
	P x B	107,333.63	7	15,333.38	1.96	0.06
	A x B	9,737,283.90	21	463,680.18	59.15	0.00
	M x P x A	114,040.75	3	38,013.58	4.85	0.00
	M x P x B	138,971.96	7	19,853.14	2.53	0.01
	M x A x B	2,640,535.20	21	125,739.77	16.04	0.00
	P x A x B	282,152.79	21	13,435.85	1.71	0.02
	M x P x A x B	291,606.08	21	13,886.00	1.77	0.02
	Error	199,700,000.00	25,471	7839.38		

the corrected load measure performs better than the direct load measure when different parameters are applied. In general, the impact of the granularity of the workload is stronger, which leads to a stronger reduction in the average utilization for the direct load when compared to the corrected load measure. This effect is even more amplified in the pure flow shop, i.e. with directed routings in a fixed sequence, as can be seen from Table 9. It prohibits a fair comparison for the direct vs. corrected load measures under ODD dispatching since the average utilization is lower for the direct load for all adjustment levels considered. A comparison is however possible for the best-performing setting, which is under MODD dispatching and with the use of different parameters. For this setting, the corrected load measure leads to a lower mean tardiness whilst maintaining similar percentage tardy and lead time results. Overall, these results extend the findings of Land et al. (2015) on the positive effect of capacity

Table 6 Results for scheffé multiple comparison procedure

	Experimental factor	Rule (x)	Rule (y)	PJS		GFS		PFS	
				lower[1)]	upper	lower	upper	lower	upper
Lead time	Dispatching	MODD	ODD	−24.3	−21.3	−24.8	−22.8	−52.5	−48.0
	Measure	Corrected	Direct	16.9	19.9	11.4	13.5	38.9	43.4
	Parameter	Different	Equal	Not applicable		−1.3*	0.8	−8.0	−3.5
Percentage tardy	Dispatching	MODD	ODD	−0.164	−0.154	−0.170	−0.163	−0.179	−0.170
	Measure	Corrected	Direct	0.063	0.074	0.042	0.049	0.092	0.101
	Parameter	Different	Equal	Not applicable		−0.007*	0.000	−0.005*	0.004
Mean tardiness	Dispatching	MODD	ODD	−23.4	−20.5	−24.0	−22.0	−51.2	−46.9
	Measure	Corrected	Direct	16.0	18.9	10.8	12.8	37.0	41.3
	Parameter	Different	Equal	Not applicable		−1.1*	0.9	−7.9	−3.6

[1] 95% confidence interval; * not significant at $\alpha = 0.05$

adjustments to shops where the routing direction is more directed. However, threshold levels need to take the routing position of a station into account.

4.3 Assessment of results: The impact of worker fatigue

To answer the third research question – what is the performance impact of worker fatigue that results from a prolonged increase of the service rate? – Fig. 2 depicts the results obtained for different levels of the fatigue factor in the general flow shop. Only one shop type and one level of dispatching (MODD), the triggering parameter/threshold (different parameters) and the load measure (corrected workload) are presented since the performance impact was qualitatively similar across these factors.

The following can be observed from Fig. 2:

- *Impact of fatigue in isolation*: This can be observed from the curve obtained for an adjustment factor of zero. There is a direct detrimental effect even if the fatigue factor is relatively small. Note that the decrease in service rate that occurs is limited to avoid unstable simulations. Therefore, the percentage tardy remains below 50%. The maximum utilization is 92.4% as can be observed from Table 10, which gives the utilization corresponding to the experimental results in Fig. 2. To assess the impact of fatigue specifically in high load periods, data for an average utilization rate of 92.4% was also collected, increasing the average processing time by 2.66%. Compared to these values (lead time = 26.8 time units; percentage tardy = 23.5%; and mean tardiness = 2.48 time units) it can be

Table 7 Performance Results for the Pure Job Shop

Dispatching rule	Load measure	Adjust. size	Lead time	Percentage tardy	Mean Tardiness	Average utilization	Number of adjust.[a]
ODD	Direct	0	22.90	20.33%	1.81	90.02%	8.6
		10%	21.10	14.10%	0.75	89.82%	8.6
		20%	20.58	11.70%	0.52	89.74%	8.5
		30%	20.32	10.47%	0.43	89.70%	8.4
	Corrected	10%	21.34	15.64%	0.78	89.84%	4.5
		20%	20.87	13.53%	0.54	89.77%	4.3
		30%	20.63	12.16%	0.44	89.73%	4.2
MODD	Direct	0	21.62	9.67%	0.84	90.02%	10.5
		10%	20.45	6.75%	0.37	89.79%	10.5
		20%	20.05	5.67%	0.26	89.70%	10.4
		30%	19.85	5.09%	0.21	89.66%	10.1
	Corrected	10%	20.69	7.44%	0.41	89.83%	5.1
		20%	20.35	6.40%	0.28	89.75%	4.9
		30%	20.17	5.77%	0.23	89.70%	4.8

[a] Number of Adjustments per 1000 time units

Table 8 Performance results for the general flow shop

Dispatching rule	Trigger Parameter	Load Measure	Adjust. size	Lead time	Percentage tardy	Mean tardiness	Average utilization	Number of Adjust.[a]
ODD	Equal	Direct	0	23.73	23.87%	2.40	90.02%	7.8
			10%	21.64	17.18%	1.16	89.77%	8.2
			20%	21.02	14.80%	0.88	89.67%	8.4
			30%	20.73	13.67%	0.76	89.62%	8.5
		Corrected	10%	22.44	20.24%	1.59	89.87%	3.3
			20%	22.10	19.07%	1.42	89.82%	3.2
			30%	21.94	18.49%	1.34	89.79%	3.2
	Different	Direct	10%	21.45	16.54%	1.08	89.73%	13.7
			20%	20.77	13.91%	0.77	89.62%	13.8
			30%	20.45	12.60%	0.66	89.56%	13.7
		Corrected	10%	22.12	19.99%	1.27	89.85%	3.7
			20%	21.66	18.28%	0.99	89.79%	3.5
			30%	21.44	17.30%	0.87	89.76%	3.4
MODD	Equal	Direct	0	21.94	12.04%	0.96	90.02%	8.5
			10%	20.67	8.67%	0.48	89.74%	9.2
			20%	20.23	7.44%	0.36	89.64%	9.5
			30%	20.01	6.84%	0.31	89.59%	9.5
		Corrected	10%	21.22	10.36%	0.66	89.86%	3.6
			20%	20.98	9.77%	0.58	89.80%	3.6
			30%	20.86	9.49%	0.55	89.76%	3.6
	Different	Direct	10%	20.44	7.97%	0.43	89.68%	17.4
			20%	19.92	6.57%	0.30	89.56%	17.7
			30%	19.66	5.93%	0.26	89.49%	17.7
		Corrected	10%	21.01	9.67%	0.51	89.82%	4.8
			20%	20.67	8.62%	0.38	89.74%	4.5
			30%	20.49	7.97%	0.32	89.70%	4.5

[a] Number of Adjustments per 1000 time units

observed that the lead time and percentage tardy increase by roughly 50% while the mean tardiness literally explodes to above 30 time units.

- *Combined effect of an increase in the service rate and fatigue*: This can be observed from the three remaining curves in Fig. 2. While the detrimental effect continues, the prior increase in the service rate softens the negative effect. A fatigue factor of 0.006 increases the percentage tardy by 50% if the adjustment size is 20%. In this scenario, the original average processing time is reached after 34 job completions (approximately 34 time units), and the maximum decrease in the service rate after 50 job completions. For an adjustment size of 30%, the average processing time is reached after 50 job completions for a fatigue factor of 0.006. So, the decrease in the service rate is delayed by approximately 16 time units resulting in a percentage tardy that is approximately 50% lower. While a proportion of this lower percentage tardy is explained by the higher initial adjustment, the majority is explained by the delay to the decrease in the service rate. This emphasizes the importance of avoiding (or postponing) fatigue given that the performance impact is stronger than the gain received from the initial speeding up behavior.

4.4 Discussion of results

Visibility in make-to-order shops with multiple stations, and potentially highly variable routings, is a different matter to the more simplistic problem of visibility when there is only one station (as in Shunko et al. (2018)). In the former case, an indirect load inevitably exists. The further downstream a station is typically positioned in the routing of jobs, the higher this indirect load. Therefore, thresholds need to be adjusted for downstream stations or the workload contribution adapted. Adapting the workload contribution leads to equal thresholds for the direct load and corrected aggregate load (i.e. the aggregate of the direct and indirect load). The simulations have demonstrated the positive effect of the corrected aggregate

Table 9 Performance results for the pure flow shop

Dispatching Rule	Trigger Parameter	Load Measure	Adjust. Size	Lead Time	Percentage Tardy	Mean Tardiness	Average Utilization	Number of Adjust.[a]
ODD	Equal	Direct	0	36.73	22.90%	3.06	90.00%	7.3
			10%	33.49	15.76%	1.43	89.76%	7.9
			20%	32.62	13.58%	1.09	89.66%	8.3
			30%	32.24	12.60%	0.96	89.62%	8.4
		Corrected	10%	35.24	19.89%	2.24	89.88%	2.7
			20%	34.98	19.31%	2.14	89.84%	2.7
			30%	34.86	19.09%	2.09	89.81%	2.7
	Different	Direct	10%	32.84	14.35%	1.25	89.70%	17.4
			20%	31.83	11.95%	0.91	89.59%	17.6
			30%	31.39	10.94%	0.78	89.53%	17.5
		Corrected	10%	34.47	19.06%	1.50	89.87%	2.0
			20%	34.08	18.07%	1.27	89.85%	1.9
			30%	33.91	17.58%	1.18	89.82%	1.9
MODD	Equal	Direct	0	34.15	11.26%	1.07	90.01%	8.0
			10%	32.03	7.65%	0.51	89.74%	8.4
			20%	31.32	6.41%	0.37	89.60%	9.5
			30%	31.00	5.84%	0.32	89.54%	9.7
		Corrected	10%	33.27	9.90%	0.79	89.85%	3.1
			20%	33.06	9.61%	0.75	89.80%	3.2
			30%	32.98	9.48%	0.73	89.78%	3.1
	Different	Direct	10%	30.77	5.84%	0.36	89.52%	36.7
			20%	29.69	4.38%	0.23	89.34%	37.3
			30%	29.18	3.70%	0.18	89.25%	36.9
		Corrected	10%	31.73	4.17%	0.16	89.58%	7.6
			20%	31.03	2.02%	0.05	89.43%	7.5
			30%	30.72	1.44%	0.03	89.32%	7.6

[a] Number of Adjustments per 1000 time units

load – i.e. increased visibility compared to only using the direct load – if different parameters are applied, and this extends the findings in Land et al. (2015) to shops with directed routings. Further, it extends prior research on state-dependent service rates in shops with several stations that only considered the load queuing at a station. The corrected aggregate load is also less granular, which leads to lower system nervousness and thus fewer adjustments. While making workload visible is a major challenge in high-variety shops with complex routings, new technology can be used to provide this information (e.g. Kim 2017).

Meanwhile, this study has highlighted the need for different load thresholds that trigger behavioral change in shops with several stations, specifically if typical upstream and downstream stations exists since the distribution of workload queuing at a station differs across stations. If an equal parameter is applied, capacity adjustments are focused on upstream stations; and this is further illustrated by Table 11, which gives the realized utilization levels in a pure flow shop with ODD dispatching across stations. In fact, while the granularity of the direct load still leads to some adjustments at Station 6 (the most downstream station), no adjustments take place for the corrected load. This may explain why the direct load outperforms the corrected load in this study if an equal parameter is applied. This finding extends previous literature on state-dependent service rates that have focused on serial production lines (e.g. Powell and Schultz 2004; Öner-Közen et al. 2017). To the best of our knowledge, it has been assumed in this existing work that triggering thresholds are equal for all stations, which may have led to an imbalance in the realized service rate – with higher service rates being realized at upstream stations than at downstream stations. This imbalance impacts performance (Hudson et al. 2015) and should be taken into account when determining threshold values. Creating appropriate thresholds speaks to the literature on goal management systems that captures the way an organization shapes employees' decisions and actions through incentives and rewards (Galeazzo et al. 2017). This also extends Land et al. (2015) and Thürer et al. (2016), since

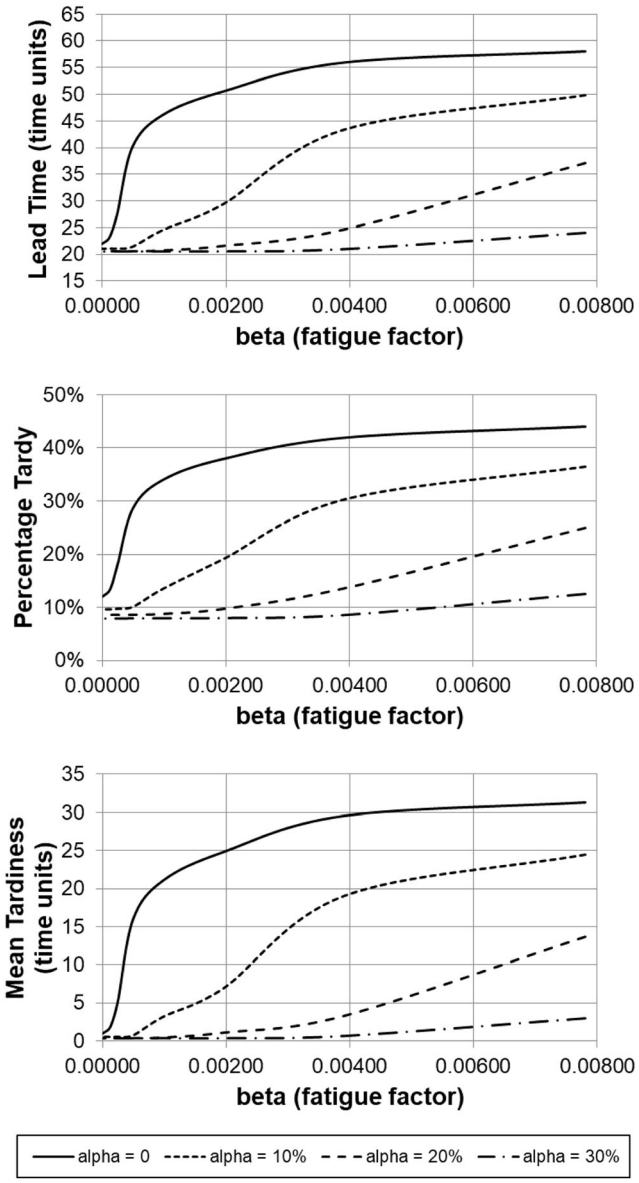


Fig. 2 Impact of the fatigue factor on performance: general flow shop, modd dispatching, different parameters and corrected workload measure

findings from these studies are not directly transferable to shops with directed routings.

Finally, results have demonstrated the important impact of worker fatigue. While capacity adjustments in high load periods lead to significant performance improvements, as demonstrated in Land et al. (2015), this is more than off-set if worker fatigue increases. Öner-Közen et al. (2017) have previously demonstrated that the positive impact of speeding-up behavior has been overestimated in previous studies because of simplifying assumptions, such as a disregard for worker fatigue. However. Öner-Közen et al. (2017) modelled worker fatigue as the diminishing ability of the worker to increase the service rate. In other words, the service rate would never decrease, but rather the increase in the service rate would be reduced. This study extends this finding by showing that, if worker fatigue leads to a decrease in the service rate in response to the initial increase (as in Jaber and Neumann (2010) or Delasay et al. (2016)) then performance may in fact deteriorate.

5 Conclusions

Humans are different from machines. While service rates at machines are typically state-independent, the service rates of humans often depend on the system state, e.g. the workload queuing at a station. This important behavioral phenomenon has received significant research attention, but the literature has typically focused on single/parallel servers or on production lines, and it has assessed the impact on efficiencies. In contrast, the literature has neglected make-to-order shops, often characterized by high routing complexity, where due dates exist and consequently tardiness is arguably the most important performance criterion. Only recently has literature begun to emerge that provides some insight into the impact of state-dependent service rates in make-to-order contexts. Yet this literature has neglected any understanding that can be taken from the literature on state-dependent service rates. It has instead used a different workload measure to trigger adjustments to the service rate: the corrected aggregate load, which provides global information on the load directly queuing at a station and the load still upstream, i.e. the indirect load. Moreover, this emerging research stream has neglected the impact of worker fatigue.

In response to the above, the first research question asked: Should the service rate be dependent on global information on

Table 10 Impact of the fatigue factor on utilization: general flow shop, modd dispatching, different parameters and the corrected workload measure

Adjustment size	Beta (Fatigue Factor)							
	0	0.00012	0.00024	0.00049	0.00098	0.00195	0.00391	0.00781
alpha = 0	90.02%	90.14%	90.67%	91.70%	92.31%	92.82%	91.78%	92.39%
alpha = 10%	89.82%	89.83%	89.83%	89.86%	90.19%	90.73%	89.94%	90.27%
alpha = 20%	89.74%	89.74%	89.74%	89.75%	89.75%	89.82%	89.83%	89.83%
alpha = 30%	89.70%	89.70%	89.70%	89.70%	89.70%	89.71%	89.78%	89.78%

Table 11 Utilization across stations for the pure flow shop with ODD dispatching

Trigger Parameter	Load Measure	Adjust. Size	Average Utilization	Utilization at					
				Station 1	Station 2	Station 3	Station 4	Station 5	Station 6
Equal	Direct	0	90.00%	90.00%	90.03%	90.00%	90.02%	90.96%	90.02%
		10%	89.76%	89.56%	89.76%	89.77%	89.82%	89.89%	89.85%
		20%	89.66%	89.39%	89.66%	89.68%	89.76%	89.82%	89.79%
		30%	89.62%	89.30%	89.60%	89.63%	89.72%	89.78%	89.76%
	Corrected	10%	89.88%	89.47%	89.89%	89.96%	89.99%	89.95%	90.02%
		20%	89.84%	89.26%	89.85%	89.95%	89.99%	89.95%	90.02%
		30%	89.81%	89.15%	89.84%	89.94%	89.99%	89.95%	90.02%
Different	Direct	10%	89.70%	89.90%	89.83%	89.71%	89.66%	89.55%	89.54%
		20%	89.59%	89.87%	89.76%	89.60%	89.54%	89.40%	89.37%
		30%	89.53%	89.86%	89.72%	89.55%	89.47%	89.32%	89.28%
	Corrected	10%	89.87%	89.89%	89.89%	89.87%	89.88%	89.83%	89.89%
		20%	89.85%	89.85%	89.86%	89.84%	89.86%	89.81%	89.87%
		30%	89.82%	89.83%	89.84%	89.82%	89.84%	89.78%	89.84%

the corrected aggregate load or only based on local information on the direct load queueing at a station? Using simulation, this study demonstrated the superior performance of the corrected aggregate load approach. However, the parameter that triggers a service rate adjustment should be different across stations for shops with more directed routings. In response to the second research question – how does routing direction influence performance in shops with state-dependent service rates? – this study found that although performance improvements are robust to changes in the routing direction, an imbalance is created if the same threshold is applied. Upstream stations experience stronger adjustments than downstream stations since the distribution of the load queuing at a station is dependent on the position of the station in the routings of jobs. Finally, and in response to the third research question – what is the performance impact of worker fatigue that results from a prolonged increase of the service rate? – this study found that worker fatigue can offset all initial performance gains if they lead to a decrease of the service rate during high load periods. This has important implications both for practice and research, as will be discussed next.

5.1 Managerial implications

The major message for practice from this study is that humans appear to be able to improve performance beyond that expected from theory. In fact, they may provide an essential buffer for bad planning by increasing the service rate if this is required. However, the contrary is also true, and workers may decrease the service rate even when their superior performance is needed the most. This study has shown that this negative effect, caused by for example worker fatigue, is likely to outweigh the initial performance gains. So, a major concern for practice is how to motivate workers so they exhibit the desired behavior (an issue that was also highlighted in Shunko et al. (2018)), and how to avoid fatigue. In other words, workers should react to surges in workload in a timely manner whilst not being over-motivated or creating overproduction waste. To reap the most benefits out of state-dependent behavior, a balance has to be found that ensures there is a reaction in time, without resulting in an overreaction. To ensure a reaction that is just-in-time, managers should transform state-dependent service rates from an implicit behavior into an explicit behavior. In other words, they should first determine thresholds that require service rate adjustments and then both incentivize and educate their workers on how to react when a threshold is triggered. This study has provided a first indication on how these thresholds should be set in the context of different degrees of routing directedness. However, there is no recommendation for a specific value since this is dependent on idiosyncratic firm characteristics, including the fatigue factor, which is also likely to vary across individual workers.

5.2 Limitations and future research

A first limitation of this study is the relatively narrow environmental setting. This study did not consider factors such as processing time variability or due date tightness in order to keep this study focused. Future research could however explore how these factors affect performance in state-dependent make-to-order shops. This includes research on shops where the service rate is dependent on the urgency of orders. A second limitation of this study is that it assumed state-dependency to be a discrete phenomenon. In other words, service rate adjustments of a certain size are triggered as soon

as a predetermined threshold is violated. In practice, there may be more than one threshold. In general, while there has been extensive empirical research and experiments that have assessed the impact of state-dependent service rates, research that explores the nature of the actual threshold level is scarce. This situation becomes even more challenging if human operators are reallocated across stations. Finally, given the positive impact of state-dependent service rates, future research could explore how state-dependency can be integrated into machines. This is specifically relevant in the context of autonomous machines and cyber physical systems.

Acknowledgements This work was supported by National Natural Science Foundation of China [grant number 71872072]; Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme 2017.

References

Altendorfer K, Hübl A, Jodlbauer H (2014) Periodical capacity setting methods for make-to-order multi-machine production systems. Int J Prod Res 52(16):4768–4784

Assen V (2018) The moderating effect of management behavior for lean and process improvement. Operation Management Review 11:1–13

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

Batt RJ, Terwiesch C (2017) Early task initiation and other load-adaptive mechanisms in the emergency department. Manag Sci 63(11):3531–3551

Bendoly E, Croson R, Goncalves P, Schultz K (2010) Bodies of knowledge for research in behavioral operations. Production & Operations Management 19(4):434–452

Bertrand JWM, van Ooijen HPG (2002) Workload control order release and productivity: a missing link. Prod Plan Control 13(7):665–678

Childerhouse P, Aitken J, Towill DR (2002) Analysis and design of focused demand chains. J Oper Manag 20(6):675–689

Delasay M, Ingolfsson A, Kolfal B (2016) Modeling load and overwork effects in Queueing systems with adaptive service rates. Oper Res 64(4):867–885

Dode P, Greig M, Zolfaghari S, Neumann WP (2016) Integrating human factors into discrete event simulation: a proactive approach to simultaneously design for system performance and employees' well being. Int J Prod Res 54(10):3105–3117

Folgado R, Pecas P, Henriques E (2015) Mapping workers' performance to analyse workers heterogeneity under different workflow policies. J Manuf Syst 36:27–34

Galbreth MR, Philipoom PR, Malhorta MK (2012) Planning with uncertain materials availability: the value of workday flexibility. Oper Manag Res 5:91–100

Galeazzo A, Furlan A, Vinelli A (2017) The organizational infrastructure of continuous improvement – an empirical analysis. Oper Manag Res 10:33–46

Gstettner S, Kuhn H (1996) Analysis of production control systems kanban and CONWIP. Int J Prod Res 34(11):3253–3273

Hopp WJ, Iravani S, Yuen GY (2007) Operations systems with discretionary task completion. Manag Sci 53(1):61–77

Hudson S, McNamara T, Shaaban S (2015) Unbalanced lines: where are we now? Int J Prod Res 53(6):1895–1911

Jaber MY, Neumann WP (2010) Modelling worker fatigue and recovery in dual-resource constrained systems. Comput Ind Eng 59:75–84

Kenne JP, Gharbi A (2001) A simulation optimization approach in production planning of failure prone manufacturing systems. J Intell Manuf 12:421–431

Kim HJ (2017) Information technology and firm performance: the role of supply chain integration. Oper Manag Res 10:1–9

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

Mason S, Baines T, Kay JM, Ladbrook J (2005) Improving the design process for factories: modeling human performance variation. J Manuf Syst 24(1):47–54

Mincsovics GZ, Dellaert NP (2009) Workload-dependent capacity control in production-to-order systems. IIE Trans 41(10):853–865

Neumann WP, Medbo P (2009) Integrating human factors into discrete event simulations of parallel flow strategies. Prod Plan Control 20(1):3–16

Öner-Közen M, Minner S, Steinthaler F (2017) Efficiency of paced and unpaced assembly lines under consideration of worker variability – a simulation study. Comput Ind Eng 111(2017):516–526

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

Powell SG, Schultz KL (2004) Throughput in serial lines with state-dependent behavior. Manag Sci 50(8):1095–1105

Roser C (2016) Faster, better, cheaper in the history of manufacturing: from the stone age to lean manufacturing and beyond. Productivity Press

Samson D, Kalchschmidt M (2019) Looking forward in operations management research. Operation Management Research 12:1–3

Schultz KL, Juran DC, Boudreau JW, McClain JO, Thomas LJ (1998) Modeling and worker motivation in JIT production systems. Manag Sci 44(12):1595–1607

Schultz KL, Juran DC, Boudreau JW (1999) The effects of low inventory on the development of productivity norms. Manag Sci 45(12):1664–1678

Shunko M, Niederhoff J, Rosokha Y (2018) Humans are not machines: the behavioral impact of Queueing design on service time. Manag Sci 64(1):453–473

Sugimori, Y., Kusunoki, K., Cho., F., and Uchikawa, S., 1977, Toyota production system and Kanban system materialization of just-in-time and respect-for-human system, Int J Prod Res, 15, 6, 553–564

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

Thürer M, Stevenson M, Land MJ (2016) On the integration of input and output control: workload control order release. Int J Prod Econ 174:43–53

Van Ooijen HPG, Bertrand JWM (2003) The effects of a simple arrival rate control policy on throughput and work-in-process in production systems with workload dependent processing rates. Int J Prod Econ 85:61–68

Weber M (2014) *Wirtschaft und Gesellschaft: Soziologie*, Studienausgabe der MaxWeber Gesamtausgabe Band I/23. Mohr Siebeck, Tubingen