4OR-Q J Oper Res (2016) 14:173–200
DOI 10.1007/s10288-016-0307-1

RESEARCH PAPER

Pooling heterogeneous products for manufacturing environments

Tanja Mlinar[1] · Philippe Chevalier[2]

Received: 19 January 2015 / Revised: 30 January 2016 / Published online: 3 March 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract In a stochastic environment pooling naturally leads to economies of scale, but heterogeneity can also create variability. In the article, we investigate this trade-off in the case of a manufacturing environment. Pooling for queueing systems has been widely investigated in the literature on the design of service systems; however, much less attention has been given to manufacturing systems where jobs are given a due date upon arrival. In such systems it is not the elapsed time until the actual completion of the job that counts, but rather the due date lead time that can be promised to the customer in order to guarantee a high service level. The purpose of the article is to get a deeper understanding about how pooling strategies and lead-time decisions can be implemented to attain a high due-date performance. To this end, we develop a simulation and analytical study to determine the benefits of pooling manufacturing systems with heterogeneous demand streams. Our work allows managers to identify the characteristics of production systems such that a pooling strategy would be beneficial. Our results demonstrate that when a due-date setting and scheduling mechanism is implemented, heterogeneity does generally not lead to deterioration of performance, as previously observed in service environments. Our studies also reveal that the benefits of pooling in terms of the expected sojourn time obtained by a simple analytical treatment can serve as a good prediction of the benefits of pooling on the average due date lead time in a wide range of situations.

✉ Tanja Mlinar
t.mlinar@ieseg.fr

Philippe Chevalier
philippe.chevalier@uclouvain.be

1 IÉSEG School of Management (LEM-CNRS), Socle de la Grande Arche, 1 Parvis de La Défense, 92044 Paris La Défense Cedex, France

2 Louvain School of Management, Center for Operations Research and Econometrics, Université catholique de Louvain, Place des Doyens 1, 1348 Louvain-la-Neuve, Belgium

Keywords Production systems · Capacity pooling · Due-date scheduling · Queueing · Simulation

Mathematics Subject Classification 60K30 · 60K25 · 90B30 · 68M20

1 Introduction

There is a clear trend in manufacturing towards product variety and heterogeneity. For example, a study about the retail sector in the food industry of several European countries showed that the number of different products increased by 53 % from 2000 to 2010 (Swinnen and Van Herck 2011). Moreover, the number of companies declined while the size of the survivors increased. As a consequence, manufacturers must find ways to diversify their production without losing competitiveness. One way to deal with the expansion of product variety is leveraging economies of scale across different products. The advantages of economies of scale can be viewed in terms of cost savings or better operational performance such as a quick and reliable delivery.

Pooling i.e., the replacement of several facilities/operations by a functionally equivalent single facility/operation (Mandelbaum and Reiman 1998), can represent an important source of economies of scale. Pooling productive activities is becoming a common practice among independent decision-makers across many industries. As an example, we cite the joint venture Behr-Hella Service whose major objective is pooling resources and expertise devoted to the after-market sales and service of their products (Hella Australia Pty Ltd 2007). It is more likely that non-competing companies with similar business and processes will collaborate through pooling. One reason lies in avoiding potential antitrust problems that can occur when competing companies share their operations or resources. The pooled design of non-competing companies faces an increased degree of heterogeneity and variability in customer and process requirements which can diminish the benefits of pooling. It is thus necessary to get a deeper understanding about the interplay between economies of scale and variability in manufacturing.

In this article, we shed a new light on this dilemma by studying a manufacturer whose objective is to quote due dates to its customers with the shortest possible due-date lead times (difference between the due date of the order and its arrival time) while being able to respect those due dates with a high probability. The due date quotations have to be coherent with the scheduling that the manufacturer applies in the production system (Wein and Chevalier 1992). Thus, the objective of this paper is to analyze whether pooling two or more such types of systems allows manufacturers to obtain economies of scale.

By quoting short and reliable due-date lead times, the manufacturer tend to have a powerful marketing advantage (Li and Lee 1994; Duenyas 1995; Easton and Moodie 1999) and can obtain a price premium, attract customers, and ensure long term profitability (Keskinocak and Tayur 2004; Yu et al. 2013). Moreover, on-time delivery of parts enable the manufacturer to maintain the long-term relationships with its customers (see Li et al. 2015). If the manufacturer provides short delivery times by a scheduling rule which shortens flow times, then work-in-process and inventory carrying costs may be also reduced. As a consequence, the company may win more orders (Easton and Moodie 1999), and gain the competitive advantage. In our model we suppose that the manufacturer determines the due-dates unilaterally such as to promise the shortest possible average lead-time. We suppose also that all orders are not differentiated in terms of lead-time importance as is typically the case in large volume manufacturing environments. This would not be applicable to the case where the due-dates of orders are negotiated individually.

Our findings bring both theoretical insights and practical applications. From a theoretical point of view, our contribution lies in providing the first approach to bridge the gap between the literature on pooling queueing systems and the literature on due-date management. The literature on economies of scale associated with pooling queueing systems has considered mainly service environments. In such environments, it is often desirable to maintain a First-Come-First-Served (FCFS) discipline as in e.g., banks, shops, and restaurants in order to treat all customers in a fair manner. The conclusion of these studies is that when customers are heterogeneous the economies of scale tend to disappear because of the additional variability brought by the heterogeneity. In contrast to these studies, we analyze how the additional flexibility brought by the due-date setting and scheduling policies can be used to deal with the variability resulting from pooling production systems with heterogeneous demand streams. Our results reveal that, for a wide range of instances, manufacturers should not be afraid of diversity when pooling production resources. In other words, companies can create economies of scale by pooling their operations even in the case of high heterogeneity and increased product variety.

The remainder of the paper is structured as follows. In Sect. 2, we present an overview of the existing literature. In Sect. 3, we describe the production configurations considered and the due-date setting and scheduling problem. In Sect. 4, we perform a simulation study to evaluate the due date performance of a pooled production system. In particular, we present the experimental design (Sect. 4.1), and describe the results of experiments performed (Sect. 4.2). The simulation study has the following purposes. First, we show that the main insights hold for different pooling scenarios (Sect. 4.2.1). We then demonstrate how much the additional flexibility can be exploited by due-date setting and scheduling policies. To this end, we contrast our results with those for the base case studied extensively in the literature on pooling service operations (Sect. 4.2.2). Finally, we examine the impact of pooling on each type of system/product that is pooled under the different due-date setting policies (Sect. 4.2.3). As simulation models are often expensive and time-consuming to develop (Law 2007, p. 77), in Sect. 5 we develop an analytical study for simplified settings to obtain deeper insights into pooling in manufacturing. Interestingly, we show that our simple analytical treatment can serve as a good indicator of the benefits of pooling on the due-date performance in a wide range of situations. In Sect. 6, we conclude and discuss limitations and future directions.

2 Literature review

In the following subsections, we give an outline of the existing works on related topics and our contribution for each topic.

2.1 Pooling queueing systems

The benefits of pooling for queueing systems have been extensively studied in the operations research and operations management literature starting from Stidham Jr. (1970), Smith and Whitt (1981), Rothkopf and Rech (1987). These benefits are assessed mainly by comparing the expected waiting time (Smith and Whitt 1981; Benjaafar 1995; Dijk and Sluis 2008) or the mean sojourn time (Mandelbaum and Reiman 1998; Enns and Grewal 2010) for pooled and dedicated systems. These studies mainly concern service environments restricted to the FCFS priority. Some applications are in call centers (Dijk and Sluis 2008), communication networks (Smith and Whitt 1981) and computer systems (Kelly et al. 2009).

Research on pooling for manufacturing environments is still scarce. Pooling in production and inventory systems is considered in Iyer and Jain (2004), Benjaafar et al. (2005). Enns and Grewal (2010) investigate pooling in batch manufacturing environments with setups and the FCFS priority. Zhang et al. (2013) consider pooling of make-to-stock (MTS) and make-to-order (MTO) operations and its impact on the three performance measures: average finished goods inventory, MTS stockout probability and MTO queue length. The authors implement a congestion-based switching policy in such a way that machines can dynamically switch between MTS and MTO production. Given the complexity of the problem, the MTS production is controlled by a base stock policy and the MTO production follows a FCFS rule. Motivated by the importance of improving the due-date performance for the competitiveness of a company, we consider a different problem in which we examine the impact of pooling and lead time decisions on the due-date performance.

There are two papers closely related to our work. Dijk and Sluis (2008) investigate the benefits of pooling on the expected waiting time for a multi-server call center which serves two job types according to the FCFS policy. The authors point out that the resource pooling can be counter-productive under high utilization, customer heterogeneity and a small number of servers. In order to reduce the impact of uncertainty, the authors deal with an alternative multi-server flexible system with an overflow and threshold policy. Their results show that such a configuration can outperform two extreme cases: a strictly dedicated and pooled configurations. Tekin et al. (2009) further investigate benefits of partial pooling on the expected waiting time in call centers under the FCFS and a non-preemptive priority. Although, the authors consider a common practice in call centers according to which utilization rates are identical across servers, in manufacturing environments this is often not the case. Given the complexity of the problem considered, the authors use queueing approximations for the expected sojourn time in a multi-server system under the non-preemptive priority, which may limit their model to be accurate only for low levels of heterogeneity. In contrast to the two previous works, we investigate the benefits of pooling in a manufacturing environment with due dates and heterogeneous order types. In such an environment the objective of the company differs and an important measure of performance is due-date related rather than the time that customers spend waiting before being served. In order to minimize the average due-date lead time (DDLT), a policy which prioritizes orders according to their due dates rather than their arrival times is employed. Moreover, we investigate the impact of increased heterogeneity in service times and traffic loads on the decision to pool resources.

The benefits of a pooled single-server system have been investigated in the literature. For example, Iyer and Jain (2004) investigate these benefits for inventory costs when a pooled single capacity is owned by two companies which serve demands from two warehouses with separate ownerships. In a more general setting, Mandelbaum and Reiman (1998) investigate the benefits of pooling in queueing networks on the mean sojourn time. We build upon these works to investigate the due-date performance of a pooled single-facility system in the presence of different due-date setting and scheduling policies. Moreover, we demonstrate that the assumption on a single-facility system does not limit our study nor insights obtained.

2.2 Due-date management

A wide body of research on the due-date management problems focuses on due-date quotation and scheduling issues. The literature on this topic distinguishes two types of due-date setting practices, namely: static and dynamic (see e.g. Keskinocak and Tayur 2004; Kapuscinski and Tayur 2007). The first one assigns a constant DDLT to each incoming customer (as in e.g., Baker and Bertrand 1982; Cheng and Gupta 1989; Azaron et al. 2011); the second one assigns a DDLT dynamically based on the state of the system (as in e.g. Bookbinder and Noor 1985; Sridharan and Li 2008; Kaman et al. 2013; Slotnik 2014). Wein (1991), Wein and Chevalier (1992), and Hopp and Sturgis (2000) develop models which quote due dates dynamically in order to satisfy a global service level such as the percent of orders satisfied on time, the fraction of tardy jobs, and the average job tardiness. Although it is common to quote fixed due dates in practice, a vast body of research finds that dynamic due-date setting methods provide better performance for the measures based on the flow time and tardiness (see e.g. Wein and Chevalier 1992; Vinod and Sridharan 2011).

Most works on due-date quotation and scheduling investigate the influence of different scheduling policies such as the earliest due date, the slack, the shortest processing time, the shortest remaining processing time and the FCFS on the DDLT performance. Examples include Wein (1991), Kaminsky and Lee (2008) and Altendorfer and Jodlbauer (2011) for a single-machine due-date management problem; and Baykasoğlu et al. (2008), Vinod and Sridharan (2011), Chen and Matis (2013) and Hübl et al. (2013) for a job-shop due-date management problem. For instance, Vinod and Sridharan (2011) compare different combinations of existing due-date setting methods and scheduling rules in a dynamic job shop production system. These works consider the global due-date performance measures such as the long-run average flow time, the long-run average DDLT, the percentage of tardy jobs and the average tardiness over all order types. However, to decide to pool or not operations for heterogeneous customers, the impact of such decision should be investigated for each operation that is pooled. Therefore, we investigate how this decision influence the overall performance and the performance of each order type that is pooled.

Another body of research on due-date management takes into account the order selection and customer preferences towards the quoted lead times. Duenyas and Hopp (1995) consider a due-date setting problem with a single customer class and single-facility when the proportion of customers that place orders depend on the quoted lead time. Duenyas (1995) extends the previous study for a single-facility due-date setting problem with multiple customer classes. The customer classes in the study differ in their preferences towards lead times and prices. The author shows that the earliest due date setting policy is optimal in the absence of heterogeneity in order sizes or processing times. The efficiency of the proposed heuristics are evaluated under the variety of different conditions such as different levels of utilization, and different types of demand functions. Kapuscinski and Tayur (2007) derive the optimal policy for the due-date quotation problem with stochastic demand from two customer classes. The classes penalize with different contribution margins the quoted lead times. The authors study the trade-off in which quoting a shorter lead time for a low-margin customer could increase the cost from the high-margin class which would wait longer. Keskinocak et al. (2001) consider a company that decides whether to accept or reject an arriving order. If the order is accepted, a reliable lead time is quoted. The order, however, will be lost if it is not processed within its interval of availability. Li et al. (2012) consider a firm that faces stochastic demands sensitive to price and lead times. The company needs to make decisions regarding the order acceptance or rejection, the trade-off between price and lead-time, and the potential for increased demand against capacity constraints.

In some real-life situations, due dates and prices are determined as a result of sale negotiations with customers, or due dates are imposed by the customer. Such situations have also attracted substantial academic interest. Easton and Moodie (1999) develop a probabilistic model for the optimal pricing and due date setting problem, where customers may decide to accept, reject or to initiate negotiations for more favorable terms. Watanapa and Techanitisawad (1999) extend the model to deal with multiple customer segments. Altendorfer and Minner (2015) develop a queueing model for an MTO production system to treat the situation where customers decide whether to place the order or not based on their required lead time and some information from the manufacturer about the queue state dependent production lead time. Altendorfer (2015) considers an MTS production system of a company with multiple items and random customer required lead times. The company makes decisions concerning a lot size and planned lead time taking into account the customer lead times, setups between the production orders, and service level. The company can satisfy the customer order from the stock, already released production order, or by releasing the new production order. The author shows that in order to increase the service level, the lot size and planned lead time should increase, and thus the pre-production on stock.

Our focus is on manufacturing environments, where repeated purchases are made in a typical MRP type environment. Short due-dates are important as this will improve the overall response time of the supply chain and the reliability is important to avoid cascading perturbations. Also in such environment orders are placed and confirmed on a routine basis, and special requirements in terms of due-date are handled as exceptions. Our objective is to find the due-date quotation and scheduling rules that will minimize the average DDLT under the service level constraint on percentage of orders satisfied on-time as in Wein (1991), Wein and Chevalier (1992), Hopp and Sturgis (2000). Although, there may be situations where a customer may want to negotiate an earlier or later due date than the one quoted, this is considered as an exception and should not alter significantly the overall performance of the system.

Our paper, in which we investigate the potential value of having a more flexible manufacturing environment, represents a direct extension of the existing works on due-date management. In such a flexible environment, manufacturing operations can be pooled under different degrees of heterogeneity in customer requirements, variability, product variety and utilization. Despite the large body of research on due-date management problems, economies of scale for such systems have, to our knowledge, not been explored.

3 Problem description

3.1 A model for dedicated and pooled manufacturing operations

The reviewed literature has evaluated benefits of pooling by comparing the performance of the pooled system with the traditional ”stand-alone” situation in which each order type is assigned to its own facility (see e.g., Mandelbaum and Reiman 1998; Iyer and Jain 2004; Dijk and Sluis 2008; Tekin et al. 2009). Thus, we follow the same idea by considering the two benchmark configurations with equal capacity, namely the dedicated configuration (denoted by d) and the pooled configuration (denoted by p).

We consider that orders from n demand streams arrive to each production configuration dynamically and stochastically along the planning horizon. Each demand stream corresponds to one product type. An order type is characterized by its arrival process and processing time distribution. As soon as an order arrives, a due date must be assigned according to a specific due-date setting policy.[1] Thus, the objective of the company is to quote due dates with the shortest possible average DDLT while being able to provide a high service level of each order type. The service level is defined as the percentage of orders finished on time.

The due-date management problem for the dedicated and pooled configurations with two order types is illustrated in Fig. 1a, b, respectively. We have kept our models quite simple in order to better emphasize the effect of the pooling decision. With such models, we are able to provide a deeper understanding of the relationship between the DDLT performance and pooling strategy.

In the dedicated configuration, each demand stream is handled by an independent facility M_i^d. Whenever a facility is occupied, the orders that arrive in the meantime will form a queue. We denote by q_i^d the queue for orders of type i ($i = 1, 2, \dots, n$) that

[1] We restrict our attention to the case in which a manufacturer assigns a due date to the incoming order (as opposed to customer requested due dates).

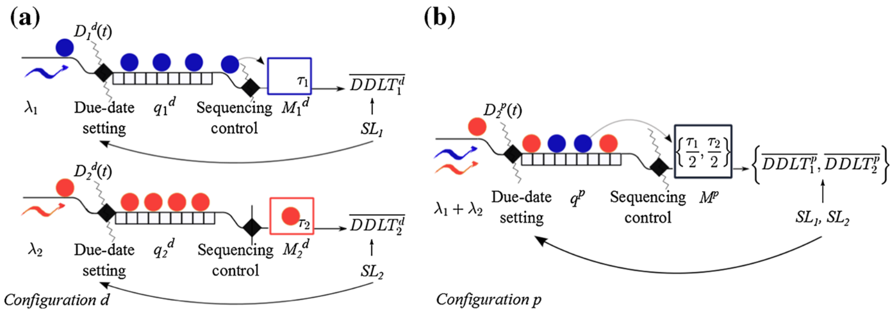


Fig. 1 Due-date management problem for the two production configurations: **a** dedicated and **b** pooled

wait for service at M_i^d (see Fig. 1a). In order to determine which job from the queue to process next, the company employs a specific scheduling policy. The production facility in configuration d could involve a series of processes, but we model it as a single-server queueing system. This is done in order to isolate the effect of heterogeneity in demand patterns and processing times from other effects that can occur in such complex environments. We follow in this a long stream of literature (see e.g., Jordan and Graves 1995; Buzacott 1996; Iyer and Jain 2004; Benjaafar et al. 2005).

In the pooled configuration, we consider that all demand is treated together in a joint facility M^p (see Fig. 1b). We denote q^p the queue for the pooled demand streams that wait for processing at M^p. As for the dedicated configuration, the next job to process is determined by a scheduling policy employed; however, in the pooled configuration, there is more freedom in scheduling because the facility can choose a job from different order types to process next. The pooled configuration results from the aggregation of the different production facilities in order to handle all demand types. We suppose that the production and demand characteristics are not affected by pooling. The operations of the pooled configuration can potentially be even more complex than in the dedicated case, however, we model the resulting facility as a single server queueing system with n order types in order to isolate the effect of the pooling decision. In Sect. 4.2.1, we show that the main insights obtained for the single-server pooled configuration hold for the pooled configuration modelled as a multi-server parallel queueuing system.

We assume that order type i has its own exponential processing time distribution with mean τ_i and its Poisson arrival process with rate λ_i. We make such assumptions because a coefficient of variation close to 1 corresponds to typical values for variability in manufacturing settings (Hopp and Spearman 2008). The traffic load of order type i is $\rho_i = \lambda_i \cdot \tau_i$. To ensure the stability of the system $\rho_i < 1$. Therefore, configuration d is modeled as n independent M/M/1 queueing systems. Similarly, configuration p is modeled as a multi-class M/M/1 queueing system. This could be the case, for example, when two partners aim to determine whether to build their own facility or to invest into a shared one. Thus, in order to make a comparison between the two configurations, we consider that the global capacity of the pooled configuration is equivalent to the sum of the production capacities of the n dedicated production systems. Consequently, the processing times are n times faster in configuration p. The assumption about the equal capacities is made in a large number of previous articles about pooling (see e.g., Smith and Whitt 1981; Benjaafar 1995; Mandelbaum and Reiman 1998; Benjaafar et al. 2005; Dijk and Sluis 2008).

We define $D_i^s(t)$ as the due date assigned to the order of type i arriving at time t to configuration s ($s \in \{d, p\}$). In configuration d, the due date is assigned to the incoming order so as to minimize the average $DDLT$ of type i, denoted by $\overline{DDLT_i^d}$, under the constraint on service level of order type i, denoted by SL_i (the percentage of orders of type i finished on time). In configuration p, due dates are assigned in order to minimize the average $DDLT$ of all order types while respecting the service level of each order type.

3.2 Operating policies implemented in the manufacturing operations

In this section, we describe the scheduling policies (SP) and the due-date setting policies (DDSP) considered.

3.2.1 Scheduling policies

In the dedicated production configuration with due dates, the orders are processed according to the order that they arrived i.e., FCFS policy. In the pooled production configuration with due dates, we consider three scheduling policies mostly studied in the literature and generally preferred in practice (see e.g., Keskinocak and Tayur 2004; Baykasoğlu et al. 2008; Kaminsky and Lee 2008; Hübl et al. 2013): the shortest expected processing time (SEPT) scheduling policy which assigns higher priorities to order types with shorter expected processing times and it minimizes the expected sojourn time, earliest due date policy (EDD) which assigns the priority to the job with the earliest due date, and the least remaining slack policy (SLACK) which assigns the priority to the job with the smallest slack (where the slack is the difference between the due date of the order and its mean processing time plus time at which the scheduling decision is being made).

3.2.2 Due-date setting policies

There are two common practices to assign due dates in production systems. The choice of using either practice is often imposed to the manufacturer by the context in which he/she is operating; thus, we examine the decision to pool resources for both practices.

One practice, the *static policy* (CON), assigns a constant DDLT to orders irrespective of the state of the facilities. Thus, $D_i^s(t)$ can be expressed by the expression (1).

$$D_i^s(t) = t + f_i^s, \tag{1}$$

where f_i^s represents the due-date lead time of an order of type i in configuration s. Parameter f_i^s is obtained via simulation such as to satisfy the service level of order type i.

Because of its simplicity for both the manufacturer and the customer, the static policy is widespread among the MTO and MTS manufacturers. In particular, in the MTS settings, a shorter lead time implies a lower inventory level.

Another practice, the *dynamic due-date policy* (DYN), is more sophisticated because it assigns a due date to the incoming order based on the state of the system. This practice is mainly implemented in pure MTO manufacturing settings and leads to higher efficiency compared to the static policy, but the variability of the resulting lead times tends to complicate the planning process for the customers. Following the literature (see e.g., Bookbinder and Noor 1985; Wein 1991; Hopp and Sturgis 2000; Vinod and Sridharan 2011), we employ a parametric DYN. Values of the parameters are determined in such a way to obtain a high due date performance. Thus, $D_i^s(t)$ can be expressed by the following expression.

$$D_i^s(t) = t + r_i \cdot E[S_i^s | X(t)], \tag{2}$$

where r_i represents a parameter obtained via simulation in order to satisfy the service level of order type i, and $E[S_i^s | X(t)]$ represents the expected sojourn time of the order of type i that arrived at time t conditional on the state of the system $X(t)$ in configuration s.

In configuration d, the dedicated facility M_i^d can process only orders of type i; therefore, $E[S_i^d(t) | X(t)]$ is given by:

$$E[S_i^d | X(t)] = \tau_i + m_i^d(t) \cdot \tau_i + E[R_i^d(t)], \tag{3}$$

where $m_i^d(t)$ represents the number of orders of type i at time t at configuration d waiting to be processed, and $E[R_i^d(t)]$ is the expected residual processing time of the order of type i being processed by M_i^d at time t. For the exponential processing times, the value of $E[R_i^d(t)]$ is defined as follows:

$$E[R_i^d(t)] = \begin{cases} \tau_i, & \text{if } M_i^d \text{ is busy,} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

In order to assign a due date with a short lead time, the cycle time should be as short as possible. Therefore, to minimize the average DDLT, the due-date setting policy which relies on the scheduling policy that minimizes the expected sojourn time should be implemented. Therefore, in configuration p, $E[S_i^p | X(t)]$ can be obtained by expression (5) assuming the SEPT policy (Wein 1991).

$$E[S_i^p | X(t)] = \frac{\tau_i}{n} + \frac{\sum_{j=1}^{n} m_j^p(t) \cdot \frac{\tau_j}{n} + E[R^p(t)]}{1 - \sigma_{i-1}}, \tag{5}$$

where $m_j^p(t)$ represents the number of orders of type j at time t at configuration p waiting to be processed, $E[R^p(t)]$ is the expected residual processing time of the order being processed at time t by M^p, and parameter $\sigma_i = \frac{1}{n}\sum_{j=1}^{i}\rho_j$. For the exponential processing times, the value of $E[R^p(t)]$ is defined by expression (6).

$$E[R^p(t)] = \begin{cases} \dfrac{\tau_i}{n}, & \text{if at time } t \text{ an order of type } i\text{(i=1,2,...,n) is being processed,} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

3.3 Evaluation criteria for pooling

We evaluate the gains that arise from pooling n dedicated production systems for a specific performance measure for order type i, denoted by $\varepsilon(PM_i)$, by the following expression:

$$\varepsilon(PM_i) = \frac{PM_i^d - PM_i^p}{PM_i^d}, \tag{7}$$

where PM_i^s is the value of the performance measure for order type i in configuration $s \in \{d, p\}$.

The pooling gains for a specific global performance measure is evaluated in a similar way by the following expression.

$$\varepsilon(PM_g) = \frac{PM_g^d - PM_g^p}{PM_g^d}, \tag{8}$$

where PM_g^s is the global performance measure for configuration $s \in \{d, p\}$ (the average value over all order types).

The pooling gain gives the proportion of improvement in the individual or global performance measure when going from the dedicated to the pooled configuration.

As the objective of the manufacturer in Sect. 3.1 is to assign due dates with the shortest possible average DDLT, the performance measures of the interest for the manufacturer are the average DDLT of order type i and the global average DDLT (the average DDLT over all order types) in configuration s, denoted by $\overline{DDLT_i^s}$ and $\overline{DDLT_g^s}$, respectively. Thus, the evaluation criteria for pooling in manufacturing with due dates are the pooling gains for the average DDLT of order type i and the global average DDLT, $\varepsilon(\overline{DDLT_i})$ and $\varepsilon(\overline{DDLT_g})$, respectively. In other words, the pooling gains are defined as the relative reductions in the average DDLTs if we pool dedicated capacities.

We also consider the expected sojourn time as a performance measure for pooling gains. In particular, we denote by $E\left[S_i^s\right]$ and $E\left[S_g^s\right]$ the expected sojourn time of order type i and the global expected sojourn time in configuration s, respectively. The corresponding pooling gains are $\varepsilon(E\left[S_i\right])$ and $\varepsilon(E\left[S_g\right])$. If the value of pooling gains for the the expected sojourn time are negative, it indicates that the performance is degraded by pooling.

There are two reasons to calculate pooling gains for the expected sojourn time. As we aim to compare pooling in manufacturing with pooling in services, we evaluate the decision to pool service operations by computing the pooling gains for the expected sojourn times. This is because the expected sojourn time represents one of the most important measures of a service system. Another reason lies in analytical tractability of the model for simplified settings where the expected sojourn time is considered instead of the due-date performance. Given that the DDLT that can be promised is influenced by an order's expected sojourn time, we aim to investigate whether the pooling gains for the expected sojourn time obtained for such simplified settings can represent good indicators of pooling gains for the DDLT.

4 Simulation study

We perform a simulation study in order to evaluate and compare the average DDLTs of the two configurations. Simulation has been used extensively in the literature on due-date setting and scheduling (see e.g., Baker and Bertrand 1982; Veral 2001; Vinod and Sridharan 2011; Chen and Matis 2013). The main reason lies in the analytical complexity of due-date setting and scheduling problems which obviate the possibility of a theoretical treatment as explained in Barman (1998).

4.1 Experimental design

4.1.1 Experiment setting

In the numerical experiment, we consider that the manufacturer processes orders of two types. We choose this setting because the contrast between orders is the largest and the pooling decision is the most difficult. Indeed, orders with intermediary characteristics would reduce the variance of the order sizes. We discuss this choice in Sect. 5.2.

The performance of the system depends on the characteristics of order types. Therefore, an instance is determined by the traffic loads of each type (ρ_1 and ρ_2) and the coefficient of heterogeneity c, defined as the ratio between the two order types ($c = \tau_2/\tau_1$). We fix τ_1 at unity and a service level of 95 % of on-time delivery. We remark that for a fixed traffic load there exists a direct link between the processing times and the arrival rates i.e. $\lambda_i = \rho_i/\tau_i$.

Utilization. To isolate the effect of heterogeneity in processing times or arrival rates from the effect of asymmetry in workloads, we initially consider the case where the workloads for the two order types are equal (balanced case). In the balanced case, we try four different values for the average utilization rate: $\rho = \{0.6, 0.7, 0.8, 0.9\}$. We then consider the unbalanced case with assymetric workloads. We represent the assymetry in workloads by the coefficient δ defined as $\delta = \rho_1/\rho_2$. We choose two values for the asymmetry in workloads and the average utilization rate: the workload of one order type is 50 % higher/lower than the workload of the other type ($\delta = \{2/3, 3/2\}$) and $\rho = \{0.6, 0.75\}$. This gives us four different combinations of ρ_1 and ρ_2 i.e. $(\rho_1, \rho_2) = \{(0.48, 0.72), (0.72, 0.48), (0.6, 0.9), (0.9, 0.6)\}$.

Coefficient of heterogeneity. In order to investigate the impact of heterogeneity in processing times on the decision to pool production systems, we set the coefficient of heterogeneity to take the following values $c = \{4, 8, 12\}$. Therefore, in our experiment, for a given traffic load, the order type will either have a fast service and frequent arrivals or a long service with infrequent arrivals. In practice this corresponds to situations where one firm would be targeting clients with small orders, and another one is targeting clients with larger orders.

4.1.2 Experimental procedure

In the simulation study, we consider six combinations of due-date setting and scheduling policies for configuration p: CON-SEPT, CON-EDD, CON-SLACK, DYN-SEPT, DYN-EDD and DYN-SLACK; and two for configuration d: CON-FCFS and DYN-FCFS (DDSP-SP refers to a combination of due-date setting policy DDSP and scheduling policy SP). Note that, we compare the due-date performance of the two configurations under the same DDSP policy.

In total, there are 144 and 48 different instances for the pooled and dedicated configurations, respectively. We construct two simulation models for the two production configurations in order to study the pooling gains for these instances. Each experiment was simulated during a period T = 200,000 time units and was replicated 40 times using unique random numbers. The system was started empty and without a warm-up period. We chose these simulation parameters in order to ensure that for the most variable settings (long service times with infrequent arrivals) the 95 % confidence interval does not exceed 5 % of the average DDLT. Thus, for each instance we run the simulation models to obtain the average DDLTs, and then compute the pooling gains for the average DDLTs by expressions (7) and (8).

4.2 Insights on pooling gains for the due-date performance measures

In Tables 1 and 2, we present the pooling gains for each order type and the global pooling gain under the SEPT scheduling policy applied in the pooled configuration for the balanced and unbalanced case, respectively.

Table 1 The pooling gains $\varepsilon(\overline{DDLT}_j)$ for the DDLT of order type $j = 1, 2$ and global system $j = g$ (in %) under the balanced case and the SEPT policy

ρ	DDSP	$c = 4$			$c = 8$			$c = 12$		
		$\varepsilon(\overline{DDLT}_j)$			$\varepsilon(\overline{DDLT}_j)$			$\varepsilon(\overline{DDLT}_j)$		
		$j = 1$	$j = 2$	$j = g$	$j = 1$	$j = 2$	$j = g$	$j = 1$	$j = 2$	$j = g$
0.6	CON	30.58	50.12	40.36	−25.12	53.53	14.25	−82.63	53.60	−15.13
	DYN	34.02	54.33	44.15	−2.22	56.92	27.32	−39.03	57.75	9.15
0.7	CON	42.15	47.15	44.62	−4.36	50.66	23.13	−50.72	52.67	1.13
	DYN	40.45	51.93	46.17	5.27	56.21	30.63	−29.63	58.04	14.47
0.8	CON	58.20	43.85	51.04	24.66	48.48	36.51	−10.70	49.56	19.23
	DYN	51.22	48.48	49.83	21.53	54.10	37.81	−8.39	55.97	23.75
0.9	CON	77.08	42.25	59.34	58.79	45.79	52.45	40.14	45.52	42.79
	DYN	68.61	41.83	55.29	48.63	48.58	48.62	29.18	50.89	39.91

Table 2 The pooling gains $\varepsilon(\overline{DDLT}_j)$ for the DDLT of order type $j = 1, 2$ and global system $j = g$ (in %) under the unbalanced case and the SEPT

ρ	δ	DDSP	$c = 4$			$c = 8$			$c = 12$		
			$\varepsilon(\overline{DDLT}_j)$			$\varepsilon(\overline{DDLT}_j)$			$\varepsilon(\overline{DDLT}_j)$		
			$j = 1$	$j = 2$	$j = g$	$j = 1$	$j = 2$	$j = g$	$j = 1$	$j = 2$	$j = g$
0.60	0.67	CON	7.99	64.62	49.54	−69.10	66.34	30.56	−148.00	65.85	8.46
		DYN	16.29	64.13	50.16	−33.22	66.59	37.54	−84.93	67.09	22.95
0.60	1.50	CON	52.80	36.36	48.46	16.50	40.67	22.87	−20.75	42.09	−4.13
		DYN	51.72	44.94	49.75	27.93	49.59	34.26	4.24	50.85	17.93
0.75	0.67	CON	19.79	78.83	70.42	−47.24	79.37	60.87	−115.15	79.77	51.28
		DYN	19.92	76.43	66.18	−32.18	77.59	57.25	−85.14	77.65	47.15
0.75	1.50	CON	79.70	11.74	69.93	64.14	19.73	57.77	48.94	23.03	45.23
		DYN	74.25	29.05	66.06	59.97	37.92	55.94	46.63	40.29	45.49

We choose omit the presentation of the results under the other two scheduling policies and to focus on managerial insights obtained only for the SEPT policy. This is done because, for the majority of instances, the SEPT policy is superior over the other two policies in terms of the global pooling gains. For the remaining instances, the performance advantage of the other two scheduling policies is negligible. Moreover, the SEPT policy outperforms considerably the other two scheduling policies under the CON due-date setting policy. The results also reveal that performance improvement in the global pooling gain brought by the SEPT policy increases compared to the due-date based scheduling policies as the coefficient of heterogeneity increases. This improvement can be attributed to the higher priority given to orders with shorter expected processing times that are the orders most easily hurt by pooling.

4.2.1 Pooled multi-server configurations

For simplicity, we assume that the pooled facility consists of a single server with an increased capacity; however, a more general case for pooling represents a multi-server pooled configuration. Thus, before discussing the results obtained, we first investigate the impact of our modeling assumption on the pooling gains by comparing it with a pooled facility with multiple servers.

It can be easily shown that the expected sojourn time will be longer in the multiple-server model than in the single-server model with equivalent capacity. On the other hand, the expected queueing time will be shorter; this gives an indication that the sojourn time will not be too different. Here, we develop an additional simulation model for the pooled two-facility configuration in order to obtain insights into the average DDLTs in the two pooled configurations. We model the pooled two-facility configuration as an M/G/2 queueing system with the mean processing time and arrival rate of order type i, τ_i and λ_i, respectively.

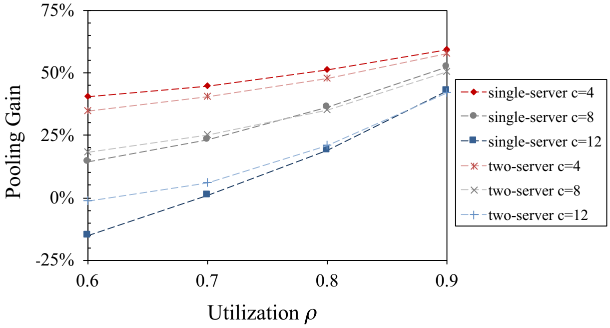


Fig. 2 Pooling gains (in %) for the global average DDLT for the single-server and the two-server configurations under the CON-SEPT, the balanced case, and different values of utilization ρ and coefficient of heterogeneity c

In Fig. 2, we present the results for the pooling gains for the global average DDLT achieved in the two pooled configurations under the balanced case and the CON-SEPT policy combination. We observe that the pooling gains of the two pooled configurations are similar, especially under the higher utilization rate. This confirms that our modeling assumption on the pooled single-server facility does not significantly influence our results. Note that, we observe a higher difference in the pooling gains only under the very high heterogeneity ($c = 12$) and low utilization rate (utilization equal to 60 %). However, in both configurations the pooling gains are negative. Moreover, we remark that majority of companies in the manufacturing sector operate with average utilization >60 %. For instance, the average capacity utilization in the manufacturing sector in the United States for August 2015 is 75.8 % according to the Board of Governors of the Federal Reserve System (2015).

4.2.2 *Pooling in manufacturing versus pooling in services*

Here, we contrast the results concerning pooled manufacturing operations with due dates with results concerning pooled service operations. Thus, to obtain insights for pooling in services, we compute the pooling gain for the global expected sojourn time under the balanced case and the FCFS policy which is the policy mostly applied in services. Note that, to compare the two operations, the dedicated and the pooled configurations in services are modeled equivalently to those in manufacturing. Thus, the dedicated configuration in services is modeled as two M/M/1 queues and the pooled configuration as two class M/M/1 queue with a two times faster server. The values of parameters for the two service configurations are the same as set in Sect. 4.1.1.

The pooling gain in services can be calculated analytically as $\varepsilon\left(E\left[S_g\right]\right) = \dfrac{1}{2} - \dfrac{\rho(c-1)^2}{8c}$. The value is obtained from expression (8), where $E\left[S_g^d\right]$ is calculated based on the expressions for $M/M/1$ queue [(see expressions (9) and (12)] and $E\left[S_g^p\right]$ is calculated as $E\left[S_g^p\right] = E\left[W_g^p\right] + \overline{\tau}^p$, where $E\left[W_g^p\right]$ and $\overline{\tau}^p$ denote the expected waiting time and the expected processing time in configuration p. We can compute

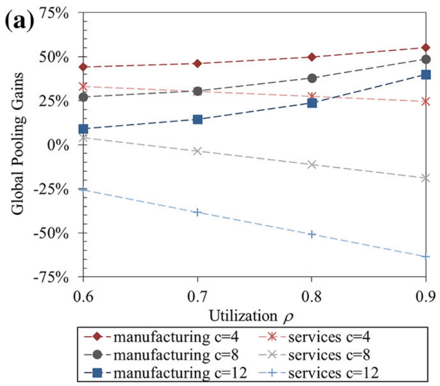
$E\left[W_g^p\right]$ and $\overline{\tau}^p$ by using the following expressions $E\left[W_g^p\right] = \frac{1}{2}(1+v^2)\frac{\rho \cdot \overline{\tau}^p}{1-\rho}$ and $\overline{\tau}^p = \frac{\sum_i^n \lambda_i \tau_i^p}{\sum_i^n \lambda_i}$, where $v^2 = \frac{\left(\sum_i^n \lambda_i\right) \sum_i^n \lambda_i (\tau_i^p)^2 (v_i^2+1)}{\left(\sum_i^n \lambda_i \tau_i^p\right)^2} - 1$ and $v_i^2 = 1$ $\forall i = 1, \ldots, n$ for the exponential processing times. We refer the reader to Dijk and Sluis (2008), Tekin et al. (2009) for more details about the computations and insights of pooling gains under the FCFS policy.

Based on our results we make the following observations.

Impact of heterogeneity. The global pooling gains in manufacturing are less sensitive to changes in the coefficient of heterogeneity c than the global pooling gains obtained in services (see Fig. 3). Moreover, whenever it is preferable to pool in services, it is also preferable in manufacturing. The opposite is not the case; for any level of heterogeneity, the pooling gains are greater in manufacturing.

Impact of the global utilization rate. Pooling brings drastically different gains in the two environments studied due to the different scheduling policies implemented. On the one hand, the pooling gains increase with utilization in the context of manufacturing, while, on the other hand, the pooling gains decrease in the context of service. We note from Fig. 3 that as utilization decreases the pooling gains are more similar. This is because as the utilization rate decreases, the pooled configuration faces fewer arrivals; thus, there are less opportunities to prioritize between orders based on their sizes or due dates i.e., jobs are more often processed in the order in which they arrive to the system which reduces the impact of the analyzed scheduling policies.

Finally, we conclude that the positive effect of a due date setting and scheduling mechanism implemented in the manufacturing environment which increases with utilization will prevail over the negative effect of heterogeneity in a pooling environment.



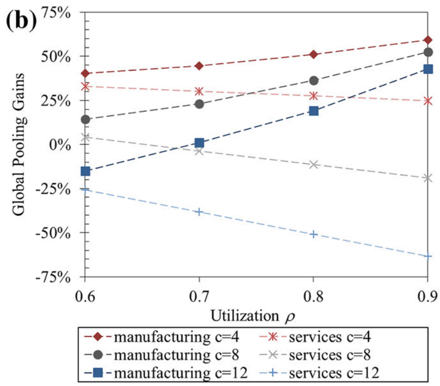


Fig. 3 Pooling gains (in %) in manufacturing ($\varepsilon\left(\overline{DDLT_g}\right)$) versus pooling gains in services ($\varepsilon\left(E\left[S_g\right]\right)$) for the balanced case, different values of utilization ρ and coefficient of heterogeneity c, and the two policy combinations: **a** DYN-SEPT and **b** CON-SEPT

4.2.3 Impact of the due-date setting and values of the individual performance measures on the pooling decision

Impact of the due-date setting practice. Our results on the global pooling gains reveal that companies which operate under the DYN due-date setting policy should pool their operations. In particular, pooling is advantageous for all instances (see Tables 1, 2). On the other hand, companies which operate under the constant due-date setting practice should note that the pooling gains under the CON due-date setting policy are more sensitive to changes in the coefficient of heterogeneity c than under the DYN policy.

Significance of individual and global performance measures for making the pooling decisions. Most papers in the extant due-date management literature study only global due-date performance measures. Our results show that this can lead to considerable overestimation/underestimation of the performance of individual order types when making the pooling decision. In particular, the difference in individual pooling gains can be very large, even-though the global pooling gains are positive (see the pooling gains obtained for $c = 8$ on Table 1).

The large-size order always benefits from pooling. The pooling gain is almost constant and generally more beneficial for the order type with a long processing time and infrequent arrivals, type 2 (see Fig. 4a), compared to the pooling gain for the order type with small size and frequent arrivals, type 1 (see Fig. 4b). Heterogeneity and utilization have a more pronounced impact for the type 1.

In particular, the pooling gains for the DDLT of order type with the small size (type 1) increase as the coefficient of heterogeneity c decreases and the utilization ρ increases. If the incoming order of type 1 finds the facility processing a type 2 order, then its sojourn time will increase with the coefficient c; and, therefore, its DDLT. The pooling gains for type 1 increase with its traffic load because the efficiency of the scheduling policy which prioritizes more orders of that type increases with the utilization rate. This will consequently improve the due date performance of type 1 in the pooled configuration.

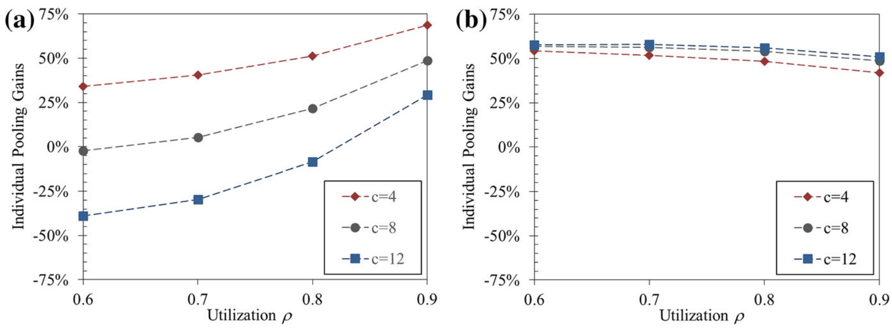


Fig. 4 Pooling gains (in %) for the DDLT of **a** order type 1 and **b** order type 2 under the balanced case, the DYN-SEPT policy combination and different values of utilization ρ and coefficient of heterogeneity c

For the order type with the large size (type 2), we observe the opposite trend although much weaker when $\rho_1 \geq \rho_2$ i.e. coefficient of asymmetry $\delta \geq 1$ (see Tables 1 , 2). For $\rho_2 > \rho_1$ i.e. $\delta < 1$ and higher differences in traffic loads (see Table 2) the pooling gain for type 2 increases in ρ. This is because the facility will face more orders of that type than of type 1 as ρ increases and δ decreases, meaning that there will be less possibility to prioritize orders of type 1; thus, orders of type 2 will be less hurt when pooling.

Moreover, the pooling gains for type 2 slightly increase when heterogeneity increases. In the pooled configuration, the incoming order of type 2 would wait less time for its processing than in the dedicated configuration when coefficient of heterogeneity c increases. This is because the incoming order can find orders of type 1 with a much smaller size in the pooled queue. The policy which gives higher priority to type 1 does not hurt considerably type 2. We later show in the analytical study on the pooling gains for each order type (see Sect. 5.1) that, for a multi-type system, order types with a large size always benefit from pooling.

5 Analytical study

In this section, we address two questions that arise when analyzing the pooling gain on the DDLT performance measures. The first question aims to provide additional insights about the impact of heterogeneity and the utilization rate on the pooling gains. In order to answer this question, we will study analytically the pooling gains for the expected sojourn time when using the SEPT policy.[2] The next question deals with the key assumption we made in our preliminary analysis with regard to the number of pooled operations and order types. Thus, in Sect. 5.2, we examine the impact of this assumption.

5.1 Pooling gains for the expected sojourn time under the SEPT policy

Let $E\left[S_i^d\right]$ be the expected sojourn time of orders of type i in configuration d with n order types. We calculate $E\left[S_i^d\right]$ based on the formula for an M/M/1 queueing system:

$$E\left[S_i^d\right] = \frac{\tau_i}{(1-\rho_i)}. \tag{9}$$

For configuration p with the SEPT policy, let $E\left[S_i^p\right]$ be the expected sojourn time of orders of type i in p. In order to calculate $E\left[S_i^p\right]$, we assume the order types $1, 2, \ldots, n$ are sorted in ascending order of their expected processing times so that $\tau_1 \leq \tau_2 \ldots \leq \tau_n$. Given that $E\left[S_i^p\right] = E\left[W_i^p\right] + \tau_i^p$, where $E\left[W_i^p\right]$ and τ_i^p denote the expected waiting time and expected processing time of type i in the configuration p, respectively, $E\left[S_i^p\right]$ can be computed based on the formula for the expected waiting time (Cox and Smith 1961):

[2] Given that the DDLT that can be promised is influenced by an order's expected sojourn time, we implement the SEPT policy because it minimizes the expected sojourn time and its computation is analytically tractable.

$$E\left[W_i\right] = \left[\sum_{j=1}^{n} \rho_j \cdot E(R_j)\right] \Big/ \left[\left(1 - \sum_{j=1}^{i} \rho_j\right) \cdot \left(1 - \sum_{j=1}^{i-1} \rho_j\right)\right], \tag{10}$$

where $E(R_j)$ represents the expected residual processing time of order type j and is equal to τ_j for the exponentially distributed processing times.

Noting that, in configuration p, the traffic load and the mean processing time of type i are ρ_i/n and τ_i/n, respectively, $E\left[S_i^p\right]$ is calculated with the following expression:

$$E\left[S_i^p\right] = \frac{\sum_{j=1}^{n} \rho_j \tau_j}{\left(n - \sum_{j=1}^{i} \rho_j\right) \cdot \left(n - \sum_{j=1}^{i-1} \rho_j\right)} + \frac{\tau_i}{n}. \tag{11}$$

The global (total) expected sojourn time for production configuration $s \in \{d, p\}$ with n order types, $E\left[S_g^s\right]$, is calculated as follows:

$$E\left[S_g^s\right] = \frac{\sum_{i=1}^{n} \lambda_i E\left[S_i^s\right]}{\sum_{i=1}^{n} \lambda_i}. \tag{12}$$

We obtain the pooling gains for the expected sojourn time of type i and global expected sojourn time, $\varepsilon\left(E\left[S_i\right]\right)$ and $\varepsilon\left(E\left[S_g\right]\right)$, respectively, by expressions (7) and (8).

In the following two subsections, we start by computing the pooling gains for the expected sojourn times under the SEPT policy when the workloads of the different order types are equal (balanced case) and then, we show how our results generalize to the case where the workloads are different (unbalanced case).

5.1.1 Balanced case

For the balanced production configuration with n order types, the traffic loads across order types are equal i.e. $\rho = \rho_i \ \forall i = 1, \ldots, n$. Then, based on the Eqs. (9)–(12), the pooling gain for the global expected sojourn time $\varepsilon(E\left[S_g\right])$ can be expressed in terms of utilization ρ, n order types, and mean processing time of type i (τ_i) by the following expression:

$$\varepsilon\left(E\left[S_g\right]\right) = 1 - \frac{1-\rho}{n}\left(1 + \frac{\rho \sum_{i=1}^{n} \tau_i}{\sum_{i=1}^{n} \tau_i (n - i\rho)(n - (i-1)\rho)}\right). \tag{13}$$

Pooling outperforms the dedicated configuration when $0 < \varepsilon\left(E\left[S_g\right]\right) \leq 1$.

In order to better understand how heterogeneity in processing times/arrival rates influences the pooling gain represented by Eq. (13), we consider that the coefficient of heterogeneity c, defined as the ratio of the expected processing times or arrival rates between successive order types, is constant. In other words, $c = \tau_i/\tau_{i-1}$ and $c = \lambda_{i-1}/\lambda_i$, where $c \geq 1$. As a consequence, $\tau_i = c^{i-1} \cdot \tau_1$ and $\lambda_i = \lambda_1/c^{i-1}$. In this way, lower values of c (i.e. close to 1) reflect the case when order types are homogeneous; higher values of c reflect higher heterogeneity between processing times or arrival rates. This stylized model enables us to summarize a large variety of instances with only three parameters ρ, c and n. The pooling gain can now be expressed in terms of these three parameters:

$$\varepsilon\left(E\left[S_g\right]\right) = 1 - \frac{1-\rho}{n}\left(1 + \frac{\rho \cdot \sum_{i=1}^{n} c^{i-1}}{\sum_{i=1}^{n} c^{i-1}\left(n - i\rho\right)\left(n - (i-1)\rho\right)}\right). \quad (14)$$

Similarly, based on Eqs. (9) and (11) and replacing τ_i by $c^{i-1} \cdot \tau_1$, we obtain the pooling gain for the expected sojourn time of order type i, $\varepsilon(E\left[S_i\right])$:

$$\varepsilon\left(E\left[S_i\right]\right) = 1 - \frac{1-\rho}{n}\left(1 + \frac{\sum_{j=1}^{n} c^{j-1}}{c^{i-1}} \cdot \frac{n\rho}{(n - i\rho)(n - (i-1)\rho)}\right). \quad (15)$$

Our numerical tests obtained by using Eq. (14) indicate that the SEPT policy leads to a significant improvement in the global expected sojourn time of the pooled system as the capacity utilization increases. Moreover, as the utilization increases, pooling becomes advantageous even for very high coefficients of heterogeneity. For instance, when pooling two order types ($n = 2$), pooling is advantageous for both order types when $c < 8.33$ under the utilization of $\rho = 0.6$; however, it is advantageous for much higher values of c (i.e., $c < 22.22$) under the utilization of $\rho = 0.9$. This is because the impact of scheduling on the expected sojourn time under low and medium utilization is lower than under high utilization. In other words, the queue length is reduced so the scheduling policy loses its efficiency and it tends to perform like the FCFS policy. Given that jobs with fast services are mixed with jobs with long services, the expected waiting time of the fast service class will increase, and thus, its expected sojourn time. As for our simulation results on pooling gains for the average DDLTs, we can conclude that an increase in the performance of the pooled system is driven by the scheduling policy implemented.

Figure 5 shows the difference between the pooling gains for the expected sojourn time for the SEPT policy and the average DDLT for the DYN-SEPT policy combination. We observe a remarkable correspondence between the two curves. This is because when quoting due dates, sojourn times are taken into account. We stress here that actual values of the performance measures can be significantly different, eventhough the global pooling gains are similar. In particular, the global average DDLT is greater than the expected sojourn time in order to ensure that 95 % of orders are satisfied within the promised due dates. Finally, we observe that the gap between the two curves is very small and it decreases when ρ and c decrease.

5.1.2 Unbalanced case

For the unbalanced production configuration with n order types, the traffic loads across order types are unequal i.e. $\rho_1 \neq \rho_2 \neq \cdots \neq \rho_n$. In order to obtain the insights into the pooling gains for different levels of heterogeneity in traffic loads, we assume that the traffic load of order type i can be represented in terms of the traffic load of its preceding

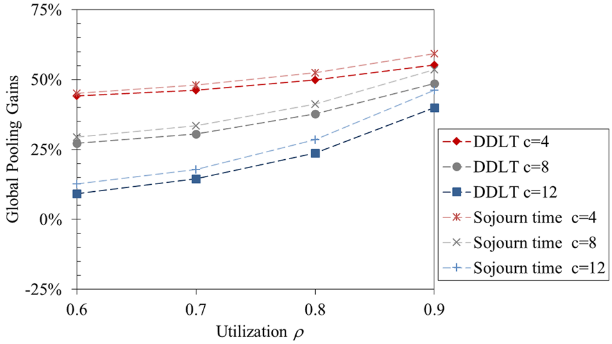


Fig. 5 Pooling gains (in %) for the global average DDLT ($\varepsilon\left(\overline{DDLT_g}\right)$) under the DYN-SEPT versus pooling gains for the global expected sojourn time ($\varepsilon(E\left[S_g\right])$) under the SEPT for the balanced case and different values of utilization ρ and coefficient of heterogeneity c

type $i-1$ and coefficient of asymmetry $\delta > 0$, such that $\rho_i = \rho_{i-1}/\delta$. Then, ρ_i can be represented in terms of ρ_1 i.e. $\rho_i = \rho_1/\delta^{i-1}$. In other words, by setting $\delta > 1$ the order types will be ranked in decreasing order of traffic load. Furthermore, when $\delta > 1$ the highest traffic load corresponds to the type with the shortest processing time (recall also that $\tau_i = c^{i-1} \cdot \tau_1$).

Then, presenting the workload and the mean processing time of type i, ρ_i and τ_i, in terms of the coefficients of asymmetry and heterogeneity, δ and c, the pooling gain for the global expected sojourn time, $\varepsilon(E\left[S_g\right])$, under the unbalanced case can be calculated by using the expressions (9), (11), and (12). Thus, $\varepsilon(E\left[S_g\right])$ is provided by the following expression:

$$\begin{aligned}&\varepsilon\left(E\left[S_g\right]\right)\\&\quad= 1 - \frac{\rho_1 \sum_{i=1}^{n} \left(\frac{c}{\delta}\right)^{i-1}}{\sum_{i=1}^{n} \left(\delta^{i-1} - \rho_1\right)^{-1}} \cdot \sum_{i=1}^{n} \frac{(c\delta)^{1-i}}{\left(n - \rho_1 \sum_{j=1}^{i} \delta^{1-j}\right) \cdot \left(n - \rho_1 \sum_{j=1}^{i-1} \delta^{1-j}\right)}\\&\qquad - \frac{1}{n} \cdot \frac{\sum_{i=1}^{n} \delta^{1-i}}{\sum_{i=1}^{n} \left(\delta^{i-1} - \rho_1\right)^{-1}}. \end{aligned} \tag{16}$$

Similarly, the pooling gain for the expected sojourn time of order type i is given by Eq. (17).

$$\begin{aligned}&\varepsilon\left(E\left[S_i\right]\right)\\&\quad= 1 - \frac{\delta^{i-1} - \rho_1}{(c\delta)^{i-1}} \cdot \left(\frac{\rho_1 \cdot \sum_{j=1}^{n} \left(\frac{c}{\delta}\right)^{j-1}}{\left(n - \rho_1 \sum_{j=1}^{i} \delta^{1-j}\right) \cdot \left(n - \rho_1 \sum_{j=1}^{i-1} \delta^{1-j}\right)} + \frac{c^{i-1}}{n} \right). \end{aligned} \tag{17}$$

The pooling gains $\varepsilon(E\left[S_g\right])$ and $\varepsilon(E\left[S_i\right])$ depend on c, δ, ρ_1 and n.

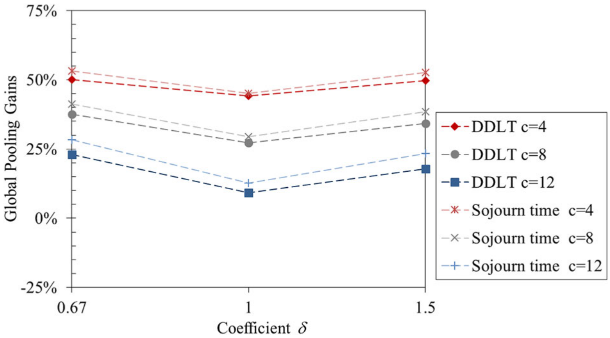


Fig. 6 Pooling gains (in %) for the global average DDLT ($\varepsilon\left(\overline{DDLT_g}\right)$) under the DYN-SEPT versus pooling gains for the global expected sojourn time ($\varepsilon(E\left[S_g\right])$) under the SEPT for the unbalanced case, and different values of coefficients of asymmetry δ and heterogeneity c

In Figure 6, once again, we contrast our simulation results for the global pooling gains for the average DDLT with the analytical results for the expected sojourn time. There is, however, a small increase in the gap between the two curves when asymmetry in workloads increase, we observe a similar correspondence between the two pooling gains as in Fig. 5.

Finally, we conclude that even-though it is a very difficult task to obtain the pooling gains in terms of the due-date performance measures; there exists a simple analytical treatment which can be used as a good indicator of these gains.

5.2 Impact of increasing the product variety on the pooling gain

Given that we could observe the high degree of similarity between the pooling gains for the expected sojourn time and the DDLT, we will base our further analysis on the expected sojourn time which is analytically tractable for the pooled single-facility configuration.

We define the ratio k as the ratio between the longest and the shortest expected processing times, $k = \tau_n/\tau_1 > 1$. In order to study how increasing product variety influences the pooling gain for the expected sojourn time, we represent coefficient of heterogeneity c in terms of the ratio k and the number of order types n, i.e., $c = \sqrt[n-1]{k}$. In this way, we can compare the pooling gains if we increase the number of order types between the slowest and the fastest order types. In a similar way, let us define ratio b which represent the ratio between the workloads of the type with the shortest and the longest expected processing times, i.e., $b = \rho_1/\rho_n$. Then, the coefficient of asymmetry δ can be represented in terms of the ratio b and n, i.e., $\delta = \sqrt[(n-1)]{b}$. Then, to study the pooling gains for different values of b and n under the same global utilization $\rho = \sum_{i=1}^{n} \rho_i/n$, ρ_1 can be replaced by $\sum_{i=1}^{n} \rho_i b / \sum_{i=1}^{n} b^{\frac{i-1}{n-1}}$ in Eq. (16). The additional constraint for the stability is $\sum_{i=1}^{n} \rho_i < \min\left\{\sum_{i-1}^{n-1} b^{\frac{i-1}{n-1}}/b, \sum_{i-1}^{n-1} b^{\frac{i-1}{n-1}}\right\}$.

In Fig. 7a, b, we illustrate the global gains when pooling different number of order types (n) for different levels of heterogeneity (k) and asymmetry (b), respectively. The two-order type system gives consistently the smallest improvement on the expected

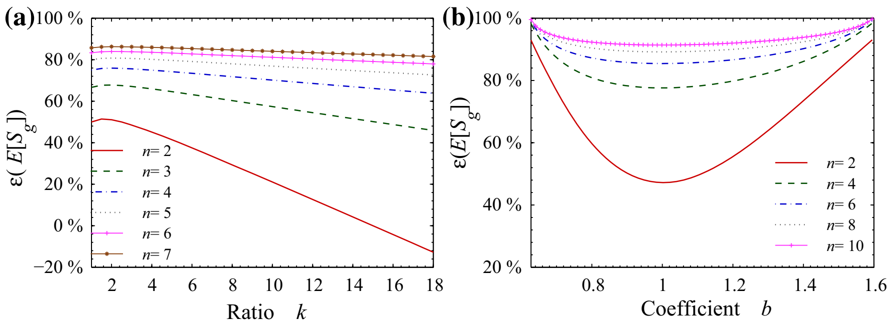


Fig. 7 Pooling gain (in %) for the global expected sojourn time when $\rho = 0.6$ for **a** the balanced case (when n and k vary and $\delta = 1$) and **b** the unbalanced case (when n and b vary and $k = 2$)

sojourn time when pooling. Another observation is that the sensitivity to the heterogeneity of the orders is also maximized in the two-order type case. Hence, the two-order type model is a lower bound on the benefit of pooling; everything else being equal, economies of scale will always increase with the number of different order types that are pooled. By adding an order type which expected processing time is between the two extreme values (minimum and maximum expected processing times), the expected sojourn time will increase in the dedicated configuration and decrease in the pooled configuration (a proof is presented in "Appendix"). Thus, the global pooling gain increases in n.

It is noteworthy that Fig. 7b is almost symmetric in terms of the global pooling gain, i.e., the pooling gains when $\rho_2 = \delta\rho_1$ and $\rho_1 = \delta\rho_2$ are similar. A similar tendency is observed when the average DDLT is considered (see Table 2) where the difference in pooling gains when $\rho_2 = \delta\rho_1$ and $\rho_1 = \delta\rho_2$ decreases as c decreases. This is because the dedicated configuration has the shortest expected sojourn time when the traffic loads are balanced, while the pooled configuration is less sensitive to unequal traffic loads. We performed additional simulation runs (unreported in this study) that show an increase in pooling gains with the number of different order types in the case of the pooled configuration which contains one parallel machine per order type.

6 Conclusions and implications for future research

Reaping economies of scale is a tempting idea for companies in their continuous struggle to increase their competitiveness. Pooling operations with uncertain demand can be an important source of economies of scale. Indeed, in addition to the direct scale effect that usually brings cost savings, the pooling of demand uncertainty has been widely shown to bring important economies of scale in terms of waiting time and inventory. Nevertheless, the lean management philosophy warns us that any increase in variability should be restrained. For firms pondering over the possibility of pooling resources (internally or externally) of heterogeneous product lines, it is thus necessary to get a deeper understanding about the interplay between economies of scale and variability.

In the article, we shed a new light on this dilemma by taking into account the fact that in practice most firms operate with due dates. For the global planning of a supply chain, reliable due dates are imperative. The goal is then to keep those due dates as short as possible in order to increase the attractiveness of the firm. The good news is that we could show how the additional flexibility provided by the due-date setting and scheduling mechanism can be used to overcome the variability introduced by the heterogeneity of the pooled products. This is not as obvious as might seem because, to the best of our knowledge, the literature so far studied pooling queues in service sectors, and consequently reached very different conclusions.

We showed that heterogeneity does not have a significant negative impact on the due-date performance if the system is adequately managed. The conclusion is even stronger when the utilization is high—the case where pooling should be particularly attractive—this means the economies of scale will, in general, largely dominate the heterogeneity impact. These insights represent a complete reversal from the conclusion reached so far for FCFS queues, that is, in the presence of heterogeneity, pooling should not be pursued especially when the utilization is high.

Another conclusion is that in terms of workload, heterogeneity brings benefits. Indeed, pooling products with different workloads will in general bring benefits that increase with the workload imbalance between the products. Finally, one might say that when pooling manufacturing operations with due dates, opposites should attract each other.

Finally, we discuss our main assumptions, limitations, and scope for future research.

In our work, we considered a simple situation with fully pooled resources as we focused on the effect of heterogeneity and product variety. Although, we modelled a pooled system as an n-times faster single-server queue, we also showed that our managerial insights can be generalized for a multi-server pooled configuration. An interesting research avenue may be to consider different flexibility configurations (see e.g. Jordan and Graves 1995; Gurumurth and Benjaafar 2004; Bassamboo et al. 2010a, b). However, scheduling rules for such configurations are much more complex. The question is to find the extent to which it is still possible to overcome the heterogeneity hurdle when operations are partially flexible.

We focused on pooling manufacturing system with two product types. The two product type system seems to be the most relevant case from a practical point of view because the coordination cost for pooling will increase very quickly with the number of product types. We showed analytically that any other multiple product-type pooled system will achieve larger benefits of pooling.

We considered the most prevalent processing time distribution in stochastic scheduling research, the exponential distribution, as it generates useful insights (see “Appendix”, Baker and Trietsch 2009). Increasing the variability of individual facilities measured by the coefficient of variation of processing times, pooling gains will increase as already studied in the literature on queueing theory (see e.g., Hopp and Spearman 2008) . We also performed additional tests assuming a log-normal distribution, one of the most prevalent distributions in practice. In the tests, we considered two product types with identical coefficients of variation. Coefficients of variations are determined in such a way to explore low, medium, and high variability in processing times. Our results, although, not presented in the paper, revealed that the difference in pooling gains between different levels of variability decreases with the average utilization.

We did not incorporate switchover times in our study in order to focus specifically on the effect of due dates. If a switchover is incurred between orders of the two firms this will of course reduce the pooling gain. On the other hand if the families of products needing a switchover span across the firms, it is not clear how our conclusion would be affected.

The main goal of the paper was to investigate pooling gains for quoted due dates under various operating conditions. In our study the manufacturer is setting the due-dates, this is the likely situation when orders are frequent. Although in such environments it is possible that the due date of an order is negotiated with the buyer, this is rather exceptional and should not affect the overall performance level. For some products due date negotiation for each order is the norm, it would be interesting to study further benefits of pooling in such environments. The adaptation of the models proposed in Altendorfer (2015) or Ivanescu et al. (2002), Ivănescu et al. (2006), Chevalier et al. (2015) would constitute an interesting research avenue.

Compliance with ethical standards

Funding Research that the authors conducted at Center for Operations Research and Econometrics, Université catholique de Louvain was partially supported by the ARC project Managing Shared Resources in Supply Chains [ARC 08-13-008].

Conflicts of interest The authors declare that they have no conflict of interest.

Appendix

For the comparison of the production configurations with different number of order types, we add an additional superscript n to the appropriate notation. In order to show that $\varepsilon(E\left[S_g^n\right]) \leq \varepsilon(E\left[S_g^{n+1}\right])$, it is sufficient that $E\left[S_g^{d,n}\right] \leq E\left[S_g^{d,n+1}\right]$ and $E\left[S_g^{p,n}\right] \geq E\left[S_g^{p,n+1}\right]$.

The inequality $E\left[S_g^{d,n}\right] \leq E\left[S_g^{d,n+1}\right]$ can be written as follows:

$$\frac{n}{n+1} \leq \frac{\sum_{i=1}^{n} k^{\frac{i-1}{n-1}}}{\sum_{i=1}^{n+1} k^{\frac{i-1}{n}}}. \tag{18}$$

The right side of inequality (18) is an increasing function of k and it achieves the minimum (i.e. $\frac{n}{n+1}$) when $k = 1$. Then, the inequality is satisfied for any value of $k \geq 1$.

Based on Eqs. (11) and (12), the global expected sojourn time for system p with n order types can be expressed as:

$$E\left[S_g^{p,n}\right] = \rho\tau_1 k \sum_{i=1}^{n} \left[k^{\frac{i-1}{n-1}}(n - i\rho)(n - (i-1)\rho)\right]^{-1} + \tau_1 k \left[\sum_{i=1}^{n} k^{\frac{i-1}{n-1}}\right]^{-1}. \quad (19)$$

Equation (19) shows that increasing the number of order types between the order types with the minimum and maximum expected processing times will decrease the variance of the processing times of the pooled facility and will consequently lead to a decrease in the global expected sojourn time. Therefore, $E\left[S_g^{p,n}\right]$ is decreasing in n.

References

Altendorfer K (2015) Influence of lot size and planned lead time on service level and inventory for a single-stage production system with advance demand information and random required lead times. Int J Prod Econ 170:478–488

Altendorfer K, Jodlbauer H (2011) An analytical model for service level and tardiness in a single machine MTO production system. Int J Prod Res 49(7):1827–1850

Altendorfer K, Minner S (2015) Influence of order acceptance policies on optimal capacity investment with stochastic customer required lead times. Eur J Oper Res 243(2):555–565

Azaron A, Fynes B, Modarres M (2011) Due date assignment in repetitive projects. Int J Prod Econ 129(1):79–85

Baker KR, Bertrand JWM (1982) A dynamic priority rule for scheduling against due-dates. J Oper Manag 3(1):37–42

Baker KR, Trietsch D (2009) Principles of sequencing and scheduling. Wiley, Hoboken (chap Appendix A: Practical Processing Time Distributions)

Barman S (1998) The impact of priority rule combinations on lateness and tardiness. IIE Trans 30(5):495–504

Bassamboo A, Randhawa RS, Van Mieghem JA (2010a) A little flexibility is all you need: asymptotic optimality of tailored chaining and pairing in queuing systems. Working paper, Northwestern University, Evanston, IL

Bassamboo A, Randhawa RS, Van Mieghem JA (2010b) Optimal flexibility configurations in newsvendor networks: Going beyond chaining and pairing. Manag Sci 56(8):1285–1303

Baykasoğlu A, Göçken M, Unutmaz ZD (2008) New approaches to due date assignment in job shops. Eur J Oper Res 187(1):31–45

Benjaafar S (1995) Performance bounds for the effectiveness of pooling in multi-processing systems. Eur J Oper Res 87(2):375–388

Benjaafar S, Cooper WL, Kim JS (2005) On the benefits of pooling in production-inventory systems. Manag Sci 51(4):548–565

Board of Governors of the Federal Reserve System (2015) Industrial production and capacity utilization-G 17. [statistical release]. http://www.federalreserve.gov/releases/g17/Current/. Accessed 28 Sept 2015

Bookbinder JH, Noor AI (1985) Setting job-shop due-dates with service-level constraints. J Oper Res Soc 36(11):1017–1026

Buzacott JA (1996) Commonalities in reengineered business processes: models and issues. Manag Sci 42(5):768–782

Chen B, Matis TI (2013) A flexible dispatching rule for minimizing tardiness in job shop scheduling. Int J Prod Econ 141(1):360–365

Cheng TCE, Gupta MC (1989) Survey of scheduling research involving due date determination decisions. Eur J Oper Res 38(2):156–166

Chevalier P, Lamas A, Lu L, Mlinar T (2015) Revenue management for operations with urgent orders. Eur J Oper Res 240(2):476–487

Cox DR, Smith WL (1961) Queues. Methuen Co. Ltd, London

Duenyas I (1995) Single facility due date setting with multiple customer classes. Manag Sci 41(4):608–619

Duenyas I, Hopp WJ (1995) Quoting customer lead times. Manag Sci 41(1):43–57

Easton FF, Moodie DR (1999) Pricing and lead time decisions for make-to-order firms with contingent orders. Eur J Oper Res 116(2):305–318

Enns ST, Grewal CS (2010) Lot size optimisation under pooled and unpooled scenarios in batch production system. Int J Prod Res 48(20):6085–6101
Gurumurth S, Benjaafar S (2004) Modeling and analysis of flexible queueing systems. Nav Res Logist 51(5):755–782
Hella Australia Pty Ltd (2007) Behr Hella service-bringing the world's best thermal management products to the Australian automotive aftermarket. [press release]. http://www.hella.com/produktion/HellaAU/WebSite/Channels/Company/Press/National_Press/Company_News/BehrHellaPR.jsp. Accessed 15 Nov 2012
Hopp WJ, Spearman M (2008) Factory physiscs. McGraw-Hill, New-York
Hopp WJ, Sturgis MLR (2000) Quoting manufacturing due dates subject to a service level constraint. IIE Trans 32(9):771–784
Hübl A, Jodlbauer H, Altendorfer K (2013) Influence of dispatching rules on average production lead time for multi-stage production systems. Int J Prod Econ 144(2):479–484
Ivanescu CV, Fransoo JC, Bertrand JWM (2002) Makespan estimation and order acceptance in batch process industries when processing times are uncertain. OR Spectr 24(4):467–495
Ivănescu VC, Fransoo JC, Bertrand JWM (2006) A hybrid policy for order acceptance in batch process industries. OR Spectr 28(2):199–222
Iyer AV, Jain A (2004) Modeling the impact of merging capacity in production-inventory systems. Manag Sci 50(8):1082–1094
Jordan WC, Graves SC (1995) Principles on the benefits of manufacturing process flexibility. Manag Sci 41(4):577–594
Kaman C, Savasaneril S, Serin Y (2013) Production and lead time quotation under imperfect shop floor information. Int J Prod Econ 144(2):422–431
Kaminsky P, Lee ZH (2008) Effective on-line algorithms for reliable due date quotation and large-scale scheduling. J Sched 11(3):187–204
Kapuscinski R, Tayur S (2007) Reliable due-date setting in a capacitated MTO system with two customer classes. Oper Res 5(1):56–74
Kelly FP, Massoulié L, Walton NS (2009) Resource pooling in congested networks: proportional fairness and product form. Queue Syst Theory Appl 63(1–4):165–194
Keskinocak P, Tayur S (2004) Due-date management policies. In: Simchi-Levi D, Wu SD, Shen ZJM (eds) Handbook of quantitative supply chain analysis: modeling in the e-business era. International series in operations research and management science. Kluwer Academic Publishers, Norwell, pp 485–553
Keskinocak P, Ravi R, Tayur S (2001) Scheduling and reliable lead-time quotation for orders with availability intervals and leadtime sensitive revenues. Manag Sci 47(2):264–279
Law AM (2007) Simulation modeling and analysis. McGraw Hill, Boston
Li L, Lee YS (1994) Pricing and delivery-time performance in a competitive environment. Manag Sci 40(5):633–646
Li M, Yang F, Wan H, Fowler JW (2015) Simulation-based experimental design and statistical modeling for lead time quotation. J Manuf Syst 37:362–374
Li X, Wang J, Sawhney R (2012) Reinforcement learning for joint pricing, lead-time and scheduling decisions in make-to-order systems. Eur J Oper Res 221(1):99–109
Mandelbaum A, Reiman MI (1998) On pooling in queueing networks. Manag Sci 44(7):971–981
Rothkopf MH, Rech P (1987) Perspectives on queues: combining queues is not always beneficial. Oper Res 35(6):906–909
Slotnik SA (2014) Lead-time quotation when customers are sensitive to reputation. Int J Prod Res 52(3):713–726
Smith DR, Whitt W (1981) Resource sharing for efficiency in traffic systems. Bell Syst Tech J 60(1):39–55
Sridharan V, Li X (2008) Improving delivery reliability by a new due-date setting rule. Eur J Oper Res 186(3):1201–1211
Stidham S Jr (1970) On the optimality of single-server queueing systems. Oper Res 18(4):708–732
Swinnen J, Van Herck K (2011) Poland, Bulgaria and Romania: social impact of discount and organized food retail formats on remote regions. Report series 16, FAO Investment Centre/EBRD Cooperation Programme, Licos
Tekin E, Hopp WJ, Van Oyen MP (2009) Pooling strategies for call center agent cross-training. IIE Trans 41(6):546–561
van Dijk NM, van der Sluis E (2008) To pool or not to pool in call centers. Prod Oper Manag 17(3):296–305

Veral EA (2001) Computer simulation of due-date setting in multi-machine job shops. Comput Ind Eng 41(1):77–94

Vinod V, Sridharan R (2011) Simulation modeling and analysis of due-date assignment methods and scheduling decision rules in a dynamic job shop production system. Int J Prod Econ 129(1):127–146

Watanapa B, Techanitisawad A (1999) Simultaneous price and due date settings for multiple customer classes. Eur J Oper Res 166(2):351–368

Wein LM (1991) Due-date setting and priority sequencing in a multiclass M/G/1 queue. Manag Sci 37(7):834–850

Wein LM, Chevalier PB (1992) A broader view of the job-shop scheduling problem. Manag Sci 38(7):1018–1033

Yu DZ, Zhao X, Sun D (2013) Optimal pricing and capacity investment for delay-sensitive demand. IEEE Trans Eng Manag 60(1):124–136

Zhang ZG, Kim I, Springer M, Cai GG, Yu Y (2013) Dynamic pooling of make-to-stock and make-to-order operations. Int J Prod Econ 144(1):44–56