International Journal of Production Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tprs20

Scheduling policies analysis for matching operations in Bernoulli selective assembly lines

Xiaoxiao Shen & Na Li

To cite this article: Xiaoxiao Shen & Na Li (2022) Scheduling policies analysis for matching operations in Bernoulli selective assembly lines, International Journal of Production Research, 60:13, 3965-3988, DOI: 10.1080/00207543.2021.1939903

To link to this article: https://doi.org/10.1080/00207543.2021.1939903

Published online: 21 Jun 2021.

INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH
2022, VOL. 60, NO. 13, 3965–3988
https://doi.org/10.1080/00207543.2021.1939903

Taylor & Francis
Taylor & Francis Group

Scheduling policies analysis for matching operations in Bernoulli selective assembly lines

Xiaoxiao Shen and Na Li

Department of Industrial Engineering & Management, Shanghai Jiao Tong University, Shanghai, People's Republic of China

ABSTRACT

In a selective assembly system, mismatched products can pass inspections due to the flexibility of product quality grades. However, they will be sold at discounted prices leading to a revenue decline. Hence, it is critical to design an appropriate scheduling policy for better matching to maximise the system quality-related revenue. In this paper, we propose a *Waiting for Closest Quality Matching Policy* (*WCQMP*), which allows postponing the assembly process within the waiting threshold. And once the postpone is finished, the closest quality parts will be selected to match. The other two policies, *Random Matching Policy* (*RMP*) and *Closest Quality Matching Policy* (*CQMP*), are also proposed as comparisons. We construct Markov chain models for small systems and develop approximation methodologies for larger systems to analyze the performance under the policies. Comparisons of different scheduling policies and the performance analysis of *WCQMP* are carried out in numerical studies. Our findings indicate that nearly in all the systems, *WCQMP*, *CQMP* performs better than *RMP*. And when system and policy parameters are properly designed, *WCQMP* is more superior by improving assembly quality without overly sacrificing system throughput, thereby increasing quality-related revenue. Managerial insights are also provided for industrial practitioners to apply *WCQMP* more appropriately.

ARTICLE HISTORY
Received 2 December 2020
Accepted 2 June 2021

KEYWORDS
Bernoulli selective assembly system; matching operations; scheduling policy; quality-related revenue; waiting threshold; production modelling

1. Introduction

Assembly quality dominates the company's ability to maximise its return on investment. Manufacturing high-quality products is a key issue for a company's success in the competitive market. Two or more components are usually assembled to make a complex final product. Because of random variations in the manufacturing process, it is impossible to produce the components with the same characteristics even if they are generated on the same production line (Ju, Li, and Deng 2017). Hence, selective assembly is employed to assemble completely interchangeable within corresponding matching groups and achieve precision assemblies from relatively low precision components (Liu et al. 2019). It has been widely applied in many manufacturing industries (i.e. machinery manufacturing, automobile, electronic industry).

One example for selective assembly systems may come from the semiconductor production line (Li et al. 2012). Semiconductor manufacturing incorporates front-end processing and back-end processing, which are usually implemented separately in different factories. Semiconductor Assembly and Test Manufacturing belong to Back-end processing. In the assembly process, dies are assembled into chipsets. In the test process, the chipsets are inspected to determine whether they can execute desired functions and then categorised into different bins based on their performance levels and product types. The chipsets in different bins are then can be assembled with other parts with different bin levels together to form packaged integrated circuit (IC) products that have different performance levels. Those IC products will be applied in high-performance computing servers, personal computers, or some simple electronic devices based on the levels of performance caused by combining different bin-level parts.

The policy for matching different bins in the assembly process dynamically is essential. Firstly, the revenue generated by different levels of final products will have a great difference, which makes that an effective policy can take significant revenue to the line. Secondly, rather than storing a large volume of Works-In-Process (WIP) in other industries, due to the high value of semiconductor

CONTACT Na Li na-li03@sjtu.edu.cn Department of Industrial Engineering & Management, School of Mechanical Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai, People's Republic of China

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

Supplemental data for this article can be accessed here. https://doi.org/10.1080/00207543.2021.1939903

products, the semiconductor production line usually has finite and relatively small buffers. It is helpful for the system to keep a relatively shorter cycle time to respond to the marketing changes and provide a quick response to customers' demands. However, with small inventories, the static matching policy (see, for example, papers by Kannan, Jayabalan, and Jeevanantham 2003; Lanza, Haefner, and Kraemer 2015; Liu et al. 2019) that needs to generate a large quantity of WIP to do deterministic optimal matching becomes unacceptable, and the dynamic control policy is plausible.

In one of the semiconductor factories we investigated, managers have struggled with finding proper dynamic matching policies. Some existing policies, such as selecting the closest quality group to match, have been applied. Improvement is achieved. But they still urgently need to develop new ways to further reduce cycle time and improve product quality, thereby improving the economic benefits of the company due to intense competition in the market. The idea that whether we can put off the assembly process for a while to improve the assembly efficiency by matching the mating components precisely is discussed. Motivated by this prospect, we began to study new types of control policies.

Actually, the discussion of the new policies is also considerable for other production lines. Another example could be in battery assembly manufacturing for electric vehicles (Ju et al. 2014). Multiple battery cells need to be welded within a tight product envelope to make batteries. Cells are partitioned into groups (or bins) according to their dimensions and stacked into sections (or modules), and then connected through welding or mechanical joints. The assembly clearance requirements for cell dimensions within a single section (or module) are various for batteries with different quality levels. In practice, batteries of flexible quality grades are employed in diverse electric vehicles with prices of discrepancy. High-quality batteries whose battery cells are perfectly aligned will be used to manufacture premium electric vehicles with a longer continue voyage course. Batteries with inaccurate matches will be used to produce lower-grade electric vehicles with a less long continue voyage course. In such a system, the selective assembly can guarantee that the cells from the same group are assembled. However, whether the selective technique will genuinely take benefits to the system is mainly determined by a proper design of the scheduling policy. Different designs of scheduling policy for matching operations may generate different quantities of various quality-level batteries, thus taking diverse quality-related revenue to the electric vehicle manufacturers.

In this work, we propose a new type of policy named *Waiting for Closest Quality Matching Policy* (*WCQMP*) for a dynamic and uncertain production environment. This policy allows putting off the assembly process within the waiting threshold so that the desirable mating part will arrive in place with a higher probability. And thus, we have more chances to obtain high-quality assemblies. When the waiting period is over, the mating part with the closest quality to the current main part is selected for assembling. To our best knowledge, there is no research to realise this idea that can address the trade-offs between productivity and quality in selective assembly systems. The other two policies, *Random Matching Policy* (*RMP*) and *Closest Quality Matching Policy* (*CQMP*) are also proposed as comparisons. *RMP* prioritises mating parts according to their random arrival sequence. *CQMP* selects the part belonging to the closest quality grade group to match if the corresponding matching part is not available. These two comparable policies are the only dynamic policies proposed in the literature (Ju, Li, and Deng 2017) for selective production lines. Specifically, we first construct mathematical models for selective assembly systems with Bernoulli machines and finite buffers for the three policies. Then, we develop the Markov chain approach and approximation methods for small systems and larger systems under three policies, respectively, to evaluate the selective assembly system quality-related revenue. And we conduct numerical studies to investigate how pivotal system and policy parameters impact the performance of the proposed policies.

The main contributions of the work include the following aspects: (1) we propose a novel scheduling policy named *WCQMP* for matching operations, which considers potential waiting in a dynamic and uncertain selective assembly environment with unreliable Bernoulli machines and finite buffers. This innovative policy has not been investigated in other research. By comparisons with the other two policies, *RMP* and *CQMP*, *WCQMP* can show its superiority nearly in all systems when a trade-off between the quality improvement and the throughput impediment is achieved. (2) we develop exact and approximation performance measurement methodologies under all policies to evaluate the system quality-related revenue. Especially, we contribute by providing approximation methods integrating decomposition and aggregation ideas for selective assembly systems implementing the scheduling policy *WCQMP*. (3) By numerical experiments, we propose some interesting insights that can provide industrial practitioners some guidelines about employing the proposed policy *WCQMP*. On the one hand, we reveal the features for the selective assembly lines where the proposed policy can help the systems to improve more of their revenue. For example, when a real selective assembly system with identical buffers and nonidentical machine reliability follows the machine reliability pattern where the main line and final assembly line machines reliability are always equal and smaller than

that of mating line machines, it could be a better choice to apply our policy *WCQMP* for getting higher revenue. On the other hand, we suggest the adjustment ways to the policy when a system's states are changed. For example, managers could set the waiting threshold with a larger value to accommodate the improvement of the mating line or final assembly line efficiency. If the main line is aging during the operation, the waiting threshold should be increased to obtain maximal revenue.

The remainder of this paper is organised as follows. In Section 2, related literature is reviewed to find out research gaps and motivation of this work. In Section 3, we formulate the problem with some assumptions and introduce the selection matching policies. The performance evaluation methods for different scales of selective assembly systems are presented in Section 4. Section 5 conducts numerical experiments to verify the convergence and accuracy of proposed methods. The results of the policies performance analysis are also summarised in this section. Finally, Section 6 gives conclusions. Some details of the performance evaluation methods and numerical examples are provided in the Appendices.

2. Literature review

Research on selective assembly systems has mainly focused on two streams in general. Literature concerning the first stream mainly deals with choosing classification criteria to partition the parts to be mated appropriately and optimising the partitioning. The second stream focuses on matching operation optimisation to achieve some goals, such as minimising assembly variation, mismatches, or maximising throughput.

The first stream has been extensively studied since the paper (Mansoor 1961) was published. By establishing tolerance specifications related to natural process tolerances, they develop the design and manufacturing plans to determine the component classification and corresponding production quantities. Kannan and Jayabalan (2001) present a two-stage grouping method for a complex ball bearing assembly with three mating parts. Chan and Linn (1998) incorporate the same concept with another idea of skipping certain portions of components to form more mating groups. Liu and Liu (2017) introduce a method of determining the number of groups and the range of dimensional tolerances for each component. Mease, Nair, and Sudjianto (2004) describe the selective assembly problem as a statistical formulation and develop optimal binning strategies under several loss functions and distributional assumptions.

The second stream of works for matching operation issues in the selective assembly has attained growing attention within the academic community. Different methods have been proposed to release the trouble taken by multiple quality levels in the selective production line. For example, the latest paper (Clottey and Benton 2020) studies the use of inexpensive intermediary components to ensure that all mating components are matched with acceptable clearance and to minimise shortage and surplus component costs. Liu et al. (2013) develop a comprehensive quality control model to improve the matchable degree of multiple components' selective assembly process. Colledani, Ebrahimi, and Tolio (2014) present an integrated quality and production logistics model to properly design selective and adaptive assembly systems, significantly reducing the defective assemblies.

A substantial amount of research effort has been devoted to static scheduling policy, which addresses the matching optimisation problem in a deterministic way. They assume that there are lots of WIP parts to be assembled and try to match batch parts at a given time. Some optimisation goals, such as minimising the assembly variation and mismatch, are achieved by determining the optimal combination of components with widely used genetic algorithms in certain work, such as Kannan, Jayabalan, and Jeevanantham (2003), Jeevanantham, Chaitanya, and Rajeshkannan (2019), and Lanza, Haefner, and Kraemer (2015). And Liu et al. (2019) propose a discrete fireworks optimisation algorithm to maximise assembly efficiency. Some non-metaheuristics algorithms are also proposed to optimise the combination of groups. Tan and Wu (2012) consider the generalised concept of selective assembly as two versions and develop related simulation methods to solve them. Siva Rezaei Aderiani et al. (2019) propose a non-metaheuristics algorithm for sheet metal assemblies' problems. This type of static policy leads to rather high inventory levels and a long cycle time certainly. Along with the popularisation of the lean manufacturing concept, production systems prefer to keep a relatively low WIP and short production cycle, which can be more flexible to adapt to customers' demand changes. In this context, the parts with different levels will remain in the buffer for a short time, which makes that the WIP in buffers captured within any step length of the optimisation period is limited. Therefore, static optimally selection of bins in batches becomes unacceptable. Other works began to study the dynamic scheduling policy to optimally match finite and constantly changing bins in a stochastic way. Existing literature suggests that there are limitations to considering selective assembly systems in such a stochastic manner. Ju, Li, and Deng (2017) propose a selection policy for a two-component selective assembly system with unreliable machines and finite buffers. Ju, Li, and Deng (2017) is the most closely related published work to ours. The policy they proposed is to choose the perfect matching parts when available. If no

part of the corresponding group is available, the part belonging to the highest quality group currently available is selected. Our work differs from theirs in the sense that we propose the idea that when a good matching is not available, waiting for a moment may achieve better matching with a tolerable sacrifice of the total throughput. We compare our method with *RP* and *CQMP*, which are proposed in the previous study by Ju, Li, and Deng (2017). And it can improve the system revenue when making a trade-off between the product quality improvement and the throughput impediment, which is bound to have important theoretical and practical significance.

In the field of online machine scheduling, some papers have considered the waiting strategy based on a rolling horizon periodic approach. The set of jobs to be scheduled in the waiting window enters the scheduling process window based on their waiting thresholds determined by the release dates. Suwa (2007) proposes a new online scheduling policy considering the time of rescheduling when the cumulative delays reach the waiting threshold. Zhang (2001) takes waiting into consideration when developing online scheduling algorithms for NP-hard problems. Numerous studies consider the problem of scheduling a set of jobs subject to release dates to achieve some optimisation objectives. Koskinen et al. (2020) investigate the scheduling of Printed Circuit Board (PCB) assembly jobs to minimise total job tardy times based on the predefined release and due dates. Yuan, Ng, and Cheng (2015) and Dover and Shabtay (2016) study two-agent scheduling on a single machine with release dates. Yuan, Ng, and Cheng (2015) also take preemption into consideration to minimise the maximum lateness. Vélez-Gallego, Maya, and Montoya-Torres (2016) and He et al. (2016) deal with a single-machine scheduling problem with release dates. The difference is that sequence-dependent setup times are also modelled in the former study to minimise the maximum makespan, while He et al. (2016) consider rejections additionally to minimise both the makespan and the total rejection cost.

Performance analysis of production systems with unreliable machines and finite buffers has been investigated far more maturely than in selective assembly, which lays the foundations for evaluating selective assembly systems. Analytical methods are extensively introduced for production systems. One of the most noteworthy works was the monograph done by Li and Meerkov (2009), which explicitly and systematically studies steady-state performance evaluation, system properties, and the improvement of various production lines. For Bernoulli serial production lines, Yan et al. (2021) propose an improved aggregation method by extending the traditional two-machine aggregation building blocks to general multimachine ones. Wang et al. (2019) deal with Bernoulli serial lines with batching machines and finite buffers and develop analytical methods for system performance and properties analysis. Existing literature, such as Jia et al. (2016) and Ching, Meerkov, and Zhang (2008), has focused on the analytical approximation approach on evaluating general assembly systems with unreliable machines and finite buffers. Zhang et al. (2021) evaluate manufacturing systems with both manual operations and collaborative robot assembly by introducing a unified model of productivity and ergonomic performances. Wang et al. (2018) investigate an approximate decomposition method for evaluating non-homogeneous assembly systems with multiple failure modes, finite buffers, and a fixed assembly proportion. Analytical methods for selective assembly system evaluation are rare except for the two-level decomposition approaches proposed by Colledani, Ebrahimi, and Tolio (2014) and Ju, Li, and Deng (2017).

3. Problem description and formulation

In this section, we consider a two-component selective assembly system. As illustrated in Figure 1, the circles are the machines, and the rectangles represent the buffers. The assembly system consists of three machines, m_1, m_2, and m_0, and two buffers, b_1 and b_2. Machine m_1 and buffer b_1 form the main line. The mating line is made up of machine m_2 and buffer b_2. Machine m_0 is the assembly machine in the final assembly line.

The concrete system description is provided as follows. We consider a typical automatic selection assembly line that can collect the group label information of the buffer parts. At machine m_1, the main parts are processed and categorised into groups 1 to G with the probabilities q_1 to q_G based on their quality levels. Independently, machine m_2 produces mating parts. Analogously, the mating parts are partitioned into groups 1 to G with the probabilities g_1 to g_G. The quality order of groups 1 to G is: $1 > 2 > 3 > \cdots > G$, which means that group 1 represents the highest quality, and group G is the lowest. In this way,

$$\sum_{i=1}^{G} q_i = \sum_{i=1}^{G} g_i = 1.$$

Only mating parts within the mating line are selected according to the main part. Selective assembly is often used in systems with high assembly precision requirements. Three possible scheduling policies for matching operations are proposed as follows to guarantee assembly quality and improve quality-related revenue further.

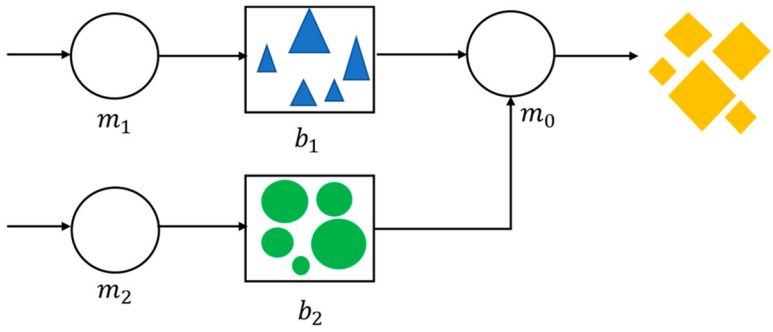

Figure 1. Selective assembly system.

- Policy I (***RP, Random Matching Policy***): select the mating part according to their random arrival sequence to match.
- Policy II (***CQMP, Closest Quality Matching Policy***): select the mating part belonging to the closest quality grade group to match.
- Policy III (***WCQMP, Waiting for Closest Quality Matching Policy***): wait for the coming of the desirable closest quality mating part to match.

The latter two kinds of policies are interpreted in detail below. We denote the current main part with quality grade i as A_i, and the mating part with quality grade i as B_i, $i = 1, 2, \ldots, G$. For Policy II, mating part B_i is selected to assemble if available. If not, the mating part with the closest quality grade to A_i is selected. High-quality mating parts B_{i-n} and low-quality mating parts B_{i+n} will be selected with equal probability. That is, the matching parts are selected for assembly according to the following priorities:

If G is odd,

$$\begin{cases} B_{i-1} = B_{i+1} > B_{i-2} = B_{i+2} > \ldots > B_1 = B_{2i+1} > B_{2i+2} > B_{2i+3} > \ldots > B_G, & \text{if } i < (G+1)/2, \\ B_{i-1} = B_{i+1} > B_{i-2} = B_{i+2} > B_{i-3} = B_{i+3} > \ldots > B_1 = B_G, & \text{if } i = (G+1)/2, \\ B_{i-1} = B_{i+1} > B_{i-2} = B_{i+2} > \ldots > B_{2i-G} = B_G > B_{2i-G-1} > B_{2i-G-2} > \ldots > B_1, & \text{if } i > (G+1)/2. \end{cases}$$

If G is even,

$$\begin{cases} B_{i-1} = B_{i+1} > B_{i-2} = B_{i+2} > \ldots > B_1 = B_{2i+1} > B_{2i+2} > B_{2i+3} > \ldots > B_G, & \text{if } i \le G/2, \\ B_{i-1} = B_{i+1} > B_{i-2} = B_{i+2} > \ldots > B_{2i-G} = B_G > B_{2i-G-1} > B_{2i-G-2} > \ldots > B_1, & \text{if } i > G/2. \end{cases}$$

For Policy III, from the time when a main part becomes the bottom (the 1st) part in the buffer b_1, the maximum allowable cycle before starting the assembly process is constrained by buffer b_2 occupancy status and the waiting threshold (WT) that is no more than buffer b_2 capacity. Note that the current main part is A_i. The mating part B_i is selected to assemble if available. The buffer b_2 is not empty, and the current occupancy is less than the waiting threshold, but there is no matching mating part B_i, then the current main part is allowed to wait. Once B_i is generated within the waiting threshold, the assembly is performed. The closest quality mating part is selected to assemble if there is still no available B_i while the buffer b_2 occupancy has reached the waiting threshold. The execution procedure is illustrated in Figure 2. And when the waiting threshold of Policy III takes the value of one, that is, not waiting, it will have the same effect as Policy II.

To further characterise the system, we introduce the following assumptions relevant to the machines, the buffers, and their interactions.

- All machines are Bernoulli machines. The status of the machine is determined at the beginning of each time slot. In each cycle, machine m_i ($i = 0,1,2$) is up with probability p_i and down with probability $1 - p_i$.
- All machines have a constant and identical processing time that is defined as the machine's cycle time. The time is slotted with the slot duration equal to the cycle time of the machines. For simplicity, we standardise the cycle time to be one- time unit without losing generality.
- The buffer capacities are N_i, $0 < N_i < \infty$, both being finite. The status of the buffer is determined at the end of each time slot.
- Machine $m_i (i = 1, 2)$ is blocked when m_i is up, and buffer b_i is full, and the mating machine m_0 does not take a part from b_i at the beginning of a time slot. Machine m_0 is never blocked.
- The assembly machine m_0 is starved when either b_1, or b_2, or both are empty. Machine m_0 is also starved when

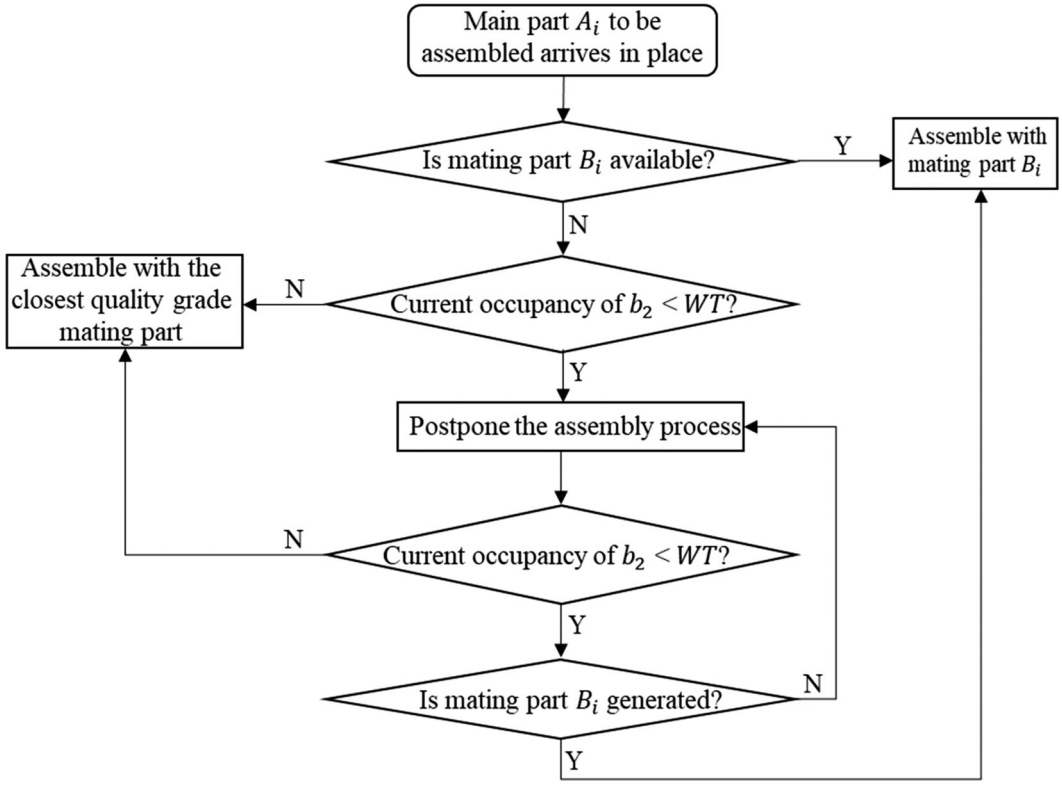

Figure 2. The operation procedure of Policy III.

buffer b_2 current occupancy is less than the specified waiting threshold and not empty, but there is no available matching mating part to assemble. It is assumed that machines m_1 and m_2 are never starved.

- The produced parts will not be scrapped throughout the process.

Remark 1: We assume that the closer the quality grades of the assembled main and mating parts, the higher the finished products' quality. Many assembly lines in factories can reflect this. Taking the assembly of power steering sleeves and shafts as an example, the clearance between the shaft and sleeve must fall within the acceptable tolerance range. Therefore, a shaft with a larger-than-average outer diameter and a sleeve larger-than-average inner diameter would be matched together (Coullard, Gamble, and Jones 1998). That is, a sleeve and shaft with a closer quality level can be assembled to make an acceptable assembly. Furthermore, we allow the main parts to wait within the waiting threshold to manufacture more matched assemblies.

Remark 2: The assembly system in many factories consists of the main line and mating line. The mating parts in the mating line need to be assembled onto the main parts in the main line. For example, in the automotive final assembly workshop, the wiring harnesses, pipes, and central control equipment in the car interiors line will be assembled to the car body. Generally, the mating subcomponents are small in size and large in quantity. So, it is easier and practical to change the mating subcomponents' sequence. Therefore, we assume that only the mating part can be selected based on the main part.

Remark 3: In this paper, we use the Bernoulli machine reliability model. Such a model is practical, especially for describing assembly systems where the machine's downtime is typically very short and comparable to the cycle time. There have been many studies on the successful application of Bernoulli models in manufacturing systems, such as Feng et al. (2018), Lee, Li, and Horst (2018a), Su et al. (2017), etc. We first apply the relatively representative but straightforward Bernoulli machine model to our study. The analysis will be extended to more complex reliability models, such as geometric, exponential machine reliability models in future work.

In such a selective assembly system, the main performance measurement index is the expected total revenue (TR) of the finished assemblies in a cycle at a steady state, which is defined as follows:

$$TR = \sum_{i=0}^{G-1} Y_i \cdot PR_i, \tag{1}$$

where PR_i represents the PR of category i product assembled with components with a gap of i quality grades,

and Y_i represents the unit price of category i product. We assume the matched product's unit price $Y_0 = 1$. By introducing a discount factor a, based on category i product's unit price, category $i+1$ product's unit price is $100(1-a)\%(0 \le a \le 1)$ off. The goal of this research is under different system parameter configurations to compare the performance of the three scheduling policies and conduct a performance analysis of *WCQMP* to maximise the total revenue.

4. The performance evaluation methods

In this section, we develop Markov chain models for three-machine three-group with small buffers systems in Section 4.1. However, for larger systems, deriving the exact solution of steady-state probabilities is not feasible because of the dimensionality curse. Therefore, approximation methods for three- and multi-machine multi-group with large buffers systems are proposed in Sections 4.2 and 4.3, respectively, which are useful for practical systems. The exact results provided by the Markov chain approach can avoid unclear observations caused by the variations of approximation methods and help us obtain more accurate theoretical findings.

4.1. Markov chain models for three-machine three-group with small buffers systems

For three-machine ($M = 3$) three-group ($G = 3$) with small buffers assembly lines, we develop Markov chain models to obtain analytical solutions under three scheduling policies. We denote the system state as $(t_1 \dots t_{N_1} n_1 n_2 n_3)$, where $t_i \in \{0,1,2,3\}, i = 1,\dots,N_1$ representing the part group number at the ith position of buffer b_1 (when there is no part at the ith position, let $t_i = 0$). The number of group j part in the buffer b_2 is denoted by $n_j = 1,2,3$. n_j is constrained by the inequality $n_1 + n_2 + n_3 \le N_2$. When $j < i$, if $t_i = 0$, we must have $t_j = 0$. Furthermore, based on the defined state space, we can obtain the number of system states as:

$$\left(4^{N_1} - \sum_{i=1}^{N_1-1} C_{N_1-1}^{i} 3^{N_1-i}\right) \times \left(\frac{1}{2}N_2^2(N_2+1)(N_2+2) + \sum_{n=0}^{N_2}(-n^2+1)\right),$$

Then, we can derive the state transition probability matrix. We take the most straightforward case $N_1 = N_2 = 2$ as an example to illustrate the solution process for the three policies. For Policy III, there are 130 states. Only partial states and transition probabilities are spread out here, as shown in Table 1. There are 131 balance equations in total, including the total probability equation. For illustration purposes, part of the balance equations is included in Section A.1 of Appendix 1.

Table 1. Partial states and transition probabilities for Policy III.

Initial state r	New state s	Transition rate P_{sr}	Condition
(ij100)	(ij200)	$p_2 g_1$	$i = 1,2,3, j = 2,3$
(ij010)	(ij020)	$p_2 g_2$	$i = 1,2,3, j = 1,3$
(ij001)	(ij002)	$p_2 g_3$	$i = 1,2,3, j = 1,2$
(ij100)	(ij100)	$1 - p_2$	$i = 1,2,3, j = 2,3$
(0j100)	(ij200)	$p_1 q_i p_2 g_1$	$i = 1,2,3, j = 2,3$
(0j010)	(ij010)	$p_1 q_i (1-p_2)$	$i = 1,2,3, j = 1,3$
(0j001)	(0j001)	$(1-p_1)(1-p_2)$	$i = 1,2,3, j = 1,2$
(0j100)	(0j200)	$(1-p_1)p_2 g_1$	$i = 1,2,3, j = 2,3$

By solving the balance equations, we can derive the analytical expressions of steady-state probabilities. We assign $P(i_1 i_2 j_1 j_2 j_3)$ to the steady-state probability of each state $(i_1 i_2 j_1 j_2 j_3)$. The steady-state probability that there is a mating part differing from the current main part by i quality grades in buffer b_2 is denoted as P_i ($i = 0,1,2$).

For Policy III, the corresponding formulas for P_i ($i = 0,1,2$) are as follows:

$$\begin{aligned}
P_0 &= \sum_{i_1=0}^{3}\sum_{j_1=1}^{2}\sum_{\substack{j_2,j_3\in\{0,1\}\\ 0\le j_2+j_3\le 2-j_1}} P(i_1 1 j_1 j_2 j_3) \\
&\quad + \sum_{i_1=0}^{3}\sum_{j_2=1}^{2}\sum_{\substack{j_1,j_3\in\{0,1\}\\ 0\le j_1+j_3\le 2-j_2}} P(i_1 2 j_1 j_2 j_3) \\
&\quad + \sum_{i_1=0}^{3}\sum_{j_3=1}^{2}\sum_{\substack{j_1,j_2\in\{0,1\}\\ 0\le j_1+j_2\le 2-j_3}} P(i_1 3 j_1 j_2 j_3), \\
P_1 &= \sum_{i_1=0}^{3}\sum_{j_2=1}^{2} P(i_1 10 j_2 (2-j_2)) \\
&\quad + \sum_{i_1=0}^{3}\sum_{j_1=1}^{2} P(i_1 2 j_1 0 (2-j_1)) \\
&\quad + \sum_{i_1=0}^{3}\sum_{j_2=1}^{2} P(i_1 3 (2-j_2) j_2 0), \\
P_2 &= \sum_{i_1=0}^{3} P(i_1 1002) + P(i_1 3200).
\end{aligned}$$

Note that details about Markov chain models analyses for Policy II and Policy II can be found in Section A.1 of Appendix 1.

So far, we have obtained the production rates of the three types of products $PR_i (i = 0,1,2)$. In turn, we can obtain the following calculation formula for the expected

total revenue of the finished products during a cycle in steady state:

$$PR_i = p_0 P_i, i = 0, 1, 2,$$

$$TR = \sum_{i=0}^{2} Y_i \cdot PR_i. \quad (2)$$

4.2. Approximation method for three-machine multi-group with large buffers systems

For three-machine ($M = 3$) multi-group ($G \geq 3$) and large buffer capacities lines, the number of system states to be enumerated increases exponentially as groups of subcomponents and the buffer capacity increase in number. Markov chain model is intractable because of state-space explosion. Therefore, in this section, we propose decomposition-based approximation methods for three-machine multi-group with large buffers systems under three scheduling policies.

Based on the framework of the two-level decomposition method proposed in Ju, Li, and Deng (2017), we develop a decomposition method for analyzing selective assembly systems considering the potential waiting in Policy III. In addition to the situation where the upstream buffer is empty, the assembly machine will also be starved and not fetch parts when the main part is waiting within the waiting threshold. In other words, the mismatch of the two sub-components will affect the overall *PR*. Consequently, the *PR* of a selective assembly system with proper delay is different from a typical selective assembly system without considering the potential waiting.

We have revised the approximate method mentioned in the literature (Ju, Li, and Deng 2017) for the typical assembly system without waiting. Analogously, the first-level decomposition decomposes the assembly line into two overlapping serial lines (see Figure 3). The machine m_0^i represents the assembly machine that is not starved by the buffer b_i and potential waiting. In this way, the reliability of virtual machines m_0^1 and m_0^2 are represented by p_0^1 and p_0^2, respectively, which need to be modified to adapt to the potential waiting. Correspondingly, its recursive procedure will also associate with the second-level decomposition. The recursive process will be delineated together with the second-level decomposition in the following part.

According to the related analysis in the literature Ju, Li, and Deng (2017), it is scientific to decompose the sub-assembly line into two sub-lines for group u and group $C_u, u = 1, 2, \ldots, G$ (we denote all groups except group u as group C_u, group C_u = groups 1&2& ... & (u-1) & (u+1) & ... &G) (Shown in Figure 4). Machine m_2 is divided into two virtual machines, m_2^u and $m_2^{C_u}$, which manufacture group u and groups C_u mating parts separately. Buffer b_2 is decomposed into virtual buffer b_2^u for storing group u parts, and virtual buffer $b_2^{C_u}$ that contains groups C_u parts. Two virtual machines m_0^u and $m_0^{C_u}$ are used to approximate machine m_0, fetching group u and groups C_u mating parts respectively. Therefore, we need to perform G times of decomposition in total. The uth decomposition is corresponding to group u and group C_u. The approximate method needs to perform the following steps to obtain the marginal probability estimated value of group u part buffer occupancy status.

PROCEDURE 4.1

- Step 1: Initialise the parameters used to estimate the marginal probability of buffer b_2 occupancy status for each group of parts.
- Step 2: For group u and group C_u, u = 1,2, ... ,G,
  (1) (1) Update the probability of virtual assembly machines m_0^u and $m_0^{C_u}$ to be up based on the marginal probability of buffer b_2 occupancy status for each group of parts.
  (2) (2) Calculate the probability that the desirable matching part is generated within the waiting threshold, and the current main part is assembled with it.

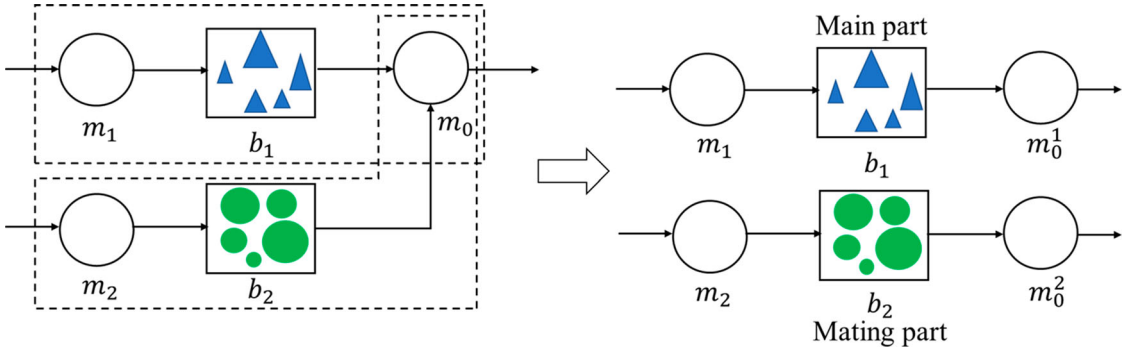

Figure 3. The first-level decomposition.

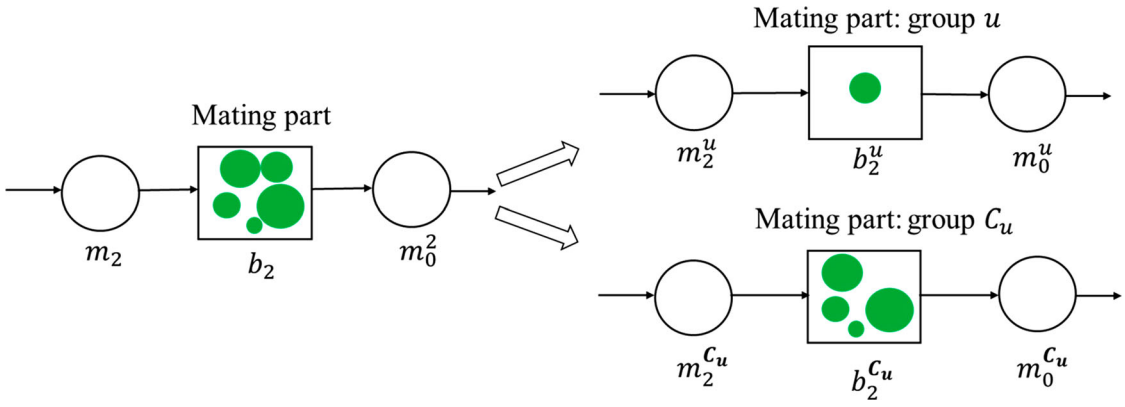

Figure 4. The second-level decomposition.

(3) (3) Evaluate the marginal probabilities of buffer occupancy status for group u parts.
(4) (4) Evaluate the probability that the waiting conditions is satisfied, that is, the desirable matching part is not available, and buffer b_2's current occupancy is less than the waiting threshold but not empty.

- Step 3: Repeat Step 2 until the convergence of the marginal probability of buffer b_2 occupancy status for each group of parts.

We take the three-group case ($G = 3$) as an example to introduce the above approximation method explicitly in the following part. When $G = 3$, the sub-assembly line is expectedly decomposed into two sub-lines for group u and groups $C_u, i = 1, 2, 3$. Group u and groups C_u mating parts are processed separately by m_2^u and $m_2^{C_u}$, which are two virtual machines obtained by decomposing m_2. Their parameters p_2^u and $p_2^{C_u}$ can be revealed as:

$$p_2^u = p_2 g_u, p_2^{C_u} = p_2(1 - g_u), u = 1, 2, 3.$$

Correspondingly, machine m_0 is decomposed into two virtual machines m_0^u and $m_0^{C_u}$, seising group u and groups C_u mating parts respectively. Their efficiency parameters are p_0^u and $p_0^{C_u}$. By the same token, p_0^u and $p_0^{C_u}$ need to be rewritten due to the potential waiting. Before making corrections, we introduce the decomposition of the buffer. Two virtual buffers b_2^u and $b_2^{C_u}$ substitute buffer b_2, which contains group u and groups C_u parts independently. To facilitate the demonstration, we first define the following symbols for the decomposed buffer state:

- $P_i^u (i = 0, 1, \ldots, N_2, u = 1, 2, 3)$: Probability of there are i group u parts in b_2.
- $P_i^{C_u} (i = 0, \ 1, \ldots, \ N_2, \ C_u \in \{(1,2), (1,3), (2,3)\})$: Probability of there are i grouC_u parts in b_2.
- $P_{j,t_{C_u}=i}^u (i = 0, \ 1, \ldots, \ N_2, \ j = 0,1, \ldots, \ N_2 - i, (u,C_u) \in \{(1,2,3), (2,1,3), (3,1,2)\})$: Probability of there are j group u parts in b_2 given i groups C_u parts occupied.
- $P_{j,t_u=i}^{C_u} (i = 0, 1, \ldots, N_2, j = 0,1, \ldots, N_2 - i, (u,C_u) \in \{(1,2,3), (2,1,3), (3,1,2)\})$: Probability of there are j groups C_u parts in b_2 given i group u parts occupied.

Following the correlated analysis and hypothesis in Ju, Li, and Deng (2017), we can obtain a closed-form formula for estimating the number of group u part in the buffer b_2. The process is as Equation (3):

$$P_{0,t_{C_u}=i}^u = Q(p_2^u, p_0^u, N_2 - i), i = 0, 1, \ldots, N_2, (u, C_u) \in \{(1, 2, 3), (2, 1, 3), (3, 1, 2)\}, \quad (3)$$

where $Q(x, y, N)$ represents the steady-state probability that the buffer is empty in a two-machine Bernoulli production line with machine parameters x and y, and a buffer capacity of N. The specific calculation formula from the monograph (Li and Meerkov 2009) is as follows:

$$Q(x, y, N) = \begin{cases} \dfrac{(1-x)(1-\alpha)}{1 - \frac{x}{y}\alpha^N}, & \text{if } x \neq y, \\ \dfrac{1-x}{N+1-x}, & \text{if } x = y, \end{cases}$$
$$\alpha(x, y) = \frac{x(1-y)}{y(1-x)}. \quad (4)$$

Again, based on the monograph (Li and Meerkov 2009), we can get the conditional probability of the group u part's buffer occupancy as:

$$P_{j,t_{C_u}=i}^u = \frac{Q(p_2^u, p_0^u, N_2 - i)}{1 - p_0^u}[\alpha(p_2^u, p_0^u)]^j, i = 0, 1, \ldots, N_2, \ j = 0, 1, \ldots, N_2 - i, (u, C_u) \in \{(1, 2, 3), (2, 1, 3), (3, 1, 2)\}. \quad (5)$$

Finally, we can derive the marginal probability of the group u part's buffer occupancy as:

$$P_j^u = \sum_{i=0}^{N_2-j} P_i^{C_u} \cdot P_{j,t_{C_u}=i}^u, j = 0, 1, \ldots, N_2, (u, C_u) \in \{(1,2,3),(2,1,3),(3,1,2)\}. \quad (6)$$

Based on the above probability calculation, we introduce the probability P_w^u that waiting conditions is satisfied, that is, the desirable matching part is not available, and buffer b_2's current occupancy is less than the waiting threshold but not empty. The expression of P_w^u is as follows:

$$P_w^u = p_0(1-x_1)q_u P_0^u \cdot \sum_{i=1}^{WT-1} P_i^{C_u}, 1 \le WT \le N_2, (u, C_u) \in \{(1,2,3),(2,1,3),(3,1,2)\}. \quad (7)$$

In the above formula, x_1 stands for the probability that buffer b_1 is empty, which can be obtained by the recursive procedure 4.2 introduced in the following part. As revealed in Lee, Zhao, et al. (2018b), the probability that buffer $b_2^{C_u}$ contains i parts, and there are j parts in buffer b_2^u when a new part is produced by m_2^u and loaded into the buffer, is formulated as $\tilde{P}_{j,t_{C_u}=i}^u$:

$$\begin{aligned}
\tilde{P}_{0,t_{C_u}=i}^u &= \frac{Q(p_2^u, p_0^u, N_2 - i)}{(1-p_2^u)(1-Q(p_0^u, p_2^u, N_2 - i)},\\
\tilde{P}_{j,t_{C_u}=i}^u &= \alpha^j \tilde{P}_{0,t_{C_u}=i}^u,\\
&i = 0, 1, \ldots, N_2, j = 1, \ldots, N_2 - i - 1,\\
&(u, C_u) \in \{(1,2,3),(2,1,3),(3,1,2)\}. \quad (8)
\end{aligned}$$

Making use of these probabilities, we can derive the steady-state probability P_{WT}^u that the desirable matching part is generated before the buffer b_2 occupancy reaches the waiting threshold, and the current main part is assembled with it as

$$P_{WT}^u = \sum_{i=1}^{WT-1} \sum_{j=i}^{WT-1} \sum_{k=i}^{j-1} \tilde{P}_{0,t_{C_u}=j}^u \cdot P_j^{C_u} (1 - \tilde{P}_{0,t_{C_u}=k}^u) P_k^{C_u}, \quad 1 \le WT \le N_2, (u, C_u) \in \{(1,2,3),(2,1,3),(3,1,2)\}. \quad (9)$$

So far, we can derive the revised parameters for virtual machines. The recursive procedure can also be modified to adapt to potential waiting to approximate marginal probabilities. To describe the recursive procedure, in addition to the symbols that have been explained above, we introduce here x_1 and x_2 to represent the probability that the buffers b_1 and b_2 are empty. Henceforth, the recursive procedure can be formally explicated as:

PROCEDURE 4.2

- Step 1: Initialisation.

$$\begin{aligned}
P_0^u(0) &= P_0^{C_u}(0) = 1, (u, C_u) \in \{(1,2,3),(2,1,3),(3,1,2)\},\\
P_j^u(0) &= P_j^{C_u}(0) = 0, j = 1, \ldots, N_2, (u, C_u) \in \{(1,2,3),(2,1,3),(3,1,2)\},\\
P_{WT}^u(0) &= 0, u = 1, 2, 3, 1 \le WT \le N_2,\\
x_1(0) &= x_2(0) = 0. \quad (10)
\end{aligned}$$

- Step 2: Calculation of the occupancy status probability of group 1 and group 2&3 parts in buffer b_2.

(1) Calculate the probability of virtual machine m_2^1, $m_2^{C_1}$ and m_0^1 to be up. Then, calculate the conditional probability that the buffer b_2 contains 0 group 1 part when a new part is processed by machine m_2^1 and loaded into the buffer given i group 2&3 in b_2.

$$\begin{aligned}
p_2^1 &= p_2 g_1,\\
p_2^{C_1} &= p_2(g_2 + g_3),\\
p_0^1(n+1) &= p_0(1 - x_1(n))(q_1 P_0^1(n)(1 - P_w^1(n)) \times (1 - P_{WT}^1(n)) + q_1(1 - P_0^1(n)) + q_2(1 - P_w^2(n)) \cdot (1 - P_{WT}^2(n)) \left(\frac{1}{2} P_0^2(n)(P_0^3(n) + 1)\right)\\
&\quad + q_3(1 - P_w^3(n))(1 - P_{WT}^3(n)) P_0^3(n) P_0^2(n)),\\
&1 \le WT \le N_2,\\
\tilde{P}_{0,t_{C_1}=i}^1(n+1) &= \frac{Q(p_2^1, p_0^1(n+1), N_2 - i)}{(1-p_2^1)(1 - Q(p_0^1(n+1), p_2^1, N_2 - i)},\\
&i = 0, 1, \ldots, N_2. \quad (11)
\end{aligned}$$

(2) Evaluate the probability that desirable matching group 1 part is generated within the waiting threshold, and the current main part is assembled with it. And further evaluate the probability that b_2 has j group 1 parts.

$$\begin{aligned}
P_{WT}^1(n+1) &= \sum_{i=1}^{WT-1} \sum_{j=i}^{WT-1} \sum_{k=i}^{j-1} \tilde{P}_{0,t_{C_1}=j}^1(n+1) \cdot P_j^{C_1}(n)(1 - \tilde{P}_{0,t_{C_1}=k}^1(n+1)) P_k^{C_1}(n), 1 \le WT \le N_2,\\
P_{0,t_{C_1}=i}^1(n+1) &= Q(p_2^1, p_0^1(n+1), N_2 - i),\\
&i = 0, 1, \ldots, N_2,\\
P_{j,t_{C_1}=i}^1(n+1) &= \frac{P_{0,t_{C_1}=i}^1(n+1)}{1 - p_0^1(n+1)} [\alpha(p_2^1, p_0^1(n+1))]^j,
\end{aligned}$$

$$i = 0, 1, \ldots, N_2 - 1, j = 1, \ldots, N_2 - i,$$

$$P_j^1(n+1) = \sum_{i=0}^{N_2-j} P_i^{C_1}(n) \cdot P_{j,t_{C_1}=i}^{C_1}(n+1),$$

$$j = 0, 1, \ldots, N_2. \quad (12)$$

(3) Determine the probability of virtual machine $m_0^{C_1}$ to be up, and figure out the probability that b_2 has j group 2&3 parts.

$$\begin{aligned} p_0^{C_1}(n+1) &= p_0(1 - x_1(n))(q_1 P_0^1(n+1) \\ &\quad \times (1 - P_w^1(n))(1 - P_{WT}^1(n+1)) \\ &\quad + \sum_{u=2}^{3} q_u P_0^u(n)(1 - P_w^u(n))(1 - P_{WT}^u(n)) \\ &\quad + q_u(1 - P_0^u(n)), 1 \le WT \le N_2, \end{aligned}$$

$$P_{0,t_1=i}^{C_1}(n+1) = Q(p_2^{C_1}, p_0^{C_1}(n+1), N_2 - i),$$

$$i = 0, 1, \ldots, N_2,$$

$$P_{j,t_1=i}^{C_1}(n+1) = \frac{P_{0,t_1=i}^{C_1}(n+1)}{1 - p_0^{C_1}(n+1)} [\alpha(p_2^{C_1}, p_0^{C_1}(n+1))]^j,$$

$$i = 0, 1, \ldots, N_2 - 1, j = 1, \ldots, N_2 - i,$$

$$P_j^{C_1}(n+1) = \sum_{i=0}^{N_2-j} P_i^1(n+1) \cdot P_{j,t_1=i}^{C_1}(n+1),$$

$$j = 0, 1, \ldots, N_2. \quad (13)$$

(4) Ascertain the probability that the waiting conditions are satisfied, that is, the desirable matching group 1 part is not available, the buffer b_2 is not empty, and its current occupancy is less than the waiting threshold.

$$\begin{aligned} P_w^1(n+1) &= p_0(1 - x_1(n)) q_1 P_0^1(n+1) \\ &\quad \cdot \sum_{i=1}^{WT-1} P_i^{C_1}(n+1), \ 1 \le WT \le N_2. \end{aligned} \quad (14)$$

- Step 3: Calculation of the occupancy status probability of group 2 and group 1&3 parts in buffer b_2 using the similar calculation philosophy as in Step 2 (See Section A.2 of Appendix 1 in details).
- Step 4: Calculation of the occupancy status probability of group 3 and group 1&2 parts in buffer b_2 using the similar calculation philosophy as in Step 2 (See Section A.2 of Appendix 1 in details).
- Step 5: Calculation of the probability that the buffers b_1 and b_2 are empty, respectively.

$$\begin{aligned} p_0^m(n+1) &= p_0(1 - x_2(n)) \left(1 - \sum_{u=1}^{3} P_w^u(n+1)\right), \\ x_1(n+1) &= Q(p_1, p_0^m(n+1), N_1), \\ p_0^s(n+1) &= p_0(1 - x_1(n)) \left(1 - \sum_{u=1}^{3} P_w^u(n+1)\right), \\ x_2(n+1) &= Q(p_2, p_0^s(n+1), N_2). \end{aligned} \quad (15)$$

- Step 6: Check of stopping condition. If

$$|P_j^u(n+1) - P_j^u(n)| < \varepsilon,$$
$$\forall u = 1, 2, 3, j = 0, 1, \ldots, N_2,$$

where ε is typically set to be 10^{-5}, then stop. Otherwise, return to Step 2.

After generating the solution algorithms for Policy III, we can derive the system performance measures. The assembly machine will be starved in the selective assembly system with potential waiting since the buffer is empty, and the main parts are waiting. Nevertheless, as a result of waiting, the main parts will be assembled with matching parts with a higher probability to produce more matched products. Consequently, we derive the estimated production rate of the three types of products $PR_i (i = 0, 1, 2)$ as follows:

$$\begin{aligned} PR_0 &= p_0(1 - x_1) \left(1 - \sum_{u=1}^{3} P_w^u(n+1)\right) \\ &\quad \times \left(\sum_{u=1}^{3} q_u(1 - P_0^u \cdot (1 - P_{WT}^u))\right), \\ PR_1 &= p_0(1 - x_1) \left(1 - \sum_{u=1}^{3} P_w^u(n+1)\right) \\ &\quad \times \left(\sum_{u=1,3} q_u P_0^u(1 - P_{WT}^u)(1 - P_0^2) \right. \\ &\quad \left. + q_2 P_0^2(1 - P_{WT}^2)(1 - P_0^1 \cdot P_0^3)\right), \\ PR_2 &= p_0(1 - x_1) \left(1 - \sum_{u=1}^{3} P_w^u(n+1)\right) \\ &\quad \times \left(\sum_{u=1,3} q_u P_0^u(1 - P_{WT}^u) P_0^2(1 - P_0^{4-u})\right), \\ &\quad 1 \le WT \le N_2. \end{aligned}$$

Note that details about approximation methods and performance measures for three-machine multi-group with large buffers systems under Policy II and II can be found in Section A.3 of Appendix 1. Then we can derive

the calculation formula of the system evaluation index as:

$$TR = \sum_{i=0}^{2} Y_i \cdot PR_i.$$

4.3. Approximation method for multi-machine multi-group with large buffers systems

For multi-machine ($M \geq 3$) multi-group ($G \geq 3$) with large buffers systems, another approximate method is proposed by adding aggregation idea into the former approximate method for performance evaluation. Specifically, the aggregation procedure introduced in Chapter 16 of Li and Meerkov (2009) can be used to simplify multiple machines line into three machines and two buffers. It has been indicated that such methods are efficacious, and an accurate estimate of system production rates can be obtained.

We consider a multi-machine selective assembly system. Its main assembly line and sub-assembly line contain M_1 and M_2 machines, respectively, followed by M_0 processing machines, as shown in Figure 5. The main idea of the aggregation procedure is that the forward and backward aggregations for machine parameters are alternately performed in the 'upper' lines and the 'lower' lines. Such aggregations will be terminated when the stop condition is met. Then we can obtain the parameters $\hat{p}_{i,M_i}$ of the machine $\hat{m}_{i,M_i}$ after the aggregation procedure. Note that superscripts 'uf' and 'ub' stand for the forward and backward aggregations in the 'upper' lines, and superscripts 'lf' and 'lb' stand for those in the 'lower' lines. The calculation formula of $\hat{p}_{i,M_i}$ is as follows:

$$\hat{x}_{1,M_1-1} = Q(p^{uf}_{M_1-1}, p^{ub}_{M_1}, N^{u}_{M_1-1}),$$
$$\hat{x}_{2,M_2-1} = Q(p^{lf}_{M_2-1}, p^{lb}_{M_2}, N^{l}_{M_2-1}), \quad (16)$$
$$\hat{p}_{i,M_i} = p_{i,M_i}(1 - \hat{x}_{i,M_i-1}), \quad (17)$$

In the above formula, $\hat{x}_{i,M_i-1}$ is the probability that buffer $b_{i,M_{i-1}}$ is empty.

We can obtain the parameter $\hat{p}_{0,1}$ of machine $\hat{m}_{0,1}$ after aggregation through the following calculation process:

$$\hat{P}_0 = Q(p^{uf}_{M_1+1}, p^{ub}_{M_1+2}, N^{u}_{M_1+1}),$$
$$\hat{P}_{N_{0,1}} = \frac{\hat{P}_0}{1 - p^{ub}_{M_1+2}}[\alpha(p^{uf}_{M_1+1}, p^{ub}_{M_1+2})]^j, \quad (18)$$
$$\hat{p}_{0,1} = p_{0,1}(1 - \hat{P}_{N_{0,1}}) \quad (19)$$

And $p_{0,1}$ is machine $m_{0,1}$ reliability. $\hat{P}_{N_{0,1}}$ is the probability that buffer B_{01} is not full and m_{02} does not take a part.

So far, we have simplified the multi-machine selective assembly system to a three-machine model in Section 4.2, as seen in Figure 6 below. p_0, p_1, p_2 are replaced by $\hat{p}_{0,1}$, $\hat{p}_{1,M_1}$, $\hat{p}_{2,M_2}$ respectively. Next, we use the approximation method for three-machine multi-group with large buffers systems to calculate the marginal probabilities of buffer occupancy status for group i parts. In turn, we can obtain the system evaluation for multi-machine multi-group with large buffers systems.

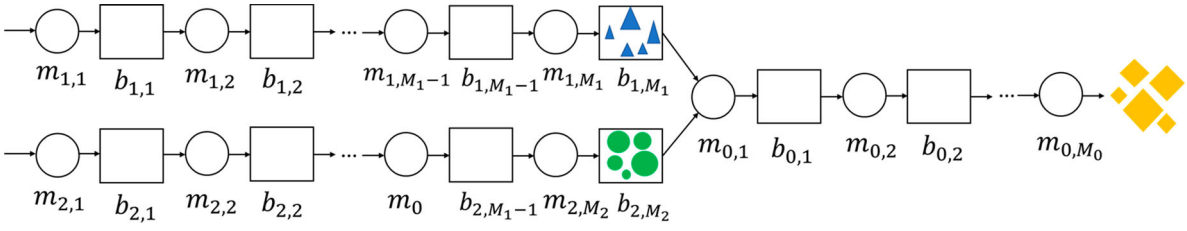

Figure 5. Multi-machine selective assembly system.

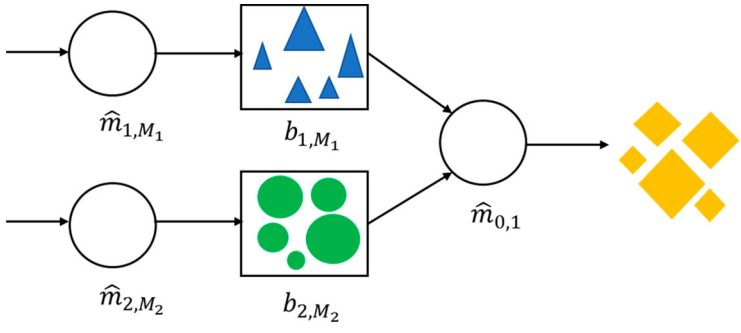

Figure 6. Simplified model of multi-machine selective assembly system.

5. Numerical studies

5.1. Convergence and accuracy discussion

In this section, first, we verify the rationality of convergence through numerous numerical studies. We take the three-machine three-group ($M_1 = M_2 = M_0 = 1, G = 3$) selective assembly system as an example to demonstrate the convergence of approximation method for three-machine multi-group with large buffers systems. The system parameters of each scenario are generated uniformly from the following sets:$p_i \in [0.5, 0.99]$, $i = 1, 2$; $p_0 \in [0.8, 0.99]$; $q_1, g_1 \in [0.5, 0.7]$; $q_2, g_2 \in [0.15, 0.25]$; $N_i \in [3, 20], i = 1, 2$. Then q_3 and g_3 can be calculated by formulas $q_3 = 1 - q_1 - q_2$ and $g_3 = 1 - g_1 - g_2$, respectively. We solve those cases by proposed methods, and every case we study converges immediately. Typically, Procedure 4.2 can terminate the iteration within 10 steps. Most cases can be solved in seconds. The convergence of the marginal probability of buffer occupancy for each group part is observed from Figures 7 and 8, which shows that convergence to the limits is quite fast. For multi-machine multi-group cases, we carry out similar numerical experiments to prove the convergence. Again, the procedure converges quickly in all the experiments.

For the purpose of verifying the accuracy of the two approximation methods we proposed, we carry out numerical experiments to compare the results of *PR* from approximation methods with the simulation results by using the *Arena* software. The warmup period of each simulation experiment is set to be 10,000 time units, and the replication length is 80,000 time units. For each line under consideration, we carry out 20 replications. And then statistically evaluate *PR*. We explore the accuracy of the two approximation methods proposed in Sections 4.2 and 4.3, respectively. Among them, to investigate the effect of buffer size on the accuracy of the methods, two experimental levels of buffer size, small and large, which are randomly generated from the uniform distribution [3,15] and [16,30], are designed. Also considering the groups, we design multiple experiments for the two methods with the following four combinations: three-group small buffers, multi-group ($G = 4, 5, 6$) small buffers, three-group large buffers, and multi-group ($G = 4, 5, 6$) large buffers to test the effect of buffer size and groups on the accuracy of the methods. A variety of input line instances are uniformly generated from the parameter ranges as listed in Appendix 2 Table A1. We assume that in the main line, all machines have identical reliability and all buffers are of equal size. So are in mating and final assembly lines. The parameters of machines and the capacities of buffers in main, mating, and final assembly lines are denoted as p_{ji} and

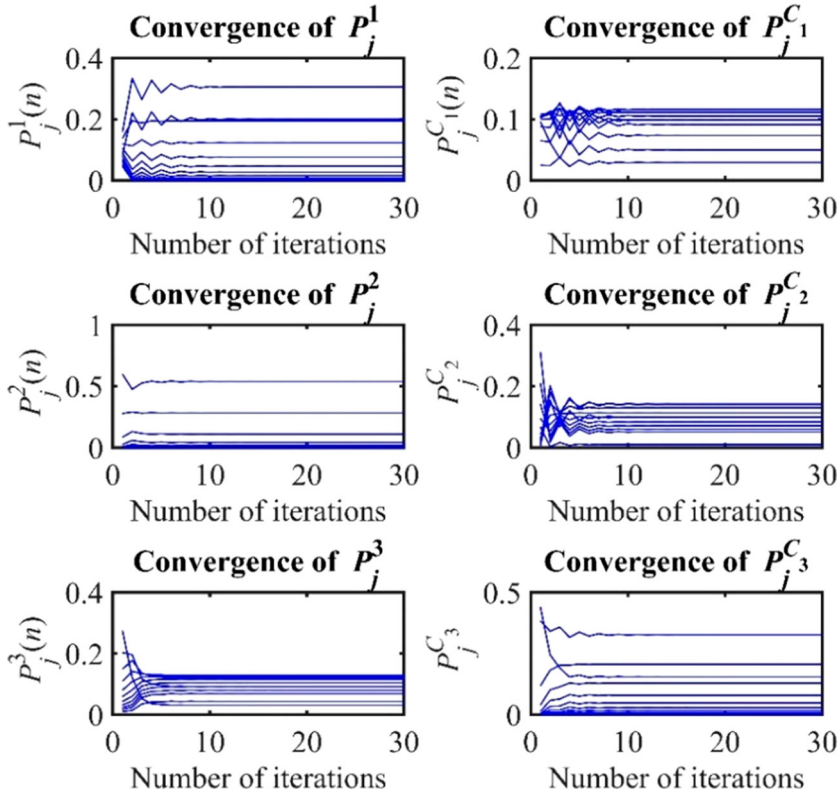

Figure 7. Illustration of convergence of Procedure 4.2:$p_1 = 0.90 p_2 = 0.94 p_0 = 0.98$, $q_1 = 0.62 q_2 = 0.20$, $g_1 = 0.54 g_2 = 0.15$, $N_1 = 8 N_2 = 10$, $WT = 6$.

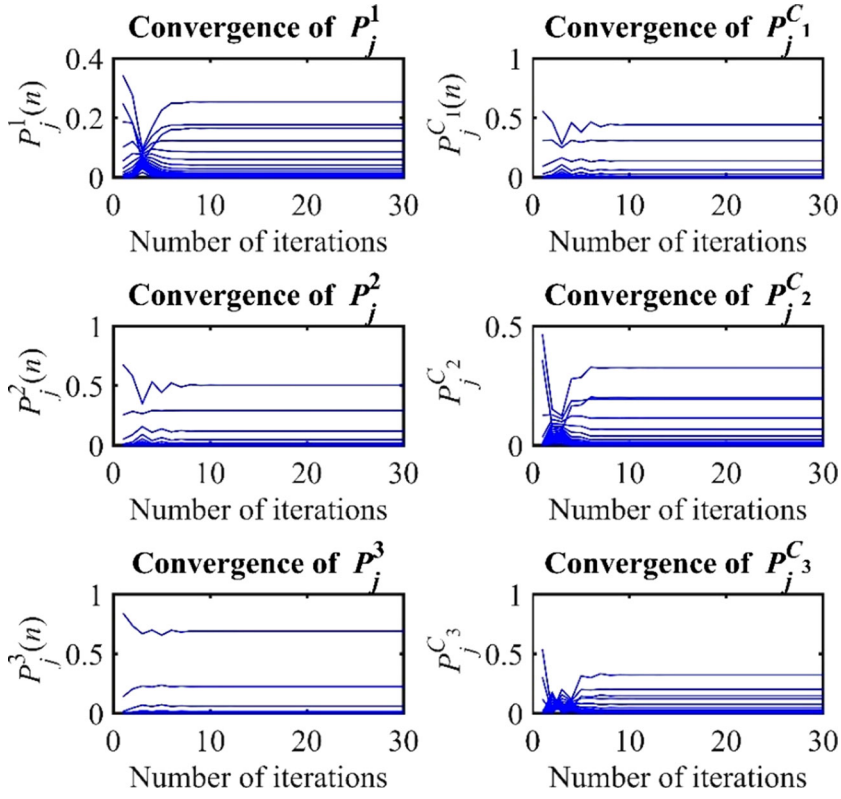

Figure 8. Illustration of convergence of Procedure 4.2:$p_1 = 0.72p_2 = 0.66p_0 = 0.96$, $q_1 = 0.51q_2 = 0.29$, $g_1 = 0.69g_2 = 0.23$, $N_1 = 14N_2 = 16$, $WT = 14$.

$N_{ji}, j = 1, 2, 0$, respectively (and in all subsequent studies). For a set of given parameters, $G, p_{1i}, p_{2i}, p_{0i}, q_i, g_i$ and WT, two groups of experiments under small and large buffers are designed, respectively. The subcomponent groups are set to be 3, 4, 5, or 6. The waiting threshold is set to be 1, $N_2/2$, or N_2. The relative error of the system performance measurement obtained by the simulation and the approximate algorithm is defined as follows:

$$Error = \frac{PR^a - PR^s}{PR^s} \cdot 100\%,$$

where superscripts $'a'$ and $'s'$ denote approximation method and simulation experiment, respectively.

Table A2 is the result of the accuracy of the approximation method designed for three-machine multi-group with large buffers systems. As seen from Table A2, the largest and the smallest relative error are 12.52% and 0.004%, respectively, and the average relative error of PR is 2.56%, which is fairly well. For multi-machine multi-group with large buffers systems, other numerical experiments are dealt with, and Table A3-A7 in Appendix 2 illustrates the examples of such experiments. The cases in Table A3-A7 has 2, 3, 4 or 5 machines in the main line, while 2, 3, 4, 6 or 8 machines are in the mating line, and 2, 4, or 5 machine follows the assembly machine. The average relative error of PR is 4.59%.

Besides, Tables A2–A7 in Appendix 2 show the examples of accuracy performance results of four combinations of cases. Considering the experiments of both approximation methods in a unified manner, we have 32 experiments for each combination, three-group small buffers, multi-group small buffers, three-group large buffers, and multi-group large buffers. We calculate the average relative errors of the four combinations in Tables A2-A7. The results show that multi-group small buffers cases have the largest mean error of 5.41%, followed by three-group small buffers: 5.27% and multi-group large buffers: 2.41%. The three-group large buffers cases have the smallest mean error of 2.23%. The accuracy of the approximation methods decreases as the number of groups increases and the buffer size decreases.

Despite this accuracy limitation, the approximation methods provide a sufficiently efficient analytical tool for performance evaluation of the selective system employing different scheduling policies, especially considering that the parameters of the machines are collected on the factory floor with an error of 5–10%.

5.2. Policies analysis

Since TR is determined by many system variables, it is not realistic to locate a particular range that is in line with a

performance manner. In this section, we first fix the waiting threshold to get some ideas about how the discount factor a will affect the three policies' performance. Then, we investigate the influence of machine efficiency on the performance of the three policies. Moreover, for Policy III, its performance under various patterns of buffer size and machine reliability is evaluated. And the optimal waiting threshold is searched to maximise system revenue. We exhibit representative scenarios and expatiate upon some necessary observations in each subsection. It is worth noting that the parameters of the numerical experiments are designed to be very similar so that the impacts of changing different parameters on the system performance can be compared with each other. These experiments will provide useful insights to help effectively apply the scheduling policy in selective assembly systems. And the patterns of buffer size and machine reliability numerical experiments also motivate the performance improvement of similar practical systems by implementing the new type of policy.

5.2.1. How does the discount factor affect the performance of the three policies?

This subsection aims to find out how the performance of the three policies depends on the discount factor and to get some idea about what other key system and strategy parameters will have a major impact on strategy performance. We consider eight examples with different system parameter configurations, as shown in Table A8. And to point out the advantages of our proposed policy over previous research, we compare the system performances using three policies for each example.

Table A9 summarises the three policies' results in terms of total revenue as discount factor a changes from 0 to 1. We can observe that nearly in all the cases, both Policy III and II perform better than Policy I. Figure 9 further visualises the results of Cases 1–4. For Cases 1–3, Policy III always performs better than Policy II for any a except one. Additionally, the revenue improvement by Policy III decreases with the increment of the discount factor. This is attributable to the fact that Policy III will increase the PR of matched assemblies and reduce mismatched assemblies. The smaller the discount factor, the larger the gap between the perfectly matched assemblies' value and the mismatched ones'. When the growth of the matched assemblies' value is greater than the decrease in mismatched ones', the result shows an increase in total revenue. For Case 4, it can be clearly noticed that there is a unique intersection of Policy II and III curves. Policy III is better than Policy II until a increases to a certain value. After that, the performance of Policy III is worse than Policy II. This is because that when a is large enough, the value of the matched assemblies is only slightly enhanced and cannot compensate for the reduced value of the mismatched ones. The overall effect is a decline in total revenue.

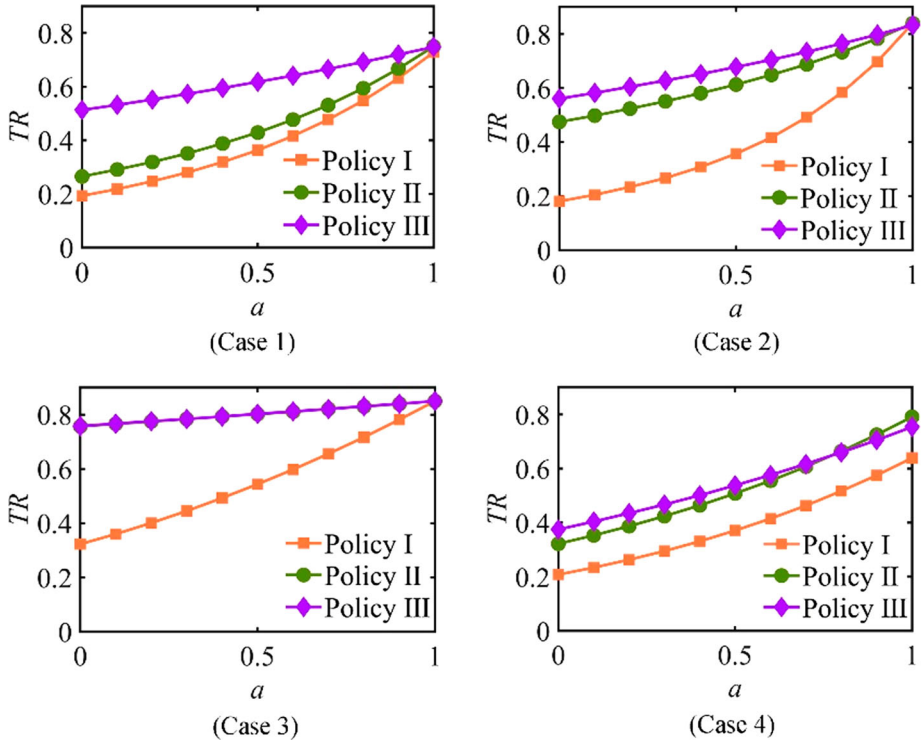

Figure 9. The comparison results for Cases 1–4.

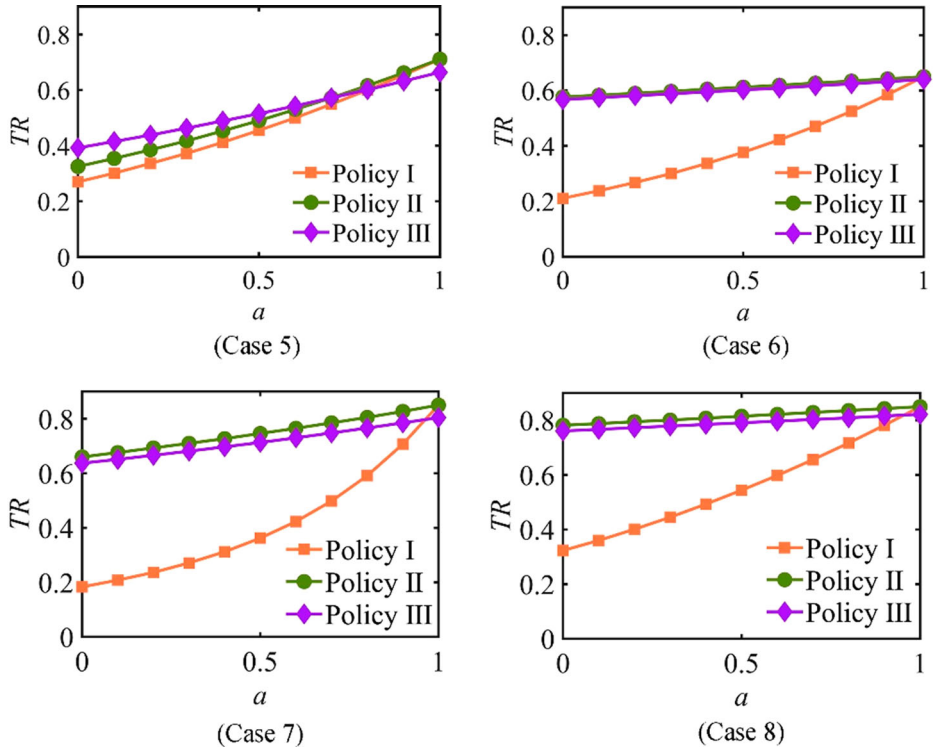

Figure 10. The comparison results for Cases 5–8.

The difference between Case 5 and 4 is that there is another intersection of Policy I and III. When a is larger than the intersection, Policy III will be the most disappointing one. For Cases 6–8, Policy II has completely surpassed Policy III, as shown in Figure 10. It is owing to the long waiting time for the matching part's arrival. The production system has been down for too long, which has caused a loss to the system efficiency. As a result, the product throughput has declined, but the product quality has not been obviously improved.

From the above numerical results, we can obtain the following conclusions.

- Nearly in all systems, as the discount factor varies, both *WCQMP* and *CQMP* perform better than *RMP*. And when pivotal system parameters, such as machine reliability, buffer size, are properly designed, *WCQMP* can stand out by comparison with *CQMP*.
- When *WCQMP* performs better (worse) than *CQMP* or *RMP*, a smaller (larger) discount factor will lead to a larger performance gap between *WCQMP* and the other two policies.

5.2.2. How does machine reliability affect the performance of the three policies?

In this subsection, we investigate the impact of machine reliability on the system performance measures. It aims to point out which scheduling policy is more conducive to system revenue improvement with different machine reliability configurations. We consider an assembly system consisting of three machines with each sub-component divided into three groups and buffer capacity $N_1 = N_2 = 2$. The discount factor is assigned as 0.3 (and in all subsequent studies). We assume that $q_i = g_i$, $i = 1, 2, 3$. Specifically, $q_1 = g_1 = 0.60$, $q_2 = g_2 = 0.24$, $q_3 = g_3 = 0.16$. The machine parameters of each machine are equal. We have $p_0 = p_1 = p_2 = 0.85$. If there are more than three machines in the assembly line, we make the aggregated machines parameters $\hat{p}_{0,1} = \hat{p}_{1,M_1} = \hat{p}_{2,M_2} = 0.85$.

First, we study how p_0 affects the three policies' performance and conduct a comparative analysis of them. Let p_0 vary from 0.49 to 0.99 while $p_1 = p_2 = 0.85$ and all the other system parameters remain unchanged. Figure 11(a) demonstrates how the system quality-related revenue contributed by three policies depends on p_0. Whatever the value of p_0 takes, the curves of Policy II and III always lie above Policy I, and the gap between them is evident. Moreover, as p_0 grows, the curve representing Policy III almost grows linearly, while the curve of Policy II first gradually rises and then tardily goes down. Therefore, the enhancement of p_0 will make Policy III perform better. In this situation, it is intuitively straightforward since the upstream machines reliability will be the main constraint for the production rate improvement when p_0 is large. Hence, allowing machine m_0 to wait for a while will

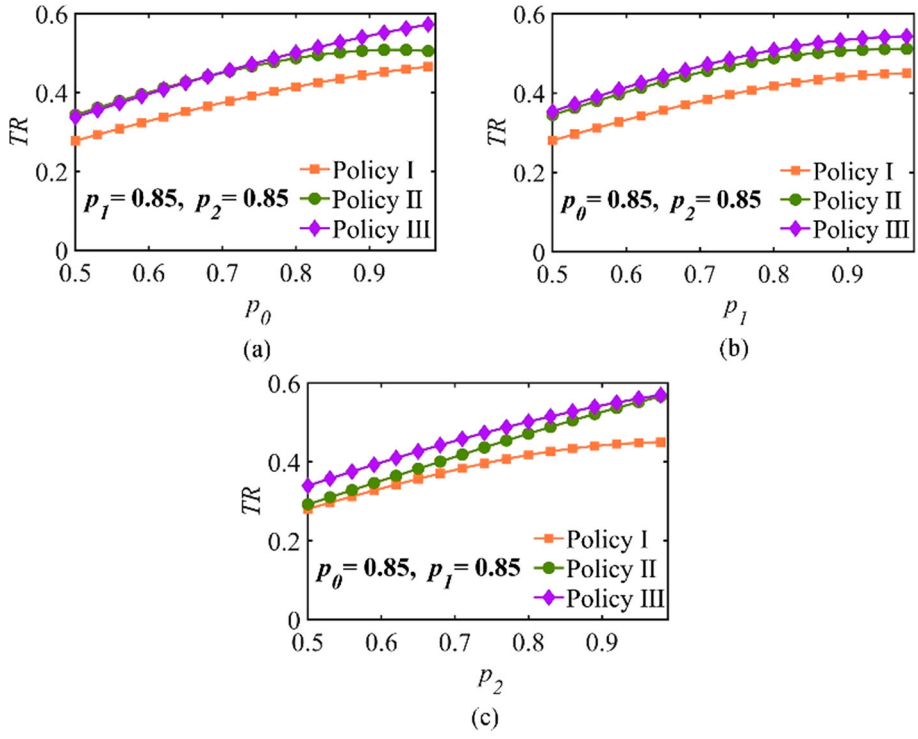

Figure 11. Impact of machine reliability (a) p_0, (b)p_1, (c) p_2 on the performance of three policies.

have a minor impact on the throughput loss, which does not exceed the effect on quality improvement caused by the waiting. Consequently, if there is a chance to improve the assembly quality by delaying the assembly process for a proper period, Policy III should be considered first.

Then, to investigate the impact of p_1, we consider the assembly lines with parameter $p_0 = p_1 = 0.85$ when p_1 changes. Likewise, the remaining system parameters of examined assembly lines keep the same. From Figure 11(b), we can observe that the order of the performance of the three policies is always: Policy III $>$ Policy II $>$ Policy I. Furthermore, the performance of the former two and the latter is obviously different on every occasion. Policy III will perform better with a larger p_1. In this context, machine m_2 will become the bottleneck of the system production rate. Starved by the empty buffer b_2, the production efficiency of the assembly machine is not high. And the probability of generating mismatches will increase. Although waiting for a desirable mating part will cause a further decline in its assembly efficiency, at the same time, the assembly quality has been greatly improved. Implementing Policy III of postponing the assembly process has a more evident impact on the quality improvement than the throughput impediment.

Finally, we study the impact of machine m_2 reliability in the assembly lines where $p_0 = p_1 = 0.85$ and p_2 changes. The rest of the system parameters keep the same. Figure 11(c) illustrates that with respect to p_2, Policy III performs best for most cases. Moreover, opposite to p_0 and p_1, a larger p_2 will inhibit the performance improvement effect brought by Policy III, resulting in a smaller performance gap between it and the other two policies. In this way, the bottleneck will be machine m_1. The probability that there is no matching part available in buffer b_2 is very low. It is not beneficial to only pursue the current individual assembly quality improvement, but ignore the fact that if the assembly process continues, more matched assemblies will be made without waiting. Relatively speaking, instead of sacrificing working time for local quality improvement, it is better to continue processing to obtain a higher level of overall throughput.

From the above observations, we can draw the conclusions as follows. Under keeping the other system configurations fixed, for systems with higher main line reliability, higher final assembly line reliability, or a lower level of mating line reliability, the benefits brought by *WCQMP* will be more in contrast to *RMP* and *CQMP*.

5.2.3. How do the patterns of buffer size and machine reliability affect the performance of the Policy III?

In this subsection, we investigate the impact of buffer size and machine reliability patterns on the performance of Policy III. First, lines with identical buffers and machine reliability are investigated. Next, we consider

Table 2. Three-machine ($M_1 = 1, M_2 = 1, M_0 = 1$) lines with identical $N_j, j = 1, 2$, and $p_1 = p_2 = p_0 = 0.85$.

N_1	N_2	PR_0	PR_1	PR_2	TR
4	4	0.589	0.15	0.022	0.669
5	5	0.627	0.146	0.021	0.705
6	6	0.65	0.139	0.019	0.724
7	7	0.666	0.132	0.018	0.737
8	8	0.68	0.125	0.017	0.747
9	9	0.692	0.119	0.015	0.755
10	10	0.7	0.114	0.015	0.761
11	11	0.709	0.109	0.013	0.767
12	12	0.716	0.105	0.012	0.771
13	13	0.722	0.101	0.012	0.775
14	14	0.728	0.097	0.011	0.779
15	15	0.733	0.093	0.011	0.782

Table 3. Multi-machine ($M_1 = 2, M_2 = 2, M_0 = 2$) lines with identical $N_{ji}, j = 0, 1, 2$, and $p_{1i} = p_{2i} = p_{0i} = 0.85$.

N_{1i}	N_{2i}	N_{0i}	PR_0	PR_1	PR_2	TR
16	16	16	0.737	0.091	0.01	0.785
17	17	17	0.741	0.088	0.01	0.787
18	18	18	0.746	0.084	0.009	0.791
19	19	19	0.75	0.081	0.009	0.793
20	20	20	0.751	0.081	0.009	0.794

Table 4. Multi-machine ($M_1 = 5, M_2 = 6, M_0 = 4$) lines with identical $N_{ji}, j = 0, 1, 2$, and $p_{1i} = p_{2i} = p_{0i} = 0.85$.

N_{1i}	N_{2i}	N_{0i}	PR_0	PR_1	PR_2	TR
21	21	21	0.755	0.078	0.008	0.796
22	22	22	0.758	0.075	0.008	0.798
23	23	23	0.761	0.074	0.007	0.8
24	24	24	0.763	0.072	0.007	0.801
25	25	25	0.765	0.071	0.007	0.802

lines with nonidentical buffers and equal machine reliability to study the impact of buffer capacity. Finally, lines with identical buffers but nonidentical machine reliability are considered to investigate the impact of machine reliability.

5.2.3.1. Lines with identical buffers and machine reliability. First, we consider three-machine lines with $N_j \in [4, 15]$, $p_1 = p_2 = p_0 = 0.85$. We assume that subcomponents have three groups and $q_1 = g_1 = 0.60$, $q_2 = g_2 = 0.24$, $q_3 = g_3 = 0.16$, $a = 0.5$, $WT = 4$ (and in all subsequent studies). As shown in Table 2, a larger buffer size always leads to a higher production rate for exactly matched assemblies and lower that of mismatched ones. And total revenue grows as buffer size becomes larger. The reason is that larger buffers contribute to a higher production rate, so that more high-quality assemblies will be produced and the total revenue will be higher. For multi-machine lines with identical buffers (capacities 16–25) and identical machine reliability, similar properties can be observed (see Tables 3 and 4).

Table 5. Three-machine ($M_1 = 1, M_2 = 1, M_0 = 1$) lines with nonidentical N_j, $j = 1, 2$, and $p_1 = p_2 = p_0 = 0.85$.

N_1	N_2	PR_0	PR_1	PR_2	TR
4	6	0.646	0.137	0.019	0.719
6	4	0.592	0.15	0.022	0.673
6	8	0.68	0.123	0.016	0.745
8	6	0.651	0.14	0.02	0.726
8	10	0.701	0.112	0.014	0.76
10	8	0.681	0.126	0.017	0.748
10	12	0.716	0.103	0.012	0.771
12	10	0.701	0.115	0.014	0.762

Table 6. Multi-machine ($M_1 = 3, M_2 = 3, M_0 = 2$) lines with nonidentical $N_{ji}, j = 0, 1, 2$, and $p_{1i} = p_{2i} = p_{0i} = 0.85$.

N_{1i}	N_{2i}	N_{0i}	PR_0	PR_1	PR_2	TR
12	10	10	0.701	0.115	0.014	0.762
10	12	10	0.716	0.103	0.012	0.771
10	10	12	0.701	0.114	0.014	0.761
12	12	14	0.717	0.104	0.012	0.772
14	12	12	0.716	0.105	0.012	0.772
12	14	12	0.728	0.095	0.011	0.779
14	14	16	0.728	0.097	0.011	0.779
14	16	14	0.739	0.089	0.01	0.785
16	14	14	0.728	0.097	0.011	0.779

Table 7. Multi-machine ($M_1 = 4, M_2 = 8, M_0 = 5$) lines with nonidentical $N_{ji}, j = 0, 1, 2$, and $p_{1i} = p_{2i} = p_{0i} = 0.85$.

N_{1i}	N_{2i}	N_{0i}	PR_0	PR_1	PR_2	TR
6	4	4	0.594	0.152	0.022	0.676
4	6	4	0.647	0.138	0.019	0.721
4	4	6	0.593	0.152	0.023	0.675
6	6	8	0.651	0.14	0.02	0.726
8	6	6	0.651	0.141	0.02	0.727
6	8	6	0.68	0.124	0.016	0.746
8	8	10	0.68	0.126	0.017	0.747
8	10	8	0.702	0.112	0.014	0.761
10	8	8	0.681	0.127	0.017	0.749

5.2.3.2. Lines with nonidentical buffers and equal machine reliability. Next, we consider three- and multi-machine lines to study the impact of buffer capacity patterns. We can observe from Table 5 that comparing to cases of $N_1 > N_2$, cases of $N_2 > N_1$ indicate increases in PR_0, and decreases in PR_1 and PR_2. Total revenue shows an upward trend. As we can see from Table 6, the buffer pattern where N_{1i} and N_{0i} are always equal and smaller than N_{2i} implies decreases in PR_1 and PR_2. Similarly, the buffer pattern $N_{1i} = N_{0i} < N_{2i}$ also leads to larger PR_0 and TR. The results of the other two patterns ($N_{1i} > N_{0i} = N_{2i}$ and $N_{1i} = N_{2i} < N_{0i}$) only have slight differences. Analogous findings are observed in multi-machine ($M_1 = 4, M_2 = 8, M_0 = 5$) lines (Table 7).

5.2.3.3. Lines with identical buffers and nonidentical machine reliability. Finally, we investigate the impact of machine reliability patterns by considering three- and multi-machine lines. In Table 8, four patterns of machine reliability are studied: $p_{1i} < p_{2i} < p_{0i}$, $p_{1i} > p_{2i} > p_{0i}$,

Table 8. Three-machine ($M_1 = 1, M_2 = 1, M_0 = 1$) lines with identical $N_j = 10$, $j = 1, 2$, and different p_{ji}, $j = 0, 1, 2$.

p_{1i}	p_{2i}	p_{0i}	PR_0	PR_1	PR_2	TR
0.65	0.8	0.95	0.572	0.071	0.007	0.609
0.95	0.8	0.65	0.573	0.07	0.007	0.61
0.7	0.8	0.9	0.612	0.079	0.009	0.653
0.9	0.8	0.7	0.611	0.079	0.009	0.653
0.85	0.65	0.85	0.509	0.123	0.018	0.575
0.65	0.85	0.65	0.558	0.065	0.006	0.592
0.7	0.6	0.7	0.479	0.103	0.015	0.534
0.6	0.7	0.6	0.505	0.065	0.007	0.539

Table 9. Multi-machine ($M_1 = 4, M_2 = 4, M_0 = 2$) lines with identical $N_{ji} = 10, j = 0, 1, 2$, and different $p_{ji}, j = 0, 1, 2$.

p_{1i}	p_{2i}	p_{0i}	PR_0	PR_1	PR_2	TR
0.55	0.75	0.95	0.487	0.057	0.006	0.517
0.95	0.75	0.55	0.486	0.058	0.006	0.516
0.65	0.75	0.85	0.566	0.075	0.009	0.605
0.85	0.75	0.65	0.565	0.075	0.009	0.604
0.9	0.7	0.9	0.544	0.135	0.02	0.617
0.7	0.9	0.7	0.605	0.07	0.006	0.641
0.95	0.85	0.95	0.666	0.161	0.024	0.752
0.85	0.95	0.85	0.745	0.085	0.007	0.79

Table 10. Multi-machine ($M_1 = 5, M_2 = 6, M_0 = 4$) lines with identical $N_{ji} = 10, j = 0, 1, 2$, and different $p_{ji}, j = 0, 1, 2$.

p_{1i}	p_{2i}	p_{0i}	PR_0	PR_1	PR_2	TR
0.6	0.75	0.9	0.529	0.066	0.007	0.563
0.9	0.75	0.6	0.527	0.065	0.007	0.561
0.75	0.85	0.95	0.659	0.083	0.008	0.703
0.95	0.85	0.75	0.657	0.083	0.009	0.701
0.95	0.75	0.95	0.582	0.146	0.022	0.66
0.75	0.95	0.75	0.652	0.075	0.006	0.691
0.9	0.8	0.9	0.632	0.146	0.021	0.71
0.8	0.9	0.8	0.694	0.083	0.008	0.738

$p_{1i} = p_{0i} > p_{2i}$ and $p_{1i} = p_{0i} < p_{2i}$. Comparing with the pattern $p_{1i} = p_{0i} > p_{2i}$, the pattern where p_{1i} and p_{0i} are always equal and smaller than p_{2i} can lead to an increasing distribution of production rate of matched assemblies and total revenue. There is an unobvious contrast between the production rate of each type of product and the total revenue of the two kinds of machine reliability distributions $p_{1i} < p_{2i} < p_{0i}$ and $p_{1i} > p_{2i} > p_{0i}$. When multi-machine lines are considered, as demonstrated in Tables 9 and 10, similar properties still hold.

Based on the extensive numerical experiments, we can draw the following conclusions.

- For selective assembly systems with identical buffers and machine reliability, larger buffers lead to a higher production rate of exactly matched assemblies and lower that of mismatched ones, resulting in higher total revenue under *WCQMP*.
- For selective assembly systems with nonidentical buffers and equal machine reliability, *WCQMP* provides more revenue to the system following the buffer pattern where the main line and final assembly line buffers capacities are always equal and smaller than those of mating line buffers.
- For selective assembly systems with identical buffers and nonidentical machine reliability, when adopting *WCQMP*, the machine reliability pattern where the main line and final assembly line machines reliability are always equal and smaller than that of mating line machines can bring about larger total revenue.

5.2.4. Optimisation waiting threshold decision for a fixed system

In this subsection, first, we focus on the change of waiting threshold constraint and its impacts on the selective assembly system performance. For this, three multi-machine three-group assembly lines with typical machine reliability settings are under consideration. The parameter configurations are summarised in Table 11. N_{1i}, N_{2i}, N_{0i}, q_i, g_i and a are assumed to be fixed values, respectively, except for a certain waiting threshold that is used to examine its impacts. We study how total revenue behaves as the waiting threshold WT increases from 1 to 20. The results of the three assembly lines are examined in Figure 12, respectively. In these figures, the total revenue is illustrated as function of the waiting threshold.

For L_1, it can be clearly observed that the assembly system's total revenue has the increasing monotonicity on WT. Therefore, the relaxation of WT is always beneficial to performance improvement. This result is attributed to the fact that longer downtime has a greater impact on assembly quality than on throughput. Therefore, the optimal WT^* is the buffer capacity 20. For L_2, the total revenue shows a tendency first to increase and then slowly fall with the change of WT. Among them, there is a sharp increase at $WT = 2$. Then the maximum TR appears at $WT = 4$. After that, the total revenue achieved by increasing WT has not significantly changed, but only a slight decrease. This is because the long waiting downtime has a major impact on the reduction in throughput, while the assembly quality has not been improved markedly. In consequence, a slight drop in the system total revenue occurs. In such a line, we can achieve maximum revenue when $WT^* = 4$. For L_3, no matter how large the WT is, the increase in total revenue brought about by the improvement of product quality is negligible for this system with high machine reliability. On the contrary, a little waste of working time will cause an apparent loss of throughput. Therefore, the total revenue has shown a continuous downward trend. In this context, $WT^* = 1$, that is, not waiting is the optimal decision for the selective assembly system.

From the above examples, it can be seen that the optimal waiting thresholds for different systems when obtaining the maximal total revenue are distinct. To figure out how to determine an optimal or the best possible waiting threshold for a fixed system, we further investigate the monotonicity properties of the system with the waiting threshold constraint by examining a small assembly system. We consider a three-machine three-group selective assembly line with $p = [0.85, 0.8, 0.95]$, $N = [10, 25]$, $q_i = g_i = [0.6, 0.3, 0.1]$ and $a = 0.3$. Six experiments are implemented in this example to study the impact of machine efficiency, the buffer size, and the discount factor on the change of optimal waiting threshold. Specifically, in Experiments 1, 2, and 3, the reliability of machines m_1, m_2, and m_0, respectively, is ranged from 0.69 to 0.99 and the other parameters are kept fixed. In Experiments 4 and 5, the buffer capacity is varied from 5 to 35 and the other parameters are kept fixed. In Experiments 6, the range of discount factor a is from 0.2 to 0.8. The configurations of these experiments are listed in Table 12.

As seen in Figure 13, for various selective assembly lines generated by setting different machine reliability parameters p_1, p_2 or p_0, when WT is increased, the monotonicity of total revenue may not hold all the time. Accordingly, the optimal waiting threshold will alter. Specifically, we can observe from Figure 13 that when other system parameters are fixed, the optimal WT^* tends to become smaller in general as p_1 increases. As p_1 increases, more main parts need to be matched, but relatively speaking, the machine m_2 is less efficient at this time, which is the bottleneck of the system. The probability of the desirable mating part's arrival is still very low though after a period of waiting. And when the cumulative waiting time is too long, postponing the assembly process has a greater impact on the throughput impediment than the assembly quality improvement. Therefore, WT^* tends to become smaller.

And a larger p_2 or p_0 leads to an increasing trend for the optimal WT^*, overall. As p_2 increases, the efficiency of machine m_1 is relatively lower at this time, which is the bottleneck of the system. And there are not so many

Table 11. Detailed parameters for three typical lines.

Lines	M_1	M_2	M_0	G	N_{1i}, N_{2i}, N_{0i}	p_{1i}, p_{2i}, p_{0i}	q_i	g_i	a
L_1	3	3	2	3	20,20,20	0.75,0.75,0.85	0.50,0.30,0.20	0.50,0.30,0.20	0.5
L_2	3	3	2	3	20,20,20	0.95,0.75,0.85	0.50,0.30,0.20	0.50,0.30,0.20	0.5
L_3	3	3	2	3	20,20,20	0.95,0.95,0.85	0.50,0.30,0.20	0.50,0.30,0.20	0.5

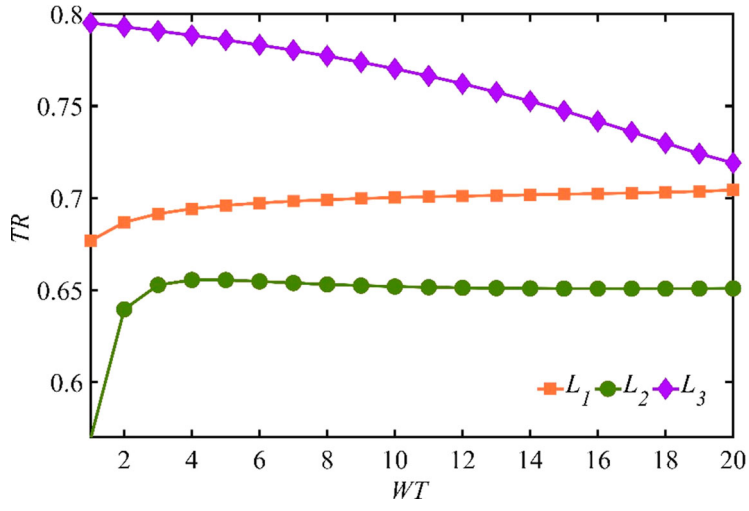

Figure 12. Performance results with respect to the waiting threshold.

Table 12. Parameters configurations of experiments.

No. of Experiments	p_1	p_2	p_0	N_1	N_2	a	q_i	g_i
1	[0.69,0.99]	0.8	0.95	10	25	0.3	0.6,0.3,0.1	0.6,0.3,0.1
2	0.85	[0.69,0.99]	0.95	10	25	0.3	0.6,0.3,0.1	0.6,0.3,0.1
3	0.85	0.8	[0.69,0.99]	10	25	0.3	0.6,0.3,0.1	0.6,0.3,0.1
4	0.85	0.8	0.95	[5,35]	25	0.3	0.6,0.3,0.1	0.6,0.3,0.1
5	0.85	0.8	0.95	10	[5,35]	0.3	0.6,0.3,0.1	0.6,0.3,0.1
6	0.85	0.8	0.95	10	25	[0.2,0.8]	0.6,0.3,0.1	0.6,0.3,0.1

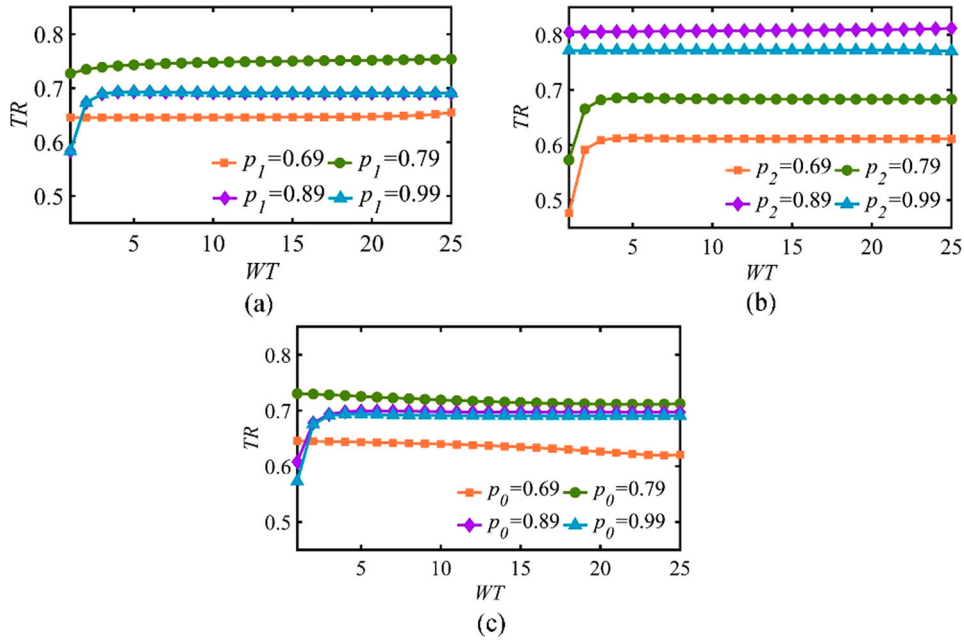

Figure 13. Impact of the machine reliability on the changes of the optimal waiting threshold.

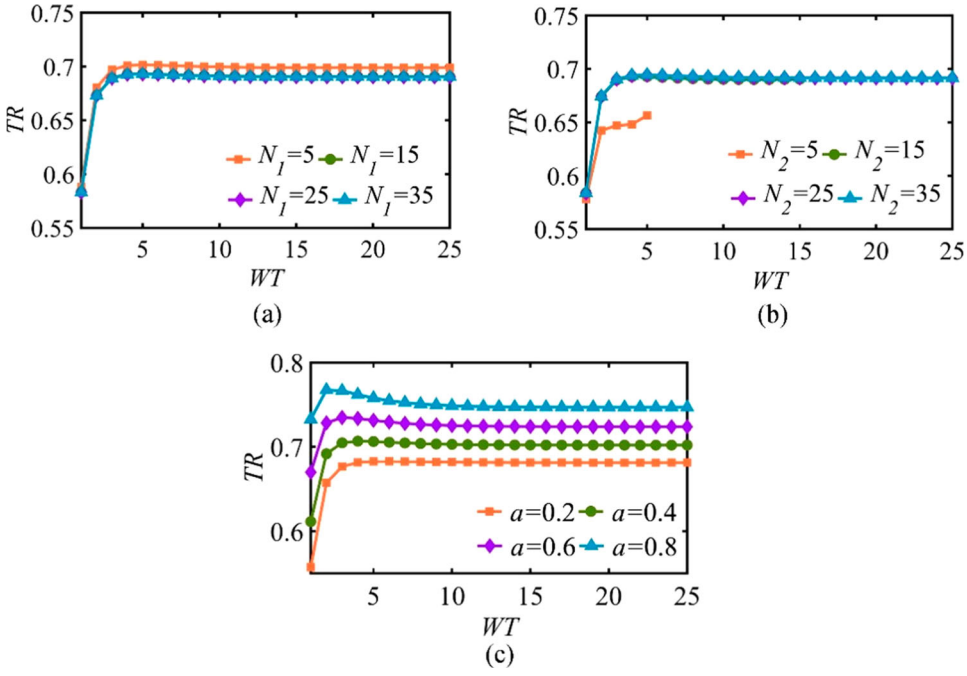

Figure 14. Impact of the buffer size and discount factor on the changes of the optimal waiting threshold.

main parts to match. Compared with the system without waiting, making the assembly machine out of work for a while will not cause too much loss in throughput. As long as a small increase in the probability of the desirable mating part's arrival, it will make an obvious assembly quality improvement. Therefore, there is a growing trend of WT^*. When p_0 is small, the assembly machine m_0 is the bottleneck of the system at this time. Waiting for a better matching will inevitably lead to a further decline in system throughput, and the main effect caused by shutdowns is to impede the throughput. Therefore, as p_0 decreases, there is a decreasing trend of WT^*. Conversely, as p_0 increases, WT^* tends to become larger.

Next, we study the monotonicity pattern exhibiting by lines differing in buffer capacity or discount factor. Similarly, the total revenue is not monotonic concerning the waiting threshold. As shown in Figure 14, buffer size has a minor effect on determining the value of the optimal threshold, and a larger discount factor leads to a smaller optimal threshold.

We can obtain the following conclusions from the results. For various selective assembly lines with different system parameters, the monotonicity of total revenue with respect to the waiting threshold may not always exist. The optimal waiting threshold exists when the assembly system makes the trade-off between assembly quality improvement and an impediment to the throughput. In general, compared to buffer size and discount factor, machine efficiency has a greater decisive effect on the value of the optimal waiting threshold. Furthermore, as main line reliability increases, the optimal waiting threshold tends to become smaller on the whole. Higher mating line reliability or final assembly line reliability leads to an increasing trend for the optimal waiting threshold. We can draw insights about the decision of optimal waiting threshold from these conclusions. When applying the proposed *WCQMP*, a manager in the industry should adjust the policy parameter to achieve a better system performance when machine efficiency has changed in a real system. Specifically, when the main line efficiency is improved, the waiting threshold should be adjusted to be smaller to obtain maximal revenue. Besides, if the mating line or final assembly line is aging during the practical production operation, the waiting threshold should be decreased. Those provide managers policy adjustment guidelines when a system has great changes or improvements.

6. Conclusions

The paper considers the quality-related total revenue maximisation problem in the selective assembly system by adopting different scheduling policies for matching operations. The *Waiting for Closest Quality Matching Policy (WCQMP)* is explicitly proposed to increase the probability of producing high-quality assemblies. Meanwhile, we propose the other two policies, *Random Matching Policy* (*RMP*) and *Closest Quality Matching Policy* (*CQMP*) as comparisons. System performance when employing the three policies is evaluated by exact and approximation methods for small and larger systems, respectively. The convergence and accuracy of the approximate methods are verified numerically. We conclude that in most cases, both *WCQMP* and *CQMP* outperform *RMP*. While for *WCQMP*, when system and policy parameters are properly designed, the superiority of *WCQMP* is more prominent by improving assembly quality without overly sacrificing system throughput, thereby increasing quality-related revenue. A few beneficial insights are also provided for industrial managers to improve system revenue by using our proposed policy *WCQMP* more appropriately in practice.

Lots of future work can be implemented furtherly. Firstly, the distribution of machine reliability, such as geometric, Weibull, and general distributions, can be further studied in selective assembly systems. Secondly, models to assembly systems that assemble products with different assembly structures (i.e. one product consisting of one X-shaped part and d Y-shaped parts) or multiple quality attributes (i.e. geometric dimensions) can be considered. Thirdly, researchers can extend the model to lines with batch machines and study good policies for batch-based matching in a selective assembly system.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the National Science Foundation of China (grant number 71871138, 71471114, 71432006, 72171144).

Notes on contributors

Xiaoxiao Shen is a PhD student at Department of Industrial Engineering and Management, Shanghai Jiao Tong University, Shanghai, China. Her current research interests include modelling and analysis of decision and optimisation problems in manufacturing and healthcare systems. Before her doctoral studies, Xiaoxiao received a B.S. degree in industrial engineering from Nanjing Agricultural University, Nanjing, China.

Na Li is an Associate Professor with the Department of Industrial Engineering and Management, Shanghai Jiao Tong University. Prof. Li received her B.S. degree in mechanical engineering and automation from Xi'an Jiaotong University, Xi'an, China, in 2003, and Ph.D. degree in industrial engineering from Tsinghua University, Beijing, China, in 2008. Her research interests include stochastic modelling in the application areas of production and the healthcare delivery system.

References

Chan, Ka Ching, and Richard J. Linn. 1998. "A Grouping Method for Selective Assembly of Parts of Dissimilar Distributions." *Quality Engineering* 11 (2): 221–234.

Ching, ShiNung, Semyon M. Meerkov, and Liang Zhang. 2008. "Assembly Systems with Non-Exponential Machines: Throughput and Bottlenecks." *Nonlinear Analysis: Theory, Methods & Applications* 69 (3): 911–917.

Clottey, Toyin, and W. C. Benton. 2020. "Sharing Quality-Distribution Information for the Selective Assembly of Intermediary Components in the Automotive Industry." *Production and Operations Management* 29 (1): 174–191.

Colledani, Marcello, Dariush Ebrahimi, and Tullio Tolio. 2014. "Integrated Quality and Production Logistics Modelling for the Design of Selective and Adaptive Assembly Systems." *CIRP Annals* 63 (1): 453–456.

Coullard, C. R., A. B. Gamble, and P. C. Jones. 1998. "Matching Problems in Selective Assembly Operations." *Annals of Operations Research* 76 (0): 95–107.

Dover, Omri, and Dvir Shabtay. 2016. "Single Machine Scheduling with Two Competing Agents, Arbitrary Release Dates and Unit Processing Times." *Annals of Operations Research* 238 (1-2): 145–178.

Feng, Yuan, Xiang Zhong, Jingshan Li, and Wenhui Fan. 2018. "Analysis of Closed-Loop Production Lines with Bernoulli Reliability Machines: Theory and Application." *IISE Transactions* 50 (3): 143–160.

He, Cheng, Joseph Y. T. Leung, Kangbok Lee, and Michael L. Pinedo. 2016. "Improved Algorithms for Single Machine Scheduling with Release Dates and Rejections." *4or* 14 (1): 41–55.

Jeevanantham, A. K., S. V. Chaitanya, and A. Rajeshkannan. 2019. "Tolerance Analysis in Selective Assembly of Multiple Component Features to Control Assembly Variation Using Matrix Model and Genetic Algorithm." *International Journal of Precision Engineering and Manufacturing* 20 (10): 1801–1815.

Jia, Zhiyang, Liang Zhang, Jorge Arinez, and Guoxian Xiao. 2016. "Performance Analysis of Assembly Systems With Bernoulli Machines and Finite Buffers During Transients." *IEEE Transactions on Automation Science and Engineering* 13 (2): 1018–1032.

Ju, Feng, Jingshan Li, and Weiwen Deng. 2017. "Selective Assembly System with Unreliable Bernoulli Machines and Finite Buffers." *IEEE Transactions on Automation Science and Engineering* 14 (1): 171–184.

Ju, Feng, Jingshan Li, Guoxian Xiao, Ningjian Huang, and Stephan Biller. 2014. "A Quality Flow Model in Battery Manufacturing Systems for Electric Vehicles." *IEEE Transactions on Automation Science and Engineering* 11 (1): 230–244.

Kannan, Sm, and V. Jayabalan. 2001. "A New Grouping Method to Minimize Surplus Parts in Selective Assembly for Complex Assemblies." *International Journal of Production Research* 39 (9): 1851–1863.

Kannan, Sm, V. Jayabalan, and K. Jeevanantham. 2003. "Genetic Algorithm for Minimizing Assembly Variation in Selective Assembly." *International Journal of Production Research* 41 (14): 3301–3313.

Koskinen, Jani, Csaba Raduly-Baka, Mika Johnsson, and Olli S. Nevalainen. 2020. "Rolling Horizon Production Scheduling of Multi-Model PCBs for Several Assembly Lines." *International Journal of Production Research* 58 (4): 1052–1073.

Lanza, Gisela, Benjamin Haefner, and Alexandra Kraemer. 2015. "Optimization of Selective Assembly and Adaptive Manufacturing by Means of Cyber-Physical System Based Matching." *CIRP Annals* 64 (1): 399–402.

Lee, Jun-Ho, Jingshan Li, and John A. Horst. 2018a. "Serial Production Lines with Waiting Time Limits: Bernoulli Reliability Model." *IEEE Transactions on Engineering Management* 65 (2): 316–329.

Lee, Jun-Ho, Cong Zhao, Jingshan Li, and Chrissoleon T. Papadopoulos. 2018b. "Analysis, Design, and Control of Bernoulli Production Lines with Waiting Time Constraints." *Journal of Manufacturing Systems* 46: 208–220.

Li, N., Z. Jiang, G. Liu, and Z. Zhang. 2012. "Analysis of Quality-Caused Re-entrance Electrical Test System in Semiconductor Manufacturing by Markov Method." *International Journal of Production Research* 50 (12): 3486–3497.

Li, Jingshan, and Semyon M. Meerkov. 2009. *Production Systems Engineering*.

Liu, Shaogang, and Longhui Liu. 2017. "Determining the Number of Groups in Selective Assembly for Remanufacturing Engine." *Procedia Engineering* 174: 815–819.

Liu, Zhenyu, Zhang Nan, Chan Qiu, Jianrong Tan, Jingsong Zhou, and Yao Yao. 2019. "A Discrete Fireworks Optimization Algorithm to Optimize Multi-Matching Selective Assembly Problem with Non-Normal Dimensional Distribution." *Assembly Automation* 39 (2): 323–344.

Liu, LiPing, Fang Zhu, Jie Chen, YiZhong Ma, and YiLiu Tu. 2013. "A Quality Control Method for Complex Product Selective Assembly Processes." *International Journal of Production Research* 51 (18): 5437–5449.

Mansoor, E. M. 1961. "Selective Assembly – Its Analysis and Applications." *International Journal of Production Research* 1 (1): 13–24.

Mease, David, Vijayan N. Nair, and Agus Sudjianto. 2004. "Selective Assembly in Manufacturing: Statistical Issues and Optimal Binning Strategies." *Technometrics* 46 (2): 165–175.

Rezaei Aderiani, Abolfazl, Kristina Wärmefjord, Rikard Söderberg, and Lars Lindkvist. 2019. "Developing a Selective Assembly Technique for Sheet Metal Assemblies." *International Journal of Production Research* 57 (22): 7174–7188.

Su, W., X. Xie, J. Li, L. Zheng, and S. Feng. 2017. "Reducing Energy Consumption in Serial Production Lines with Bernoulli Reliability Machines." *International Journal of Production Research* 55 (24): 7356–7379.

Suwa, H. 2007. "A New When-to-Schedule Policy in Online Scheduling Based on Cumulative Task Delays." *International Journal of Production Economics* 110 (1): 175–186.

Tan, Matthias H. Y., and C. F. Jeff Wu. 2012. "Generalized Selective Assembly." *IIE Transactions* 44 (1): 27–42.

Vélez-Gallego, Mario C., Jairo Maya, and Jairo R. Montoya-Torres. 2016. "A Beam Search Heuristic for Scheduling a Single Machine with Release Dates and Sequence Dependent Setup Times to Minimize the Makespan." *Computers & Operations Research* 73: 132–140.

Wang, Jun-Qiang, Fei-Yi Yan, Peng-Hao Cui, Tian Xia, Fu-Dong Cui, and Sicco Verwer. 2018. "Modeling and Analysis of non-Homogenous Fabrication/Assembly Systems with Multiple Failure Modes." *The International Journal of Advanced Manufacturing Technology* 94 (9-12): 3309–3325.

Wang, Jun-Qiang, Fei-Yi Yan, Peng-Hao Cui, and Chao-Bo Yan. 2019. "Bernoulli Serial Lines with Batching Machines: Performance Analysis and System-Theoretic Properties." *IISE Transactions* 51 (7): 729–743.

Yan, Fei-Yi, Jun-Qiang Wang, Yang Li, and Peng-Hao Cui. 2021. "An Improved Aggregation Method for Performance Analysis of Bernoulli Serial Production Lines." *IEEE Transactions on Automation Science and Engineering* 18 (1): 114–121.

Yuan, J. J., C. T. Ng, and T. C. E. Cheng. 2015. "Two-Agent Single-Machine Scheduling with Release Dates and Preemption to Minimize the Maximum Lateness." *Journal of Scheduling* 18 (2): 147–153.

Zhang, G. 2001. "An on-Line Bin-Batching Problem." *Discrete Applied Mathematics* 108 (3): 329–333.

Zhang, Ya-Jun, Li Liu, Ningjian Huang, Robert Radwin, and Jingshan Li. 2021. "From Manual Operation to Collaborative Robot Assembly: An Integrated Model of Productivity and Ergonomic Performance." *IEEE Robotics and Automation Letters* 6 (2): 895–902.