



POLITECNICO
MILANO 1863

POLITECNICO DI MILANO

Department of Management, Economics and Industrial Engineering

Master Degree in Management Engineering

Impact of output control mechanism through labour flexibility within workload control theory

Supervisor: Prof. Alberto Portioli Staudacher

Assistant supervisor: Eng. Federica Costa

MASTER THESIS

Pietro Magni 854258

Matteo Mascellani 858865

Academic year 2016 / 2017

TABLE OF CONTENT

1. ABSTRACT IN ENGLISH	5
1. ABSTRACT IN ITALIANO	6
2. INTRODUCTION	7
2.1 BACKGROUND	7
2.2 OBJECTIVE OF THE THESIS	8
2.3 RESEARCH METHODOLOGY AND THESIS OUTLINE	9
3. LITERATURE REVIEW	11
3.1. WORKLOAD CONTROL AND ORDER REVIEW AND RELEASE SYSTEMS	11
3.2. WORKLOAD CONTROL AND CAPACITY ADJUSTMENTS (OUTPUT CONTROL)	20
3.3. WORKER FLEXIBILITY	23
4. OBJECTIVES, RESEARCH METHODOLOGY AND RESEARCH FRAMEWORK	32
4.1. RESEARCH GAP	32
4.2. RESEARCH QUESTIONS	32
4.3. SIMULATION MODEL	33
4.4. PERFORMANCE MEASUREMENT	39
4.5. CONFIGURATION OF EXPERIMENT	39
5. ANALYSIS OF RESULTS	48
5.1. RESEARCH QUESTION 1	48
5.2. RESEARCH QUESTION 2	55
5.3. RESEARCH QUESTION 3	67
6. DISCUSSION AND CONCLUSIONS	78
7. REFERENCES	83
8. APPENDIX	92

LIST OF FIGURES

Figure 1 Land framework for output control (Thurer 2016)	22
Figure 2 Park 1989, flexibility matrices	25
Figure 3 Bobrowski, flexibility matrix with heterogeneous efficiency	29
Figure 4 Simulation model, schematic representation	33
Figure 5 Flexibility matrices	40
Figure 6 Homogeneous worker efficiency	41
Figure 7 Heterogeneous efficiency matrices	43
Figure 8 Warm up period with labour flexibility = 2 (R1)	47
Figure 9 Warm up period with labour flexibility = 3	47
Figure 10 Percentage GTT variation for every workload norm due to increasing level of worker flexibility respect to R0	48
Figure 11 Incremental flexibility on different saturation levels. Comparison with R0	50
Figure 12 Incremental flexibility on different saturation levels. Comparison with previous flexibility level	50
Figure 13 SFT: percentage variation respect to R0 with increasing flexibility	52
Figure 14 Tardiness: percentage variation respect R0 with increasing flexibility	52
Figure 15 Tardy: percentage variation respect R0 with increasing flexibility	52
Figure 16 GTT for increasing variability	53
Figure 17 SFT for increasing variability	54
Figure 18 Tardiness for increasing variability	54
Figure 19 Tardy for increasing variability	54
Figure 20 GTT trend with increasing efficiency for decentralized rules	55
Figure 21 GTT trend with increasing efficiency for centralized rules	55
Figure 22 GTT percentage improvement with increasing efficiency	56
Figure 23 Delta variation of GTT	57
Figure 24 SFT variation with increasing efficiency	58
Figure 25 Tardiness variation with increasing efficiency	58
Figure 26 Tardy variation with increasing efficiency	58
Figure 27 Shop performance with increasing saturation	59
Figure 28 Shop performance with increasing variability	60
Figure 29 Sum of the GTT reduction caused by flexibility and efficiency improvement separately taken compared with GTT reduction for flexibility and efficiency together. The starting point is R1 11%	61
Figure 30 Sum of the GTT reduction caused by flexibility and efficiency improvement separately taken compared with GTT reduction for flexibility and efficiency together. The starting point is R2 11%	62
Figure 31 Sum of the GTT reduction caused by flexibility and efficiency improvement separately taken compared with GTT reduction for flexibility and efficiency together. The starting point is R2 42%	62
Figure 32 Sum of the GTT reduction caused by flexibility and efficiency improvement separately taken compared with GTT reduction for flexibility and efficiency together. The starting point is R2 67%	62
Figure 33 Heterogeneous efficiency matrices. Coefficients for processing time reduction.	63
Figure 34 Comparison between heterogeneous and homogeneous efficiency pattern	64
Figure 35 Tardiness comparison of homogeneous and heterogeneous efficiency patterns	64
Figure 36 Tardy comparison of homogeneous and heterogeneous efficiency patterns	65
Figure 37 GTT-SFT comparison of homogeneous and heterogeneous efficiency patterns with increasing saturation	66

Figure 38 GTT-SFT comparison of homogeneous and heterogeneous efficiency patterns with increasing variability	66
Figure 39 Centralized and decentralized rules comparison of GTT-SFT. Low saturation.	68
Figure 40 Centralized and decentralized rules comparison of GTT-SFT. High saturation.	68
Figure 41 Centralized and decentralized rules comparison of Tardiness with high saturation.....	69
Figure 42 Centralized and decentralized rules comparison of Tardy with high saturation	70
Figure 43 Centralized and decentralized rules comparison of GTT-SFT with low efficiency	71
Figure 44 Centralized and decentralized rules comparison of Tardiness with low efficiency	72
Figure 45 Centralized and decentralized rules comparison of Tardy with low efficiency	72
Figure 46 Centralized rules comparison of GTT-SFT with high saturation	73
Figure 47 Centralized rules comparison of Tardiness with high saturation	74
Figure 48 Centralized rules comparison of Tardy with high saturation.....	74
Figure 49 Comparison of PRD and SPT dispatching rules. GTT_SFT performances in different conditions	75
Figure 50 Comparison of PRD and SPT dispatching rules. Tardiness and Tardy performances with low efficiency.....	76
Figure 51 Comparison of PRD and SPT dispatching rules. Tardiness and Tardy performances with high variability	76
Figure 52 Comparison of PRD and SPT dispatching rules. Tardiness and Tardy performances with high saturation	77

LIST OF TABLES

Table 1 ORR classification 19

Table 2 Aggregate limiting description 35

Table 3 Operator efficiency and processing time reduction relationship 42

Table 4 Description of tested rules 44

Table 5 Experimental work, list of parameters 45

Table 6 GTT improvement respect to R1 11% 61

Table 7 Centralized and decentralized "When" rules summary..... 67

1. ABSTRACT IN ENGLISH

Make To Order production planning is a challenge for managers searching for performances improvements in the shop floor. The inherent characteristics of the product make traditional lean management approach insufficient for the purpose. Workload Control methodology offers a valuable answer for this kind of environment: it is capable of speeding up the flow of products and decrease variability. Workload control can lever on two drivers: input and output control. Order Review and Release is an input control methodology, which has been vastly researched. Less research has focused on output control and on the combination of the two. Within Workload Control theory, this Thesis aims at investigating the different patterns of worker efficiency and flexibility, the different triggers for operator reallocation and their effects on production flow within different external conditions and supporting different managerial decisions. The large number of variables taken into account increases realism to the simulation and offers a valuable theoretical approach for managers facing these problems in the real world. Following the approach of many researchers, the shop is modelled through a simulation using Python. The results show that both efficiency and flexibility are beneficial for the shop performances. Efficiency has a higher impact with high flexibility. Additionally, there is a synergy of efficiency and flexibility which gives best results when they are combined respect to the sum of the effects of the two variables improved separately. Incrementing flexibility reduces the impact of an increasing variability. The best trigger for worker reallocation is also evaluated according to the different external conditions.

Keywords: workload control, worker flexibility, worker efficiency, input control, output control

1. ABSTRACT IN ITALIANO

Il Make To Order è un sistema di pianificazione della produzione che costituisce una sfida per i manager alla ricerca di miglioramenti delle prestazioni nei sistemi industriali. Le caratteristiche intrinseche del prodotto rendono insufficiente l'approccio tradizionale di gestione mediante le classiche metodologie di lean management. La metodologia Workload Control offre una risposta valida per questo tipo di contesto: è in grado di velocizzare il flusso dei prodotti e diminuire la variabilità. Il workload control può fare leva su due azioni: input e output control. L'Order Review and Release è una metodologia di input control, ampiamente studiata. Meno ricerche si sono concentrate sul controllo dei risultati e sulla combinazione delle due metodologie. All'interno della teoria del Workload Control, questa tesi mira a indagare i diversi modelli di efficienza e flessibilità degli operatori, i diversi trigger per la riallocazione dell'operatore e i loro effetti sul flusso di produzione in diverse condizioni esterne e supportare così le decisioni manageriali. Il gran numero di variabili prese in considerazione aumenta il realismo della simulazione e offre un valido approccio teorico ai manager che affrontano questi problemi nel mondo reale. Seguendo il metodo di molti ricercatori, il sistema produttivo è modellato attraverso una simulazione Python. I risultati dimostrano che sia l'incremento di efficienza che di flessibilità degli operatori migliorano le prestazioni. L'efficienza ha un impatto maggiore in contesti di elevata flessibilità. Inoltre, si rileva una sinergia tra efficienza e flessibilità per cui l'effetto della combinazione delle cause è maggiore rispetto alla somma degli effetti delle due variabili separate. L'aumento della flessibilità riduce in maniera significativa l'impatto di una variabilità crescente. Il miglior trigger per la riallocazione degli operatori viene anche valutato in base alle diverse condizioni esterne.

2. INTRODUCTION

2.1 BACKGROUND

Make To Order (MTO) is a manufacturing management system where the production of a part only starts when the order is received by the company. This results in an early Order Penetration Point (OPP) which is the point at which a product is associated with a specific order (Olhager, 2003). As a result, the matching of demand and capacity is challenging for production planning. The ever growing complexity of products is pushing companies towards a more flexible and efficient shop floor organization. The production system is severely impacted by the product features:

- Demand is very variable both in terms of total volume and in terms of variety (Lander & Liker, 2007)
- Product Bill Of Material is complex (Hicks & Braiden, 2000). Several levels and a large variety of possibility are the result of the customization degree required in order to satisfy customers' needs
- Production lead time is long (Cutler, 2005), variable and depend on the shop workload
- Product and demand variability require highly skilled workers and flexible machines (White and Prybutok,2001)

Job shop is the most used production system layout since it allows more flexibility. On the other hand, it is difficult to evaluate the actual production capacity and to understand where the bottlenecks are. Consequently, MTO companies find it difficult to estimate production lead time, to forecast delivery date for customer orders and to be punctual on delivery. These capabilities are highly valued by customers and makes them order winning performances. As a result, planning managers are pushed to release the highest possible level of workload in the shop causing a higher level of WIP, increasing lead times and late delivery (Portioli and Tantardini, 2012).

Common lean manufacturing techniques implementation is not suited for this environment and WorkLoad Control (WLC) approach is used. WLC is a planning tool which has proved to be very effective in high variety and low volume environment. WLC uses a so-called Pre Shop Pool (PSP) where orders are stored before entering in the shop. The release mechanism is governed by Order review and Release (ORR) methodology, which is mainly composed by two

decisions: a sequencing decision which defines the sequence for reviewing orders and a releasing decision that outlines the order selection criteria. The main results brought by ORR implementation are the reduction of the total lead time, of the WIP level alongside with an improvement of order due date compliance.

Output control is a method for improving the actual capacity of the shop. Several decisions can be included herein:

- Worker reallocation
- Production capacity oversize
- Outsourcing
- Overtime

2.2 OBJECTIVE OF THE THESIS

The overall objective of the Thesis is to identify the relationship between workers' capabilities and training patterns with shop performances with respect to different possible external conditions. Following the already existing literature on the topic, three research questions were considered:

1. *What is the impact of incremental flexibility on average orders gross throughput time (GTT), shop floor time (SFT), tardiness and number of tardy orders, when combined with an ORR method for input control?*
2. *How do different levels of operators cross-training and efficiency affect shop performance? What is the impact of considering a heterogeneous flexibility pattern for operator cross training?*
3. *Which "When" rule, decentralised rule transferring operators when idle or centralised rule allowing the transfer on the basis of the queue length, is most performing? And how do shop performances react to different load threshold when a decentralised "When" rule is selected?*

2.3 RESEARCH METHODOLOGY AND THESIS OUTLINE

A simulation model has been used to test the hypothesis and record system performances under every environmental condition and according to all the possible managerial decisions taken into consideration in this study. The model simulates a job shop with five work stations and five workers. The release of the orders to the shop floor is managed and controlled through an ORR mechanism. After arrival, orders are stored in a PSP before being released to the shop floor. The release is controlled by the ORR method according to the aggregate workload already released into the shop floor. Workers can be transferred among workstations according to their cross-training level. Operators' movements to another station yields a processing time reduction of the order that is being processed at that station.

Stochastic orders' arrival rate and processing time at each workstation are considered in order to create more realistic experimental condition. Job shop production, indeed, is usually characterised by different job types requiring different processing, routing within the shop floor and production time. Stochastic arrivals and processing times allow to include this peculiarity in the model.

Four system performances are used to evaluate and compare the variables tested. Two are concerned with orders flow time. The first is the Gross Throughput Time (GTT) which is the average time between order's arrival in the pre shop pool and the final departure of the finished product; the second is Shop Floor Time (SFT) which represents the time spent in the shop after order's release, which is also a good indicator for average work in process. The other two performances are related to the ability of the system to respect due dates and deliver on time: average tardiness and average number of tardy orders per day.

After this brief introductory section presenting the main characteristics and contents of the study, in section 3 a thorough literature review is presented. The first main topic addressed in the literature review is the concept of workload control (input control). An exhaustive description of ORR mechanisms presented in previous researches is reported including the ORR classification by Bergamaschi et al. (1997). A description of the main methods concerning output control studied in the literature is then introduced, with a particular attention to those researches implementing output control in combination with some order release and input control mechanism. The final paragraph of section 3 is entirely dedicated to worker flexibility

as output control method and to the various aspects affecting the effectiveness of this solution considered in past studies.

Section 4 introduces the research gap and the research questions that are addressed by this study. Then a detailed description of the simulation model developed and used for the aim of the research is made followed by the configuration of experiments carried out with all the parameters and variables considered.

In section 5, results are grouped according to each research question. Results are extensively discussed under every environmental condition and each possible managerial decision considered in the study in order to support the validity of the conclusions.

In section 6, results are summarized and the practical implication are discussed. Then, a brief conclusion is presented with the main findings of the research and the further possible future study in the field.

3. LITERATURE REVIEW

The articles cited in the following literature review have been collect by using www.scopus.com and searching “workload control” and “flexibility” as keywords. Relevant papers were selected and deepened. The references of these articles were useful in order to find more pertaining material for the literature review. Furthermore, other master thesis’ literature reviews were exploited as benchmark and as valuable sources of relevant papers.

3.1. WORKLOAD CONTROL AND ORDER REVIEW AND RELEASE SYSTEMS¹

Workload control (WLC) is a production control and planning technique which decouples job arrivals and planning phase from the production (Melnyk & Ragatz, 1989) and it has been developed primarily for high-variety, low-volume products (Silva, et al., 2015). It manages the conversion of production orders from the planning system to the execution phase. When orders are generated, either from a planning system or from customers’ orders, they are not released directly to the shop floor but they are stored in the pre-shop pool until the releasing conditions are satisfied (Bergamaschi, et al., 1997). Order release is a key component of the Workload Control concept, which includes two decisions: a sequencing decision that defines the sequence in which orders are reviewed to be released; and a selection decision that establishes the criteria for selecting orders for release (Thürer, et al., 2015). The usage of an ORR systems brings to several benefits conforming to lean management as they enable to “focus on, speed up and stabilize the flow” (Portioli- Staudacher & Tantardini, 2012). In fact, the key objectives of ORR are the control of WIP level and the workload balance among the stations, ensuring improved shop utilization and better delivery performances (Bergamaschi, et al., 1997). Thanks to the backlog of non released orders, the production is protected from the impact of demand variability and other external dynamics (Land & Gaalman, 1996) (Bertrand & Van Ooijen, 2002). The key success factor of an ORR system and the use of a pre-shop pool is the method used to select the orders to be released and it will determine the system’s performance (Tatsiopoulos, 1997) (Land & Gaalman, 1996) (Land, 2006) (Baykasoğlu & Gçken, 2010) (Lu, et al., 2010). As pointed out in literature, the ability to control the WIP level in the shop and to balance the workload among work centers, enables to reduce and

¹ Part of the following literature review is by Lo Cascio, Lo Cascio 2017

stabilize shop floor throughput times, reduce shop congestion and to estimate more reliable due dates ((Bechte, 1988) (Hendry & Wong, 1994) (Bergamaschi, et al., 1997) (Sabuncuoglu & Karapinar, 1999) (Breithaupt, et al., 2002) (Stevenson & Hendry, 2006) (Baykasoğlu & Gçken, 2010)). In addition, an effective order release system increases flexibility because by reducing the WIP it is possible to delay final decisions on production, reducing the impact of changes and specification modifications after order confirmation (Land & Gaalman, 1996) (Stevenson & Hendry, 2006). Since with a proper release rule the queues are shorter, the importance of the dispatching rule for the production's performances decreases, making it possible to adopt a simple first-come-first-served (FCFS) rule (Bechte, 1988) (Wein, 1988) (Land & Gaalman, 1996) (Kingsman, 2000). This allows to simplify the management and lower the congestion of the production system. Consequently, as highlighted by Bertrand and Van Ooijen (2002), it decreases the decision-making pressure on the operator. Furthermore, Bechte (Bechte, 1988) reports that a workload control system can also decrease the personnel involved in production planning and control by up to 40%. ORR has been largely developed in the literature. Bechte (1994) has reported that its use can lead to reducing WIP by more than 25%, lead times by 15%, and 20% in the percentage of tardy jobs. ORR systems consist of 4 main steps:

1. Customer enquiry phase
2. Order entry phase
3. Pre-shop pool
4. Order release phase

When the order release phase is initiated, an orders' subgroup enclosed in the pool is released and, once released, a job remains on the shop floor until all its operations have been completed. ORR set norms for the workload allowed on the floor. It scans all production orders stored in the pool and if a job does not suit in these norms, the release decision will hold it back (Oosterman, et al., 2000). In the first phase decisions about due date, job acceptance and rejection are taken. The second phase is the order entry (OE), phase in which the orders are prepared and introduced in the pre-shop pool. Furthermore, a check on material and tools availability is performed and, if necessary, engineering activities are carried out. In particular (Bergamaschi, et al., 1997):

- the job routing is recovered or defined and the availability of the required tooling, fixtures and CNC programs is checked;
- the required materials' picking list is created and the availability of the listed ones is checked;
- a delivery date is assigned

Most references assume that all the incoming orders will be accepted, nonetheless, some authors remove this assumption making a distinction between planned workload lengths, controlled at job entry phase and based on the workload of only accepted orders, and total workload lengths, controlled at the customer enquiry phase. The latter is based not only on the workload of the accepted orders, but it considers also a portion of unconfirmed orders based on order winning probability known as strike rate percentage (Hendry, et al., 2013) (Kingsman & Mercer, 1997) (Kingsman, et al., 1996). Once the availability of all the resources necessary to process the order is ensured, the latter can be discharged to the following phases. It is so stored in the pre-shop pool (PSP), which usually is a database (but can also be physical raw materials or just the order's paperwork) containing all the orders processed by the order entry phase but not conforming to the rules set by the ORR system so not released yet. All the orders must flow in the PSP in order to reach the shop floor (Hendry & Wong, 1994). The orders in the PSP are queued mostly according to some priority rules, which will be further explained in the following section. The last and most important phase is the order release phase. The main instrument to have a successful ORR strategy is the release decision, shifting an order from the pre-shop pool to the set of orders acceptable for production (Witte, et al., 2008). 'Triggering mechanism' or 'input control mechanism' are the name for the set of criteria used to define the orders to be released. There are three types of information that can be used (Bergamaschi, et al., 1997):

- current pre-shop pool status: number and kind of orders in the pre-shop pool
- current shop status: number and kind of orders in the shop floor, in which station they are queuing and current shop capacity
- planned shop performances: manufacturing lead times and delivery timeliness

The criterion for determining if and when the order has to be released is the level of workload, usually measured in units of processing time, at each station in the order routing sequence. The released workload for a station can be partitioned into two parts: a direct part (work from orders waiting to be processed at the considered station) and an indirect or upstream part (from jobs queuing at an upstream station) (Oosterman, et al., 2000). Literature on workload control systems can be divided into three categories:

- simulation studies
- other theoretical studies
- empirical studies

The broader category is the first one (Hendry, et al., 2013), among which many are focused on the order release stage ((Thürer, et al., 2012), (Perona & Portioli, 1998) (Oosterman, et al., 2000)) and other are focused on how the system can be developed looking at more realistic so more complex environments. For example, Lu et al. (2010) focused on the application of order release in a complex assembly job shop; Thurer et al. (2011a) through simulations compared two WLC approaches to establish which would be easier to implement in terms of the minor effort of determining the required parameters (Hendry, et al., 2013). The second category includes all the theoretical papers that instead of simulations use mathematical analysis as analysis tool. An example is given by Kingsman (2000) who developed a theory for workload control in a mathematical form to provide procedures for applying input and output control; Henrich et al. (2004) developed a framework to explore the applicability of WLC in MTO companies. Nonetheless, according to (MacCarthy, 2006) the increase of theoretical studies, including both simulations and other methods, widens the gap between theory and practice. The last category includes the studies that analyzed the use of WLC in practice. Several authors report that often the implementation of WLC systems leads to performances that differ from the ones obtained through simulations. (Stevenson, et al., 2005). Nonetheless, some empirical research papers report successful cases of WLC implementation ((Bechte, 1988) (Bechte, 1994) (Wiendahl, 1995) (Park, et al., 1999)). This last category can be further divided into four groups (Hendry, et al., 2008). First group is formed by empirical studies focused on the differences between alternative WLC and the results of their implementation, for example, Bertrand and Wortmann (1981), Bechte (1994) (1988) and Park

et al. (1999) report exceptional positive performances applying WLC. Another group is given by empirical studies that describe some aspects of the implementation process, like Hendry (1989) (1993) who investigated problems arising with the implementation. The third category includes empirical studies that developed an implementation strategy for WLC. For example, Fry and Smith (1987) proposed a procedure to implement WLC in six steps applying to the order release phase; Wiendahl(1995), instead, proposed a more complex WLC approach. The last category encompasses simulation studies considering also implementation issues. Examples are given by Henrich (2004) (2005) and Wiendahl (1995).

ORR classification

Numerous authors have classified and investigated the characteristics of the ORR systems (for example, Philipoom (1993), Bergamaschi (1997) and Sabuncuoglu (1999)) because, as highlighted by Portioli (Portioli-Staudacher & Tantardini, 2012), ORR systems are articulated and quite complex. In particular, we will focus on the classification made by Bergamaschi (Bergamaschi, et al., 1997). He classifies ORR systems, focusing only on the Order Releasing phase, on eight different axes that describe the key characteristics and properties of an ORR system. A brief description of all the eight dimensions proposed by Bergamaschi follows.

1. Order release mechanism:

According to the mechanism used to release the orders, OR systems can be divided into two major groups: load limited and time phased methodologies. The former approach is the most used among the two. It releases orders every time period so the only decision to be taken is which order in the pre-shop pool should be released. It is based on information coming from the shop floor as well as job's features. The latter, on the contrary, computes a release time for each order and release when the planned time is reached. Compared to the load limited methodology, it is much easier and usually based only upon job's information (e.g. due-date, work content, routing) without considering the shop load. The former approach is the most used because it allows to balance and limit the workload more easily bringing to a more controlled level of WIP.

2. Timing convention:

The time in which an order can be released can be continuous or discrete. When a continuous timing convention is considered, an order can be released at any time during the system's operation, while, under discrete timing convention it may happen only at periodic intervals (e.g. beginning of shifts). The former approach requires a continuous control resulting less simple than the latter, that is why most authors refer to the discrete timing convention (Ragatz & Mabert, 1988) (Melnik & Ragatz, 1989) (Glassey & Resende, 1988). Continuous approach is used in a scarce number of papers, among which the most notable are by (Melnik & Ragatz, 1989) (Hendry & Wong, 1994) (Sabuncuoglu & Karapinar, 1999) while most studies adopted the discrete method (e.g. (Cigolini & Portioli-Staudacher, 2002) (Land & Gaalman, 1996) (Oosterman, et al., 2000) (Sabuncuoglu & Karapinar, 2000)

3. Workload measure:

There are two main methods to express workload: number of jobs on the shop floor and work quantity in terms of hours or percentage of planned capacity in a certain period. Since job shops productions are characterized by high processing time variability among the different products, the first method is rarely adopted.

4. Aggregation of workload measure:

Based on the level of aggregation, it is possible to divide this category into multiple subgroups. At one extreme is the total shop load in which no information on how load is allocated among the work centers is given. An example is given by Ragatz and Mabert (1988) according to which orders are released when the total uncompleted work in the shop goes below a predetermined threshold. This method is incomplete because it does not consider the presence of bottleneck(s). A solution is given by controlling the load only for the bottleneck work center. On the other extreme, the load can be computed for each work center, bringing to a more effective control but requiring data collection and reporting for each work center. Ragatz and Mabert (1988) stated that the former approach, even if significantly easier, is less effective than the latter.

5. Workload accounting over time:

Based on the chosen aggregation of workload measure, the difficulty in the measurement of the load is different. Indeed, for the total shop load, it is easy to measure it while, when load is computed for each work center, three factors have to be considered to calculate the amount of work that is expected for the forthcoming period (Bechte, 1988): queue at each center (on hand), work that will come from upstream machines (in transit) and load in PSP that will be released. There are three methods to account load over time:

- Atemporal approach: no indication of how load is distributed over time is provided so total work for each machine is given by the sum of processing times for all jobs in the shop that will be processed by that machine (no differentiation between on hand, in transit and released load)
- Time bucketing approach: time is divided into frames like weeks and the load of the work center since a specific job is assigned to the bucket (frame) according to the schedule. It considers only load on hand and there is the need of a scheduling activity.
- Probabilistic approach: released load and load in transit are multiplied by the probability that each job arrives at the current work center in that planning period. It provides the system with a greater robustness against perturbations such as breakdowns, scraps etc.

6. Workload control:

Four methodologies are mainly used to keep the workload under control: upper workload bound (or load limit), lower workload bound, Upper and lower workload bound and workload balancing. According to the first approach, the order is released only if it does not exceed the upper limit level, which is chosen according to the measure of aggregation. To ensure a buffer to each station, an additional lower bound can be used, while, if there is a station with a very short idle time, only a lower workload bound should be adopted. The use of both an upper and a lower bounding can reduce the risk of starvation and at the same time it prevents bottlenecks (Stevenson & Hendry, 2006). The last approach is an indirect limitation of load. Indeed, it is an explicit workload balancing method, the aim of which is to reduce the sum of deviations from aggregate balance of each station. According to these methods, orders are selected from the pre shop pool with the aim of minimizing the deviation between the

availability and the load of the work center, avoiding the starvation of the latter and improving the predictability of the throughput times (Land & Gaalman, 1996). With this approach, orders could be released even if the work center is slightly overloaded, the limit can be exceeded if it will improve future performances. The majority of ORR systems adopt a limiting methodology (Portioli-Staudacher & Tantardini, 2012), nonetheless, Germs and Riezebos (2010) stated that the release system has to improve also the balance of workload on the shop floor in order to be advantageous. In the Workload balancing method, Portioli and Tantardini develop a new way of balancing which tries to release a similar amount of workload to each work center (Portioli-Staudacher & Tantardini, 2012).

7. Capacity planning:

There are two main approaches regarding this dimension: active and passive capacity planning. According to the former one, the ORR method regulates machine capacity during system's operation by assigning overtime or by reallocating operators to machine centers. On the contrary, the latter one implies that the capacity is assumed as given by the planning stage and the ORR strategy is not able to control it.

8. Schedule visibility:

This dimension defines the amount of information given at the order release phase about future planned orders. To this regard, the system can have a limited visibility, in which the release of orders is oriented at controlling workload level in the shop during next closest planning period without considering how following periods could be affected. By contrast, if order release is oriented towards general optimization of performances along longer time horizon than the mere following period, then it is called extended visibility. In this case, the orders are selected from the pre-shop pool in order to maintain a balanced workload not only among work centers but also with time.

Table 1 ORR classification

Dimensions	Options
Order release mechanism	Load limited Time phased
Timing convention	Continuous Discrete
Workload measure	Number of jobs Work quantity
Aggregation of workload measure	Total shop load Bottleneck load Load by each workcenter
Workload accounting over time	Atemporal Time bucketing Probabilistic
Workload control	Upper bound only Upper and lower bounds Workload balancing Lower bound only
Capacity planning	Active Passive
Schedule visibility	Limited Extended

Sequencing

The orders in the pre-shop pool are queued mostly according to some priority rules, as for example: earliest due date (Ragatz & Mabert, 1988), earliest release date (Bechte, 1988) (Portioli, 1991) (Perona & Portioli, 1996), capacity slack based rule (Philipoom, et al., 1993), critical ratio (Bobrowski, 1989). An exception is given by Baker (Baker, 1974) (Baker, 1984) who uses the first come first served rule. Two different sequencing rules will be considered: Shortest Processing Time (SPT) and Planned Release Date (PRD), which in this thesis due to the configuration of the model corresponds to a FCFS approach. The former rule prioritizes jobs with the shortest processing time in order to allow a quick replenishment at downstream station and avoid starvation. The latter one, instead, considers the arrival date of the job. Starting from the jobs that arrived earlier, a job is released if it satisfies the release rule, otherwise a subsequent job is considered, this process is reiterated for all the jobs within the time limit.

3.2. WORKLOAD CONTROL AND CAPACITY ADJUSTMENTS (OUTPUT CONTROL)

Output control is the set of actions and drivers that a company can put in place in order to adjust capacity and match demand. In WLC literature, the focus is on accounting the workload while the description of capacity is simplified to standard output rates (Yuan 2017). However, real production capacity may be very different from the standard one. Moreover, Yuan (2017) reports that it is common among companies not to have a precise quantification of production capacity. As a result, it is difficult to estimate real production lead times and consequently, it is hard to comply with orders due date.

Resource constraints determine overall production capacity. Yuan defines three resources

- Machine
- Manpower
- Subcontracting

Each resource can be a lever for controlling output, as follows:

1. Working overtime: the station can be operated for a number of hours of overtime in addition to the normal shift
2. Reallocating operators: the work center may have several machines, not all of which are usually in operation. Operators can be reallocated to work on the same station, which increases the amount of processing work that can be done per day
3. Subcontracting: the work is done by a sub-contractor

While input control has been vastly examined, only recently has research emerged that includes output control and uses WLC theory to guide the output control decision (Thurer 2016). Kingsman and Hendry (2002) operationalized input control by rejecting orders that did not fit within a predetermined maximum workload. However, it is rather unlikely for companies to reject orders. Instead they will try to extend capacity or at most to delay due date.

Felan 1993 starts from the consideration that most research work has been performed on operator flexibility while it is common practice in Japanese companies following JIT

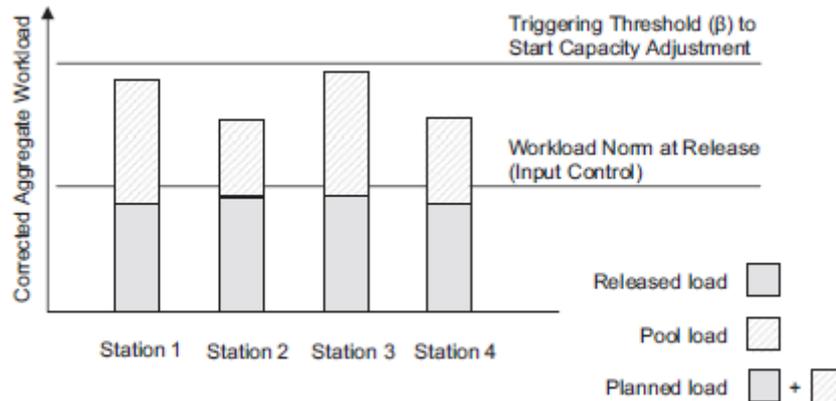
production system to oversize total capacity in order to keep up with demand peaks and avoid overproduction. Felan 1993 compares the two possible scenarios. By evaluating the results of the two possibilities with two types of criteria (cost criteria and non-cost criteria), results show that labour flexibility is more advantageous from cost criteria point of view, while increased staffing levels gives best results in term of flow time, tardiness and tardy orders. Felan (1993) suggests that it is also possible to combine the two strategies.

Kingsman (2000) states that output control considered as the ability to respect order due date and reduce queues and lead times is linked with the ability to release a balanced amount of work. Furthermore, Kingsman and Hendry (2002) describe the Lancaster University Management School (LUMS), which manages the total amount of work in the system so that it can be all completed within a preset maximum time limit Backlog Length (BL, the total amount of planned workload divided by the planned capacity). It maintains a pool of unreleased orders, which are only released if they would not cause an excess of queues' length. Through a simulation the model is tested in all the combination of presence of input and output control. (Output control is tested as: a) operator reallocation (operator flexibility), b) overtime, c) operator reallocation and overtime). The use of output control is triggered anytime the release of an order overcomes the BL limit and therefore additional capacity is needed. The results show that the LUMS provides a dynamic planning provision of extra capacity for bottleneck process. In MTO environment these bottlenecks are "wandering" which means that change from one station to another due to changing workloads. The effect of input control is to reduce lead times by reducing the accepted workload, while output control brings to lower lead time for the same workloads compared to input control alone.

Moreira and Alves (2006) combined input and output control in a job release mechanism called PIOC (Proposed Input-Output Control). The PIOC includes information about the jobs, the shop floor and the shop capacity, increasing capacity if necessary and defining the trigger for shop releasing. The output control is performed by setting an upper limit on the workload of the shop and by computing the workload corresponding to the jobs in the pre-shop pool. If the computed workload is above the upper limit then a decision to increase short-term capacity is made (like hiring temporary workers, working a second shift or overtime). PIOC improves performances especially when due dates are not tightly defined. The most visible measure improvement is mean tardiness, percent tardy, and mean queue time in shop floor.

Land (2015) proposed a framework where output control is performed when a predetermined threshold in station workload is overcome.

Figure 1 Land framework for output control (Thurer 2016)



Land (2015) also considers three parameters to guide the capacity adjustments:

1. The size of processing time reduction
2. The load threshold that triggers the commencement of the capacity adjustment
3. Percentage points below the triggering threshold at which the station returns to normal capacity conditions

However, Land 2015 list the various options for temporary capacity adjustments, however they state they are not interested in the specific adjustment mechanism, but rather in the performance impact.

Thurer et al (2012) refines LUMS and creates LUMS COR (Lancaster University Management School Corrected Order Release). LUMS COR incorporates a periodic release according to the corrected aggregate load approach, instead of the classical load approach ((Land and Gaalman 1996) the contributed load is divided by the position of a work center in the routing of a job) and a continuous release that in case a direct load of any work center falls to zero, the WLT pulls a job forward from the pool. The LUMS COR is compared to other well performing releasing methodologies: SLAR, WCPRD, FCFS, PST, ConWIP, Periodic and assessed through three categories of performance measures:

1. Traditional performance measures: Throughput time, WIP, reduction in percentage tardy

2. The robustness of release method respect to changes in flow characteristics as mean or variability
3. Practical use, which is the ease of implementation

LUMS COR proved to be the best performing in terms for category 1 measures, the most robust and the most intuitive to use.

This methodology proved to be more effective than lean management to face variability in small to medium sized MTO companies (Thurer et al 2014). Variability in this kind of environment cannot be eliminated via classic implementation of lean management due to the inherent characteristics of the demand. However, through WLC, it is possible to obtain shorter and more predictable lead times which is becoming an increasingly important criterion for customers selection and order winning.

Thurer 2016 follows Land 2015 approach. Little interest is given to the typology of capacity adjustment. Results show that input control has a stronger effect on the lead times and percentage tardy, while output control mostly impacts mean tardiness. Therefore, the two approaches shall be combined.

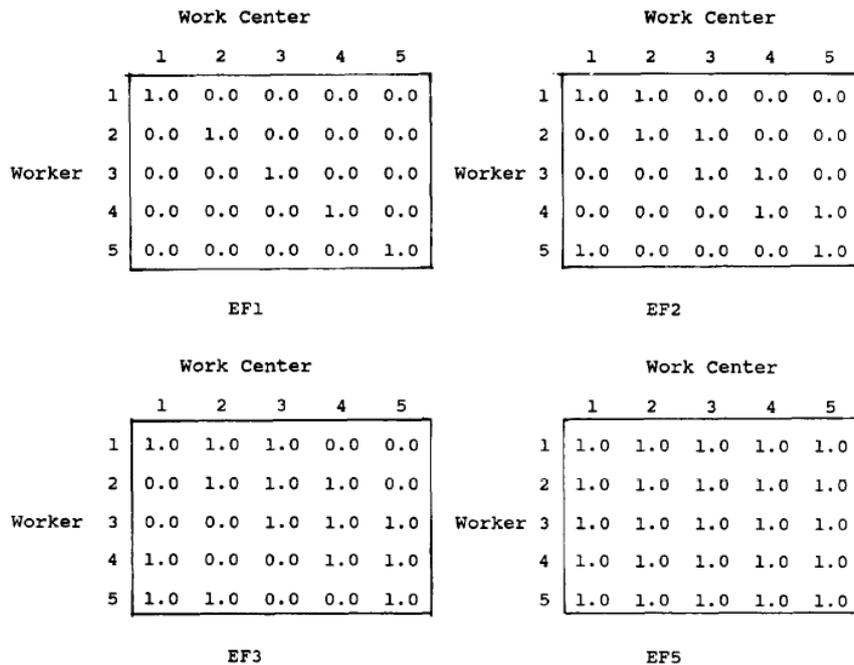
3.3. WORKER FLEXIBILITY

An alternative option to manpower staffing, overtime and capacity increase in output control is represented by worker flexibility. Worker flexibility is the ability of shop floor's operators to work in different work centers or on more than one machine. "A flexible workforce allows shop floor managers to move people around to respond to temporarily overloaded department. A higher level of staffing reduces the probability that any such department would become overloaded in the first place" (Felan 1993). Such a workforce gives to managers a greater range of possibility in staffing operators in different departments, according to company actual requirements, and facilitates the management of product mix changes enabling the transfer of operators to departments with high demand (Fry 1995). Whether all the workers are trained or only few selected operators are included in the program, achieving higher degree of worker flexibility requires the investment in cross-training the workforce and, in high worker attrition, manpower training may represent an on-

going expense rather than a one time initial cost. The cost of the training and the productivity losses due to the training period have to be considered (Kher 1994; Fry 1995). Therefore, cross-training people to learn more than one task is not an option that can be quickly implemented by companies to face periods of high load of work. For these companies needing timely solutions, the addition of labour capacity, through increasing staffing level or the exploitation of overtime, can represent the only possible way to face similar stressing situations. “For companies less willing to discard cost-based criteria or companies under less pressure to improve, increasing labour flexibility without changes in staffing levels may represent the best choice” (Felan 1993).

Despite the literature on worker flexibility in Dual Resource Constrained (DRC) systems is lacking, its beneficial effects on shops performance have been proven, enabling organization to exploit manpower in a more efficient way (Treleven 1989). A flexible workforce has a positive impact on a production system reducing Work In Process (WIP) inventory, reducing flow time and improving the performance in terms of average order lateness and average percentage of tardy orders (Park 1989; Park 1991; Felan 1993). Park (1989) introduced the concept of labour flexibility studying its effects on the performance of a job shop model composed of five work centers and five workers in conjunction with two different order release mechanism regulating the timing of the entrance of the orders to the shop floor. Four different configuration of labour flexibility were tested in combination with both the release mechanisms: backward infinite loading and forward finite loading. Matrices in Figure 2 show four configurations included in the study. $e_{ij} = 1$ represents the ability of the worker i to perform the work on work center j , while if $e_{ij} = 0$ worker i has not the required training to work on work center j .

Figure 2 Park 1989, flexibility matrices



The performance differences of the two release mechanisms, in terms of total costs including inventory holding, late penalty and worker transfer cost, in this model were not statistically significant. The introduction of flexibility by cross-training workers to operate efficiently at more than one work center, instead, showed to be beneficial. Besides, the introduction of the minimum level of worker flexibility (each operator trained to work at two different work center) determined the greater improvement in the performance of the model presented. Further efforts in cross training workforce to achieve higher degree of flexibility had a lower impact.

The sharp performance improvement due to the introduction of a minimum level of worker cross-training against the less significant impact of succeeding increase of worker flexibility was confirmed by further studies (Park 1991). Cross-trained workers' movements in the shop floor allows to respond to work overload that may occur at working stations determined by the type of work released to the shop floor and their sequencing. Therefore, increasing labour flexibility also reduces the impact on overall system results of different dispatching rules that affect work release sequence and may cause temporary bottleneck at some work center (Park 1991). This implies that in the presence of flexible workforce simple dispatching rules can be implemented in the place of more sophisticated ones resulting in similar overall results (Park 1991; Kher 2004).

Researches on labour flexibility mostly investigated the effect of cross-training workforce by considering a homogeneous distribution of skills and competences among all the operators. In the majority of studies, all workers are able to perform the job on the same number of work centers with the same maximum efficiency. It is very likely that in real production environments the actual situation of shop floor competences distribution would be quite different (Felan 2001). Operators might receive different level of training that enables them to perform different number of tasks according to production needs, to personal inclination to learning more than one job and personal aptitude. In their paper “Multi-level heterogeneous worker flexibility in a Dual Resource Constrained (DRC) job-shop” (2001), Felan and Fry introduced and analysed the performance of a DRC job shop with nine level of average labour flexibility and different combination of workers able to perform one single task with other multi-skilled workers who can move in the shop floor to work in more departments. The results of the study show that the introduction of labour flexibility is beneficial for the system. Increasing the level of flexibility improves the means of the monitored performance. Coherently with what reported before, the greater improvement is obtained shifting from no cross-training to a level of average worker flexibility of 2 (meaning that on average operators are able to work in 2 departments). Moreover, a flexibility of 1,7 defines a similar level of performance with a lower training cost. The major contribution of this study is that the configuration of the labour cross-training impacts significantly the results of the shop. A mix of cross-trained operators and operators with no flexibility gives better performance compared to a shop with equally trained workers. Dedicating some operators to one only department allows a higher specialization and a greater efficiency, while focusing high cross-training on few operators allows to achieve a good overall flexibility level at a lower training cost than giving the same level of cross-training to all the workforce.

When modelling systems where workers are required to run the machines and actively participate in the production, the inclusion of human behaviour in the model helps to more accurately predict system performance (Givi 2015). Transferring operators between departments and work centers can result in production delays and loss of production time; training periods are characterised by lower operator’s productivity and high attrition rates require a constant effort in training new workers (Kher 1994). In addition, higher level of flexibility requiring operators to alternatively perform many different tasks may cause the loss of the experience acquired in repeating an operation as soon as a transfer to another work

center is required (Givi 2015). The accumulation of fatigue has an impact on productivity, too (Givi 2015). Cross-training workers in more than one department proved to be effective in improving shops performance even when large transfer delays are present (Kher 1994). Inventory reduction and customer service improvement due to incremental labour flexibility were confirmed also in environments characterised by moderate attrition rate and labour losses caused by learning effects during cross-training (Malhotra 1993). However, with high levels of attrition and forgetting rate, learning and relearning losses do not allow workers to effectively learn two different tasks. Under these conditions, the main focus should be on retaining workers, reducing the attrition rate, rather than cross-training the workforce (Kher 1999). Givi (2015) investigated the effects of learning-forgetting and fatigue-recovery in a DRC system model concluding that a flexibility level of 3 (operators able to work on three work stations) yields better results than flexibility equal to 2 or 4.

The introduction of the concept of worker flexibility necessarily stimulated the definition and study of managerial rules aimed at the organization of the shop floor and at regulating the labour assignment to labour resources. These techniques are grouped in three sets of rules: “Where” rules, “When” rules and “Who” rules.

- “Where” rule defines which work station an eligible operator should be moved to. Several rules were tested in the literature: Random choice; first come, first served; first in system, first served; shortest operation time; greatest work station’s workload; upstream or downstream station and several other variations (Treleven 1989; Park 1989; Bokhorst 2004; Sammarco 2014).
- The “When” rule determines whether an operator can be moved to work to another station. The two most frequently used “When” rules are the centralised rule and the decentralised rule. The former is based on the length of the queue before the workstation the operator is working at. After the completion of each job the relative lengths of different queues are evaluated to decide whether an operator is available to be transferred or not. With the decentralised rule an operator can be moved to other departments or work stations only when his queue is empty (Treleven 1989; Park 1989; Bokhorst 2004; Sammarco 2014).

- The “Who” rule indicates, in the presence of more than one operators eligible to be transferred, which of them should be assigned to a particular job on a particular work station (Bokhorst 2004).

Combinations of “When” and “Where” rules obtained contrasting results in the literature mainly due to the definition of performance criteria and the inclusion of different variables in the research according to the objective of the study. The “Where” rule was found to have positive effects on orders flow time in the shop floor, however, the choice of a “Where” labour assignment decision rule should depend on criteria such as ease of use or administrative costs (Treleven 1989). Sammarco (2014) concluded that also the managerial decision of implementing a “When” rule should be based on company’s peculiarities and strategic objectives. In his study the centralised “When” rule produced the greatest improvement in terms of flow time and work in process level if combined with a “Where” rule enabling the operator to move to the downstream station. In combination with other “Where” rule, instead, the decentralised “When” rule resulted to be the most performing. In DRC shops in the presence of delays due to workers transfer, workers learning periods and worker attrition costs, a decentralised labour assignment policy is preferred to the centralised labour assignment policy (Kher and Malhotra 1994). With high learning losses during cross-training and learning process the selection of a “When” rule is more critical respect to the decision regarding the implementation of a “Where” rule (Malhotra 1993).

While several studies were conducted on the effects of “Where” and “When” rules, the majority of them does not explain how and according to which criteria operators transferred from a work center to another are selected when more than one is available (Bokhorst 2004). Bokhorst (2004) specifically addressed this topic in his paper “On the who-rule in Dual Resource Constrained (DRC) manufacturing systems”. The study included the analysis of three different “Who” rules in combination with different levels of worker flexibility and worker efficiency at the work stations. The three rules are: Longest Idle Time (LIT), Random (where an operator is randomly selected to work on a work center from the available operators) and Priority which selects from the available operators the one with the highest priority defined by a criteria related to the shop configuration. As with the “Where” and “When” rules, the effects of the “Who” rule are dependent on the specific design of the job shop model. However, the implementation of a “Who” rule proved to be more effective in shops with

operators able to work on more work stations but with different task efficiency rather than in the presence of homogenous efficiency for all operators on all work stations. Indeed, thanks to the Priority “Who” rule the operator with the highest efficiency is more often selected to work, resulting in an improved flow time performance.

WORKER EFFICIENCY

Dealing with a flexible workforce, meaning that operators are trained to work in more than one department or at more than one work center, it is very likely that operators might not have the same level of productivity. The concept of flexibility discussed until now does not consider this possibility and workers are assumed either to have the maximum level of efficiency or to be unable to work at a work station. This would be an ideal situation where operators are perfectly interchangeable without productivity losses due to different set of skills or experience. Bobrowski (1993) defined labour flexibility as the “worker efficiency at each work station”. Figure 3 shows a worker flexibility configuration used in the study.

Figure 3 Bobrowski, flexibility matrix with heterogeneous efficiency

		WORK CENTER								
		1	2	3	4	5	6	7	8	9
WORKER	1	1.00	.95	.95	.95	.95	.85	.85	.85	.85
	2	.85	1.00	.95	.95	.95	.95	.85	.85	.85
	3	.85	.85	1.00	.95	.95	.95	.95	.85	.85
	4	.85	.85	.85	1.00	.95	.95	.95	.95	.85
	5	.85	.85	.85	.85	1.00	.95	.95	.95	.95
	6	.95	.85	.85	.85	.85	1.00	.95	.95	.95
	7	.95	.95	.85	.85	.85	.85	1.00	.95	.95
	8	.95	.95	.95	.85	.85	.85	.85	1.00	.95
	9	.95	.95	.95	.95	.85	.85	.85	.85	1.00

In the matrix of Figure 3 the efficiencies of nine workers on nine work centers are shown. All values of efficiency e_{ij} are maximum on matrix’s diagonal, when $i=j$, meaning that each worker has the maximum efficiency 1.00 on one work center. On the other work centers his efficiency is lower, representing the different skills of operators. The introduction of heterogeneous workers’ efficiencies needs to be considered in the labour assignment rules, as it was confirmed by Bokhorst (2004), too, when introducing a specific “Who” rule. Bobrowski (1993) included the information of different operators’ proficiencies at different work stations in some “When” and “Where” rules that were tested against the more

traditional rules used in the literature. The outcome of the study stated that the DRC job shop system modelled performs better when labour assignment rules includes criteria that contemplate operators' efficiency. Indeed, the best performance are obtained when an operator is moved immediately if he can move to the work center where he is more efficient ("When" rule) and, when eligible to be transferred, he is moved to the work center where he is most efficient ("Where" rule). Moreover, in this specific study, when the most performing "Where" rule is implemented, the differences obtained by choosing a specific "When" rule or another are not so significant, leaving the possibility to select the most convenient and simple rule from a managerial perspective. These results contrast with the findings of most researches that attributed to the "When" rule a greater influence on shop's results respect to the "Where" rule that was secondary.

The issue of different level of operators' efficiency was addressed by Brusco and Johns (1998) with the development of an integer linear programming model aimed at the minimization of workforce staffing cost with the constrained of minimum labour requirements. Operators' flexibility is considered both in terms of number of tasks that operators are able to perform and in terms of their efficiencies in carrying them out. Results show that the largest part of cost saving is obtained even through a limited amount of cross-training. That is, a system may realize a large portion of available cost savings by training employees or workers in more activities even if the nature of the work categories precludes cross-training at 100% productivity.

WORKER FLEXIBILITY IN HETEROGENEOUS ENVIRONMENT

Job shop systems performances are affected by two main sources of variability: order arrival variance and order processing variance (Kher 2004). In a job shop system, the frequency and the type of arriving orders may present great discontinuity. The production usually includes several types of jobs requiring different routings on multiple machines with unequal average processing time (Bokhorst 2009). Moreover, variability in processing time has a dramatic impact on shop floor queues, orders lead time and customer service (Kher 2004). What is the impact of introducing a flexible workforce in such heterogeneous environment? And more, is incremental flexibility still beneficial when implemented in systems characterised by high processing time variability?

Bokhorst (2009) proved that in Dual Resource Constrained (DRC) systems workforce, cross-training yields performance gains if the differences in orders' mean processing time are limited. In the study, two types of jobs with different processing time were considered and the benefits of labour flexibility are significant until the ratio between mean processing time is between 5:1 and 10:1. In the presence of greater differences, cross-training could be introduced in smaller subset of homogeneous machines, if the dimension of the plant justifies such a solution.

Production systems' causes of variability can be addressed by proactively reducing job arrivals and processing time variance or introducing variance control strategies like dispatching rules and worker flexibility (Kher 2004). Reducing processing times variance produces the greatest improvement. However, when controlling system variance is not possible, incremental worker flexibility positively impacts inventory performance. With low processing time variability, the benefits of incremental flexibility are reduced. As well as flexible workforce, dispatching rules are more effective in highly variable environments while their impact is lower in controlled system (Kher 2004).

4. OBJECTIVES, RESEARCH METHODOLOGY AND RESEARCH FRAMEWORK

4.1. RESEARCH GAP

Literature is mainly focused on ORR research as input control is extensively discussed. Only recently have some authors shifted their attention on output control. Most output control research deals with operator reallocation, but very few articles explore the effects of other types of output control such as subcontracting, overtime or capacity increase. Additionally, the combination of input and output control is even less examined.

For what regards operator reallocation or operator flexibility, research lacks a focus on heterogeneous efficiency, which means that different operators have different efficiencies in performing different jobs. Also, different papers have contradictory results on the benefits of increasing flexibility. This results in an unclear definition of the extent to which it is advantageous to train operators.

4.2. RESEARCH QUESTIONS

Research question 1: *What is the impact of incremental flexibility on average orders gross throughput time (GTT), shop floor time (SFT), tardiness and number of tardy orders, when combined with an ORR method for input control?*

Research question 2: *How do different levels of operators cross-training and efficiency affect shop performance? What is the impact of considering a heterogeneous flexibility pattern for operator cross training?*

Research question 3: *Which “When” rule, decentralised rule transferring operators when idle or centralised rule allowing the transfer on the basis of the queue length, is most performing? And how do shop performances react to different load threshold when a decentralised “When” rule is selected?*

4.3. SIMULATION MODEL

In order to address the issues presented in the research questions and fill the literature gap in this particular field, a job shop simulation model was developed. The model is written in Python language, exploiting mainly Python’s simulation module Simpy.

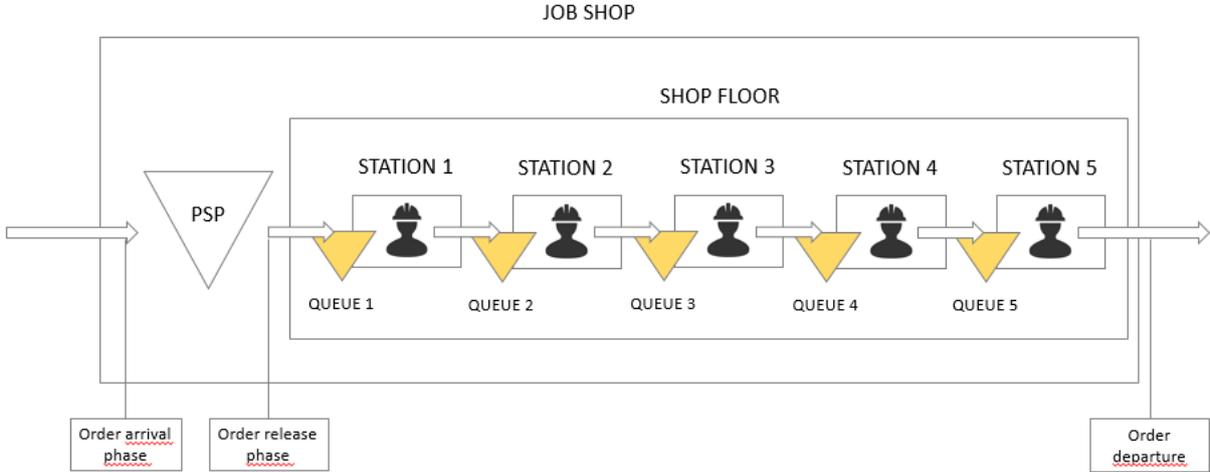
The model simulates a job shop system working on an 8 hours shift (480 minutes), 250 days per year. The two workload control systems, input control and output control, are combined in the model.

- An Order Review and Release (ORR) method controls the order arrival phase, the pre shop pool management phase and the release of orders to the Shop Floor. The introduction of this mechanism aims at the reduction of occurrence of workload peaks at workstations and reduce Work In Process (WIP) inventory (Bergamaschi 1997).
- In order to respond to workload imbalances and to enhance overall system performance an output control system is implemented in addition to the ORR method. It consists of a reallocation of cross-trained workers to other stations to help other operators performing their jobs.

The input and the output control mechanisms will be further detailed in following sections.

The system’s structure is composed by a Pre Shop Pool (PSP), where orders are stored after arrival before being processed, and a Shop Floor where orders are processed and then exit the system.

Figure 4 Simulation model, schematic representation



Pre shop pool

Orders arrival phase is managed through an Order Review and Release (ORR) method and the introduction of a Pre Shop Pool (PSP). The PSP decouples orders arrival in the system to orders release to the shop floor. This managerial tool allows to control and mitigate the impact of external demand variability on the system. It enables to monitor the workload released into the shop floor and the WIP inventory level, consequently. A further benefit of storing orders in the pre shop pool, instead of releasing them to the shop immediately after arrival, is the possibility to intervene with late modification on orders in the PSP without incurring in extra costs like reworking of semi-finished products or scraps.

Shop floor

The shop floor is composed by 5 identical work stations (Figure 4). In front of every station there is a buffer that decouples the production flow from the upstream and the downstream stage. Five workers are assigned to the shop floor, one to each station. According to their level of cross-training and the load queued at their work station, workers have the possibility to move from their station to another to help in the processing of an order and then come back to the initial position. In the “Worker reallocation” section the rule regulating operators’ transfers will be addressed and further explained.

4.3.1. Order review and release method

In this research an Order Review and Release method is implemented as an input control mechanism in combination with an output control system. An aggregate limiting ORR system was selected for the purpose of the study. The aggregate load of work in the shop floor considers the work present at a work station and the load of work currently released to the upstream stages that will have to be processed by that station. It is computed in minutes of work.

The characteristics of the of the ORR method implemented are summarized in Table 2.

Table 2 Aggregate limiting description

Aggregate Limiting (Upper-bound Limited Workload)	
Order release mechanism	Load limited
Timing convention	Discrete
Aggregation of workload	Load at each station
Workload accounting over time	Atemporal - aggregated load
Workload control	Upper-bound only
Capacity planning	Active
Schedule visibility	Limited
Release period	At the beginning of each working day

As it was anticipated before, according to Bergamaschi (1997), the Order Review and Release (ORR) manages the transition of production orders from the planning system, or directly from the customer's request, to the execution phase. It covers three consecutive phases:

- Order entry phase
- Pre shop pool management phase
- Order release phase

The following sections describe how these phases are modelled and the most significant variables and parameters affecting the performance and the outcome of the study will be presented.

Order entry phase

The orders arrival is generated at the beginning of each simulation day. It follows a Poisson distribution (that proved to be a good approximation of the arrival process (Albin, 1982)). The routing in the shop floor is supposed to be the same for all jobs. After orders arrival, a due date is assigned to each order. In the model every job is assigned a due date which is 7 days

after the arrival irrespective to the time actually required to process it. Processing time at each station is assigned through a lognormal distribution.

Pre shop pool management phase

In the pre shop pool, orders are queued after the entry phase until the release to the shop floor. In this research, two dispatching rules regulate the sequencing of jobs in the pre shop pool. These rules determine the order through which jobs are considered for the release.

The two dispatching rules are:

- First Come First Served (FCFS)

It is the simplest rule. Jobs are considered for the release according to their arrival order.

- Shortest Processing Time (SPT)

Jobs are ordered in the pre shop pool from the shortest processing time to the one with the highest processing time.

Order release phase

In the order release phase orders are evaluated to decide whether they can be introduced into the shop floor or they have to wait in the pre shop pool according to the workload currently in the shop and to the norm set as the upper workload limit. This activity is scheduled at the beginning of each simulation day. Every order in the PSP at the decision moment is considered for the release. The order follows the sequencing defined by the dispatching rule presented in the previous section.

An order is released to the shop floor if its processing time at each work station added to the load of that station already in the shop does not exceed the workload norm. The workload norm for each station is calculated as follow:

$$\text{station } k \text{ norm} = \frac{k * (\text{Workload norm} - 480)}{\text{number of stations}} - 480$$

Where k goes from 1 to 5 and represents the number of the station, the denominator is equal to 5 (which is the number of stations included in the model) and the workload norm is a

variable set at the beginning of the simulation. In this research, nine different levels of workload norm are tested.

The workload already released at a station is computed through the atemporal approach (Bergamaschi 1997), which means that the work distribution over time does not affect the release decision. So, the current load of a work station is obtained by summing up all the processing time of the jobs in the shop that have to pass through that station. The model sums the load in the queue of the workstation to the processing time on that station of the jobs present in upstream stages.

Therefore, the release is subject to the following criteria: an order i is released to the shop floor if

$$\text{Current load of station } k + \text{processing time}_{ik} \leq \frac{k * (\text{Workload norm} - 480)}{\text{number of stations}} - 480$$

where processing time $_{ik}$ is the processing time of job i at the station k .

Every time an order exits the PSP and it is queued in the shop floor, the current shop load is updated with its processing time. Orders in the backlog file of the PSP are evaluated in sequence until a station's load reaches the workload norm, in which case the order release phase is interrupted and the remaining jobs will wait in the queue until the following day's release phase.

A particular case is represented by jobs whose due date is passed. In this eventuality the orders are released to the shop floor even if the release criterion is not respected and the workload norm at same station will be exceeded.

4.3.2. Worker reallocation

There are two sets of rules that are responsible for the operator reallocation process:

1. Centralised rules, when the decision is based on the comparative length of one queue respect to the other queues.
2. Decentralised rules, when an operator is reallocated only when its own queue is idle.

After the completion of each order the operator reallocation decision is performed. For what regards the centralised rules,

- A parameter a is defined, such that $0 < a < 1$
- The direct workload of each queue is measured
- The maximum direct workload (max WL) of a queue is considered as benchmark
- A threshold $a * \max WL$ is set. If the current workload of that station is lower than the threshold, the operator will be reallocated to the station with the highest workload. As a result, the processing time of the order that the reallocated operator is contributing to perform is decreased by a percentage which depends on the efficiency of that operator.
- When the processing of that order is finished, the operator will go back to his own station

For what regards decentralised rules, the reallocation evaluation process is also performed right before a new order is processed

- The current station i to which a possible reallocation is being considered is beginning a new order
- Depending on the specific flexibility rule, a different number of operators is evaluated for reallocation. There are two types of Rules:
 - Homogeneous efficiency Rules where each operator has the same efficiency no matter which station is he supporting. In this case, in Rule1 each operator can only support downstream station, in Rule2 each operator can support both upstream and downstream stations, in Rule 3 each operator can support all the other operators
 - Heterogeneous efficiency Rule where each operator has different productivity depending on which operator goes to which station
- If the operator $j \neq i$ considered is currently idle and at the same time is on his own station, he will be reallocated to the current station i ,
- In case of more than one idle operator the simulation model selects the most downstream

- As a result of the reallocation process, the processing time of the current order in the current station i is decreased by a percentage which depends on the efficiency of that operator j in that specific station
- When the processing of that order is finished, the operator j will go back to his own station j .

4.4. PERFORMANCE MEASUREMENT

The system performance evaluation is achieved thanks to the measurement of different variables:

1. Gross Throughput Time (GTT) [hours]: the time period between the order entry in the pre shop pool and the exit from the last processing station
2. Shop Floor Time (SFT) [hours]: the time period between the order between the order release from the pre shop pool and the exit from the last processing station
3. Percentage Tardy orders (Tardy) [$\frac{piece}{day}$]: the average number of orders delivered after the due date
4. Tardiness [hours]: the average delay of the orders that do not respect due date.

These performances are among the most used in literature in assessing the performance of a system. They are particularly relevant as GTT and SFT are actually important drivers for shop efficiency and reactivity to customer request. The measure of WIP improvement allows to evaluate the benefit in terms of stock holding cost. Tardy and Tardiness assess the company capability to predict and respect order due date which, according to Thurer et al 2014, is becoming an increasingly important factor for order winning for MTO companies.

4.5. CONFIGURATION OF EXPERIMENT

The configuration of the results generation process was shaped in order to give an answer to the three research questions presented before in the section of this study dedicated to the introduction of literature's gap and research's objective.

Research question 1

The first research question aims at the study of the impact that different level of flexibility, considered as the number of different stations an operator can work on, has on shop performance. To respond to this specific issue four simulation “rules” were developed to test four corresponding flexibility levels: R0, R1, R2, R3.

- R0 corresponds to the benchmark which the other rules are tested against. It is the standard shop configuration where operators are trained to perform the job on one single station and cannot be transferred among the shop floor. With this rule no cross-training is provided to the workforce and it represents the lowest level on the flexibility scale which is flexibility = 1.
- R1 is the rule that corresponds with the introduction of the minimum level of flexibility (flexibility = 2). In this case workers are cross-trained to work on two stations. Each worker can work at his assigned position or be moved to the downstream station according to the specific when rule implemented.
- R2 simulates a level of flexibility equal to 3 in which operators can be transferred to the upstream or the downstream station.
- R3 is the maximum level of flexibility (flexibility = 5) where each worker can work at each station.

For the objective of this study only homogeneous flexibility is considered meaning that every worker can work on the same number of station. The four matrices in Figure 5 show the four rules. $e_{ij} = 1$ means that worker i can work on station j , while $e_{ij} = 0$ means that he cannot.

Figure 5 Flexibility matrices

		R0					R1					
		WORK STATION					WORK STATION					
WORKER		1	2	3	4	5		1	2	3	4	5
	1	1	0	0	0	0	1	1	1	0	0	0
	2	0	1	0	0	0	2	0	1	1	0	0
	3	0	0	1	0	0	3	0	0	1	1	0
	4	0	0	0	1	0	4	0	0	0	1	1
	5	0	0	0	0	1	5	1	0	0	0	1

		R2					R3						
		WORK STATION					WORK STATION						
		1	2	3	4	5			1	2	3	4	5
WORKER	1	1	1	0	0	1	1	1	1	1	1	1	1
	2	1	1	1	0	0	1	1	1	1	1	1	1
	3	0	1	1	1	0	1	1	1	1	1	1	1
	4	0	0	1	1	1	1	1	1	1	1	1	1
	5	1	0	0	1	1	1	1	1	1	1	1	1

Research question 2

The second issue addressed by this study is the impact of different levels of workers' productivity when moved to another station to help the co-worker assigned to that position in performing the job. A worker transferred to help another station results, in the simulation model developed and used in the study, as a reduction of the processing time of the order. In the configurations presented above for the first research question, operators either can work on other workstations with 100% efficiency or cannot work there. When an operator is transferred and works with 100% efficiency with the worker already on that station (which has the maximum 100% efficiency, too), the time to complete the job on that order at that workstation is halved. Therefore, in the model, that processing time is multiplied by the coefficient 0,5. The introduction of different efficiency levels imply different reduction of orders' processing time when workers are moved.

Two configuration are considered in the study: homogeneous workers' efficiency and heterogeneous workers' efficiency.

Figure 6 Homogeneous worker efficiency

		WORK STATION				
		1	2	3	4	5
WORKER	1	100%	a	0	0	0
	2	a	100%	a	0	0
	3	0	a	100%	a	0
	4	0	0	a	100%	a
	5	0	0	0	a	100%

With homogeneous worker's efficiency, the same level of cross-training is delivered to operators. This means that each worker has 100% efficiency in his position and a productivity at other stages he can work on that is set equal independently on which worker is moved where (Figure 6). Five level of workers' efficiency are tested. From the maximum efficiency level (100%), which results in processing time halving, to the minimum level that produces a 10% processing time reduction (it is multiplied by 0,9). In Table 3, the five coefficients representing orders processing time reduction are listed with the corresponding operators' efficiency.

Table 3 Operator efficiency and processing time reduction relationship

Processing time reduction coefficients	Operators' efficiency
0,5	100%
0,6	67%
0,7	42%
0,8	25%
0,9	11%

In this research, two configurations of heterogeneous efficiency in addition to the homogeneous efficiency just described are included. The aim of the introduction of this further set of experiments is to have a more realistic model of skills' and competencies' distribution among workers in a job shop and compare the results with the homogeneous efficiency configuration. In a productive plant it is very likely that operators do not have the same level of productivity in performing different jobs. Bobrowski (1993) introduced different levels of efficiency in the research but did not compare the results with the standard homogeneous efficiency used in the literature. In this study it is assumed that operators have maximum efficiency (100%) at their workstations and decreasing efficiency as the station they move to is further away from their assigned position. The matrices in Figure 7 show the coefficients for processing time reduction due to different worker efficiency at different workstations. R3M01 implies an average reduction of 30% of the processing time (0,7) while the average reduction with R3M005 is 40% of the processing time.

Figure 7 Heterogeneous efficiency matrices

		R3M01					R3M005						
		WORK STATION					WORK STATION						
		1	2	3	4	5			1	2	3	4	5
WORKER	1		0,6	0,7	0,8	0,9			0,55	0,60	0,65	0,70	
	2	0,6		0,6	0,7	0,8	0,55		0,55	0,60	0,65		
	3	0,7	0,6		0,6	0,7	0,60	0,55		0,55	0,60		
	4	0,8	0,7	0,6		0,6	0,65	0,60	0,55		0,55		
	5	0,9	0,8	0,7	0,6		0,70	0,65	0,60	0,55			

Research question 3

The last research question investigates the effects of a centralised when rule on the performance of the job shop system modelled compared with a decentralised when rule. For this purpose, Rule 3, which is the rule with maximum flexibility and uses a decentralised “When” rule to define operators’ eligibility for transfer, is used as a benchmark to evaluate the behaviour of the system when a centralised “When” rule is adopted. With Rule 3 operators can move to every other station only if the queue of orders corresponding to their own station is empty. To test the centralised “When” rule a workload threshold is introduced in the model to determine whether a worker can or cannot move. This threshold is calculated multiplying the load of work in the longest queue by a coefficient a :

$$\textit{Workload threshold} = a * \textit{max workload}$$

Every time an operator has finished his job on an order and has to begin processing the following one, this threshold is updated and, if the shortest queue in front of a station is lower or equal to the workload threshold, than the operator with the shortest queue is transferred to help in performing that job and the order processing time is reduced.

$$\textit{min workload} \leq \textit{Workload threshold}$$

In order to study this rule and to investigate the impact of selecting a specific threshold, simulations has been run with three levels of parameter a : 0,2; (1/3); 0,4. The corresponding rules are RS02, RS033 and RS04, respectively.

The rules and the specific system’s variables investigated in this study have been tested in different conditions to support the validity of conclusions. With this aim the simulations were run with:

- Two dispatching rules: Shortest Processing Time (SPT) and Planned Release Date (PRD).
- Nine levels of workload norm to test the behaviour of the shop according to different amount of workload released from the pre shop pool to the shop.
- Three levels of system’s saturation. The saturation of the system is controlled by imposing three levels of mean for the Poisson distribution generating daily orders’ arrival rate.
- Five levels of processing time variance that are necessary to verify if the results are confirmed in environments characterised by higher degree of processing time variability due to higher diversity in the jobs type performed in the shop floor.

All the variables and the parameters tested are summarised in Table 4 and Table 5.

Table 4 Description of tested rules

RULE	FLEXIBILITY LEVEL	CROSSTRAINED WORKERS EFFICIENCY CONFIGURATION	WHEN RULE
R0	1	-	-
R1	2	Homogeneous	Decentralised
R2	3	Homogeneous	Decentralised
R3	5	Homogeneous	Decentralised
R3M01	5	Heterogeneous	Decentralised
R3M005	5	Heterogeneous	Decentralised
RS02	5	Homogeneous	Centralised
RS033	5	Homogeneous	Centralised
RS04	5	Homogeneous	Centralised

Table 5 Experimental work, list of parameters

SYSTEM PARAMETER	VALUES
Saturation [orders/day]	14,8
	15
	15,3
Dispatching Rule	SPT
	PRD
Workload norm [min]	1440
	1620
	1800
	2040
	2340
	2700
	3300
	4800
Flexibility level	1
	2
	3
	4
	5
Workers' efficiency (coefficient for processing time reduction)	0,5
	0,6
	0,7
	0,8
	0,9
Processing time variance [min]	576
	1024
	1236
	1600
	1936
When rule	Centralised
	Decentralised
Workload threshold for centralised when rule	0,2
	1/3
	0,4

DETERMINATION OF LENGTH, WARM-UP PERIOD AND NUMBER OF RUNS²

The aim of a simulation is to predict the performance of a system at steady state. Therefore, the number of runs and their length and the length of the warm up period should be estimated and used in the simulation in order to obtain meaningful and representative data.

In this study, the length of the simulation time is 500 days. Each simulation is replicated in 50 runs (Land, 2006). This with the objective of reducing the experimental error and the variance of the parameters. Both parameters (simulation time length and number of runs) are calculated following the procedure presented in the research by Mosca, et al. (1982).

At the beginning of the simulation, the production system is assumed to be empty. This implies that before reaching the steady state some simulation time is necessary. Data in that period are inaccurate and not representative of the potential performance of the system, therefore, they cannot be considered. For this reason, the length of the initial transient period from the initial condition to the steady state, that will be called “warm up” period, has to be estimated in order to exclude not relevant data that may alter simulation’s results.

According to the literature, different methods to determine warm up period’s length exist and were used. In this thesis, the graphical method described by Welch (1983) is implemented. An experimental run is made for each level of labour flexibility. The total workload processed daily is chosen as the main performance parameter, as it represents the stability of the system. The graph in Figure 8 represents the workload distribution over 500 simulation days in the system with flexibility level 2 (R1). Data reach the steady state after the 50th day, while in the initial period the system does not have consistent reliable performance, yet. With higher flexibility (Figure 9) the system requires more time to settle down, and the warm up period is extended until the 250th simulation day.

All statistics are collected after the warm up period. And as it was mentioned before the length of one experiment is 500 days. These values will be used for all simulation runs performed in this study.

² The warm up period definition was written by L.C. Acevey (2017)

Figure 8 Warm up period with labour flexibility = 2 (R1)

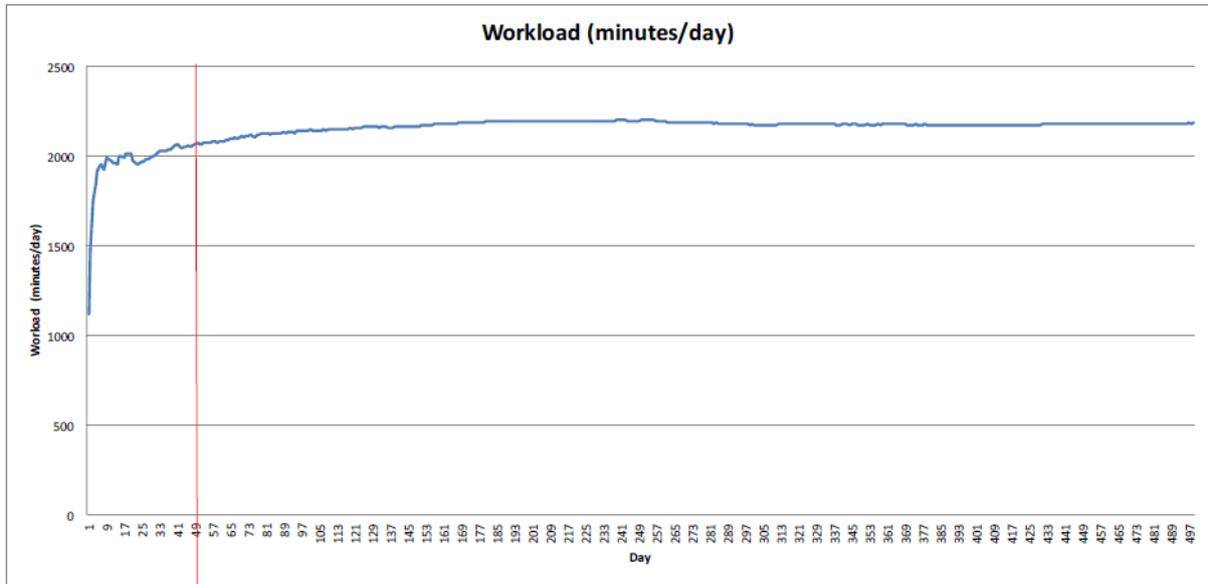
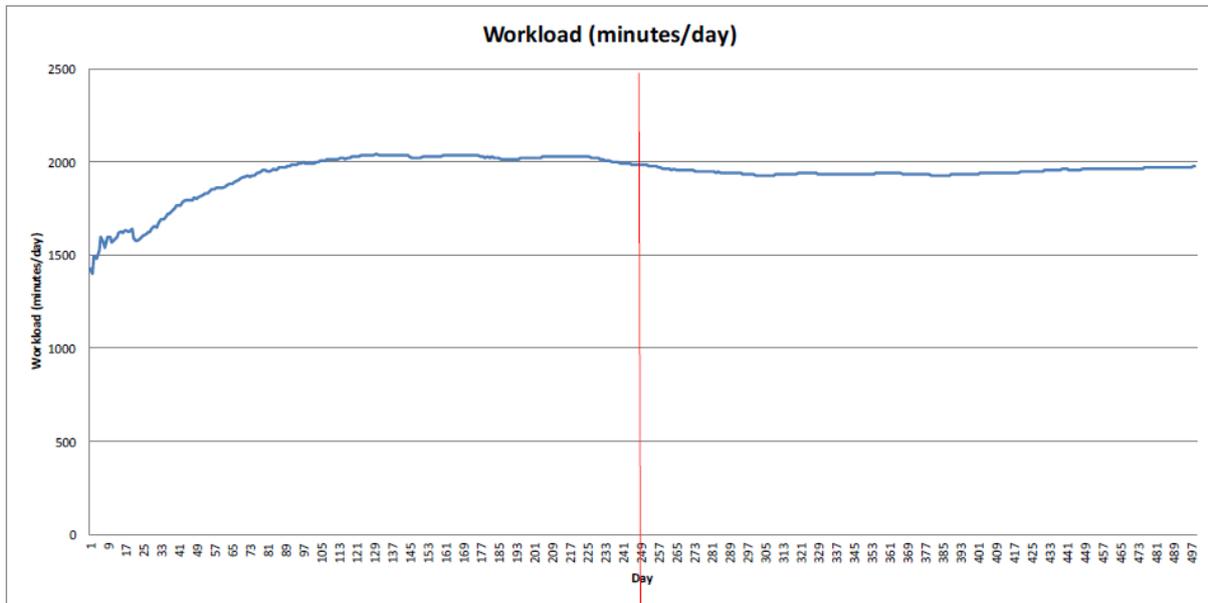


Figure 9 Warm up period with labour flexibility = 3



5. ANALYSIS OF RESULTS

5.1. RESEARCH QUESTION 1

What is the impact of incremental flexibility on average orders gross throughput time (GTT), shop floor time (SFT), tardiness and number of tardy orders, when combined with an ORR method for input control?

The first topic addressed by this study is incremental flexibility. The first objective of the study is to investigate the impact of growing level of cross-training manpower to work on more workstations, performing different types of job, combined with the aggregate limiting ORR method presented before. Four levels of flexibility (1,2,3,5) corresponding to four rules (R0, R1, R2, R3) are compared.

Figure 10 Percentage GTT variation for every workload norm due to increasing level of worker flexibility respect to R0

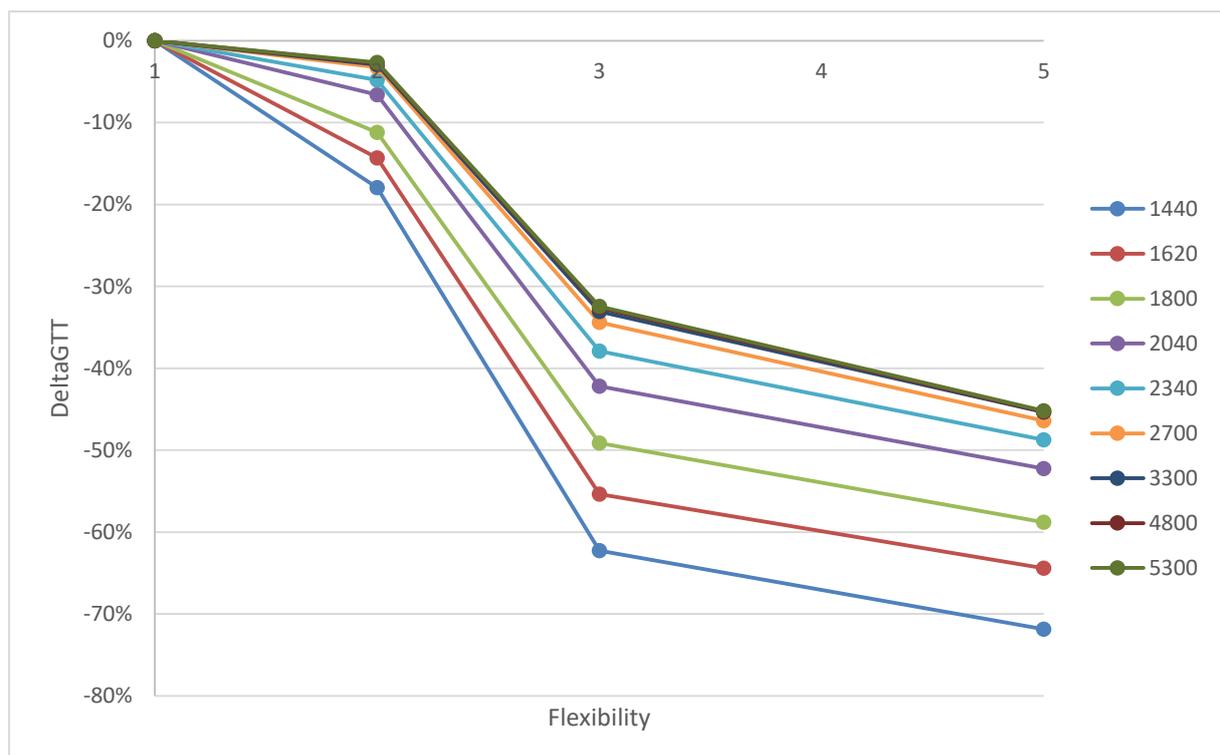


Figure 10 shows the improvement in average Gross Throughput Time (GTT) performance due to the increase in worker flexibility. GTT reduction is computed as the percentage variation respect to the standard rule R0 with flexibility = 1 that does not consider worker reallocation among workstations. On the graph it is possible to see that the lower the workload norm, the greater the reduction in average GTT. Another interesting result is that the most significant

impact on performance is obtained with a flexibility level equal to 3 where operators can move to two workstations in addition to theirs. The average variation obtained by increasing flexibility level from 2 to 3 is -35%, against a variation of -7% and an additional 11% reduction shifting from flexibility 1 to 2 and 3 to 5, respectively. These results are in compliance with the conclusions obtained in previous studies in which the major impact of cross-training operators is related to the introduction of a flexibility level equal to 2 (Park 1989; Park 1991; Felan 2001) or 3 (Givi 2015). Further efforts in cross-training to achieve the maximum flexibility when every operator can work on every workstation have a lower impact on performance, yet leading to a not negligible GTT reduction. In higher saturation environments, the benefits of incremental flexibility are confirmed and the GTT reduction curves maintain the same behaviour as it is possible to see in the graphs in Appendix A where data obtained with saturation 15 and 15,3 orders per day are plotted.

Despite the comparable trend of the curves, in contexts characterised by greater demand and consequent higher saturation, a further increase in flexibility exceeding level 3 to achieve level 5 has a slightly sharper impact on GTT. Figure 11 shows the percentage decrease of gross throughput time (DeltaGTT) and shop floor time (DeltaSFT) for each rule R1, R2 and R3 considering, for each rule, the result of the workload norm that yields the minimum GTT. The percentage variations are computed against the standard rule R0. The graph in Figure 12, instead, represents the percentage variations shifting from a flexibility level to the following one. With higher saturations the relative improvement from R2 to R3 is greater both in terms of GTT and SFT. In Figure 11 the vertical and horizontal distance between R2 and R3 on the curve with saturation 15,3 is greater than for the other two saturations 14,8 and 15. The same conclusion can be highlighted in the graph of Figure 10 where R2 and R3 on the green line have very close values, while on the other curves the differences are more evident. This implies that the performance improvement from flexibility 2 to 3 and 3 to 5 are more similar with saturation 15,3.

Figure 11 Incremental flexibility on different saturation levels. Comparison with R0

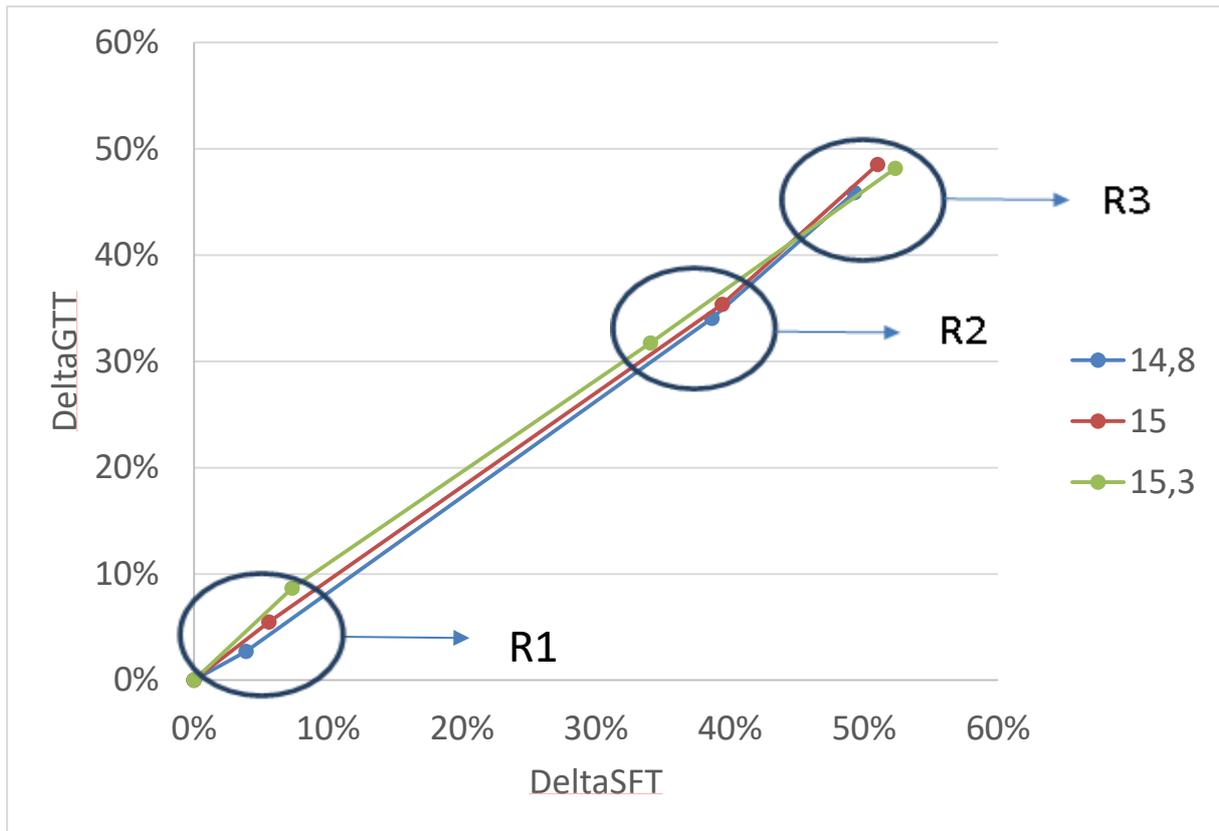
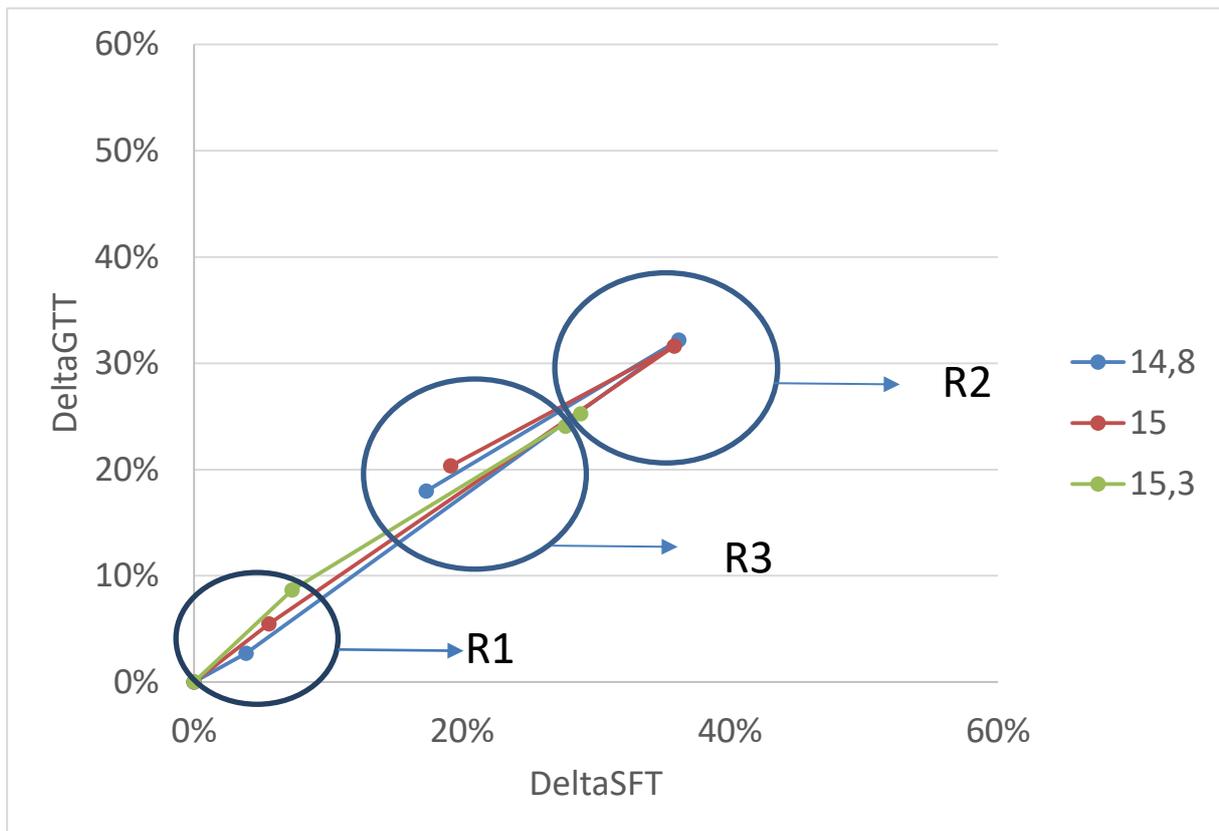


Figure 12 Incremental flexibility on different saturation levels. Comparison with previous flexibility level



Incremental flexibility is beneficial for every performance criterion considered in this study. Figure 13 shows the percentage impact of incremental flexibility on shop floor time with low and high saturation, respectively. The shop floor time reduction follows the same pattern described before for the gross throughput time: a sharper improvement from flexibility 2 to 3 and a lower marginal gain shifting from 3 to 5. As it happened for GTT, with higher saturation of the system this marginal improvement difference is reduced. Moreover, with saturation 15,3 the relative improvement is lower when a low value of workload norm is implemented in the ORR mechanism. This results in a lower average SFT reduction: -27% and -39% with flexibility 3 and 5 and saturation 14,8, against -20% and -32% with saturation 15,3.

Average tardiness of orders exceeding due dates and the average number of tardy orders per day have almost identical trends when observed in contexts with different shop floor workers' flexibility and saturation. Figure 14 and Figure 15 show how these two performance criteria are affected by incremental flexibility with saturation 14,8 and 15,3 average orders per day. The method to calculate and plot the data is the same used and presented before for GTT and SFT. In the graphs with the lowest saturation tardiness and tardy orders seem to have a surprising behaviour. The introduction of the minimum level of flexibility (2), in combination with the selection of workload norm levels higher than 2340, seems to significantly worsen the two performances and the ability of the system to respect due dates. However, with the average of 14,8 orders per day and high workload norm average tardiness and average number of tardy orders per day are negligible. Therefore, even high variations (more than 100% for tardiness and more than 20% for tardy) in these performances result in an absolute value that does not remotely compromise the efficacy and timeliness of the system. Moreover, from flexibility 3 on, the occurrence of tardy orders is completely removed. For what concerns high saturation contexts, the impact of incremental flexibility is consistent with what was presented before. Level 3 of flexibility defines a 77% reduction in average tardiness and a 71% reduction in tardy orders. The effort in cross-training until the maximum flexibility further reduces tardiness and tardy orders by 13% and 16%, respectively.

Figure 13 SFT: percentage variation respect to R0 with increasing flexibility

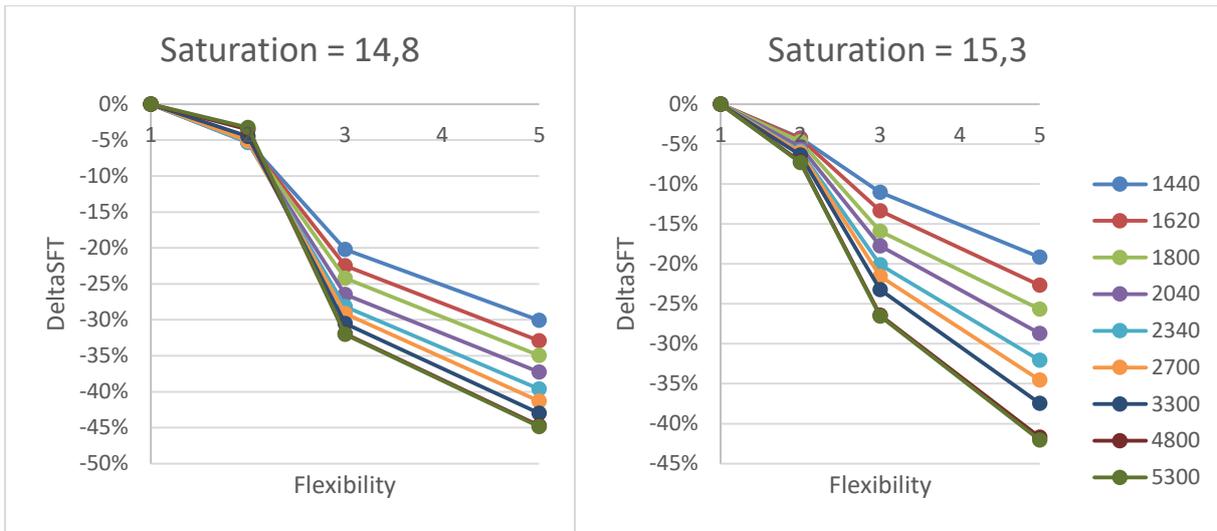


Figure 14 Tardiness: percentage variation respect R0 with increasing flexibility

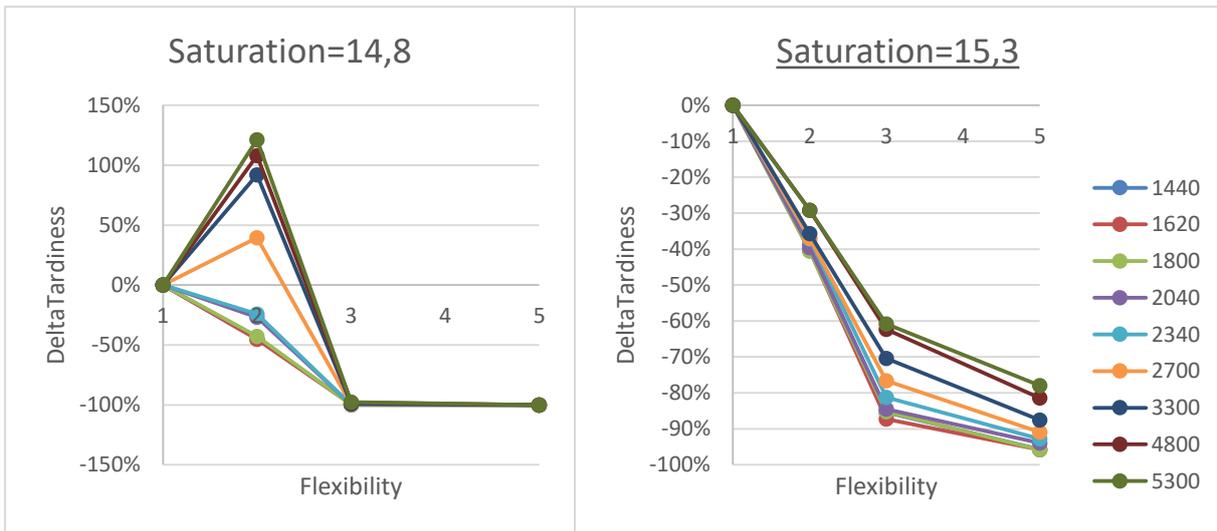
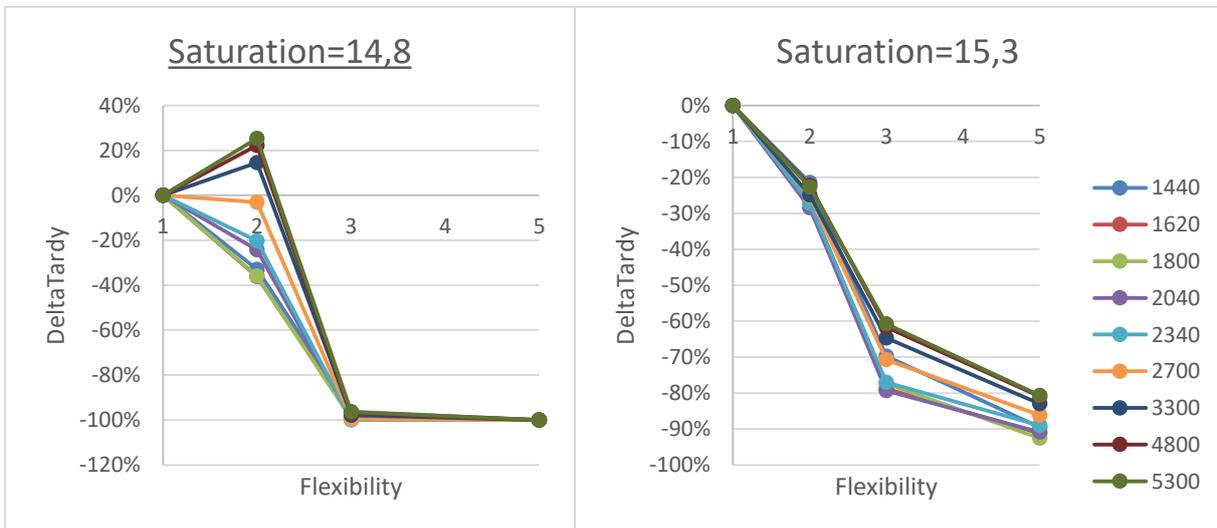


Figure 15 Tardy: percentage variation respect R0 with increasing flexibility



HIGH VARIABILITY CONTEXTS

This section is dedicated to the study of system performance when processing time variability increases. Higher processing time variance may depend on a greater heterogeneity in the job types that are in the production mix of the productive plant and that are requested by customers. The system was tested with five levels of processing time variance, as it was presented in the configuration of experiment section. Variances considered are in a range between a moderate variability level 576 and a high value of 1936.

Effect of increasing variability on system performances are presented in Figure 16, 17, 18, 19 below. Each graphs presents the behaviour of the four flexibility rules (R0, R1, R2, R3) according to the variation of the new variable introduced in this paragraph.

The worsening in overall system results consequent to greater differences in jobs processing times is rather intuitive. GTT and SFT present a positive linear trend as variability increases. However, the significant aspect emerging from these data is that the slope of the curve R0, which does not include any level of workers cross-training, is greater than the others. As flexibility increases the slope of the curves decreases. This result suggests that the introduction of worker flexibility not only improves orders flow time, but smooths the impact of processing time variability, too.

Looking at tardiness and tardy orders graph this conclusion appears even more clearly. The divergence of the curves related to higher level of variance appears more evident in Figure 16 where the average tardiness when no operators' transfer is possible grows faster than the others, while with R2 and R3 variability impact is more controlled.

Figure 16 GTT for increasing variability

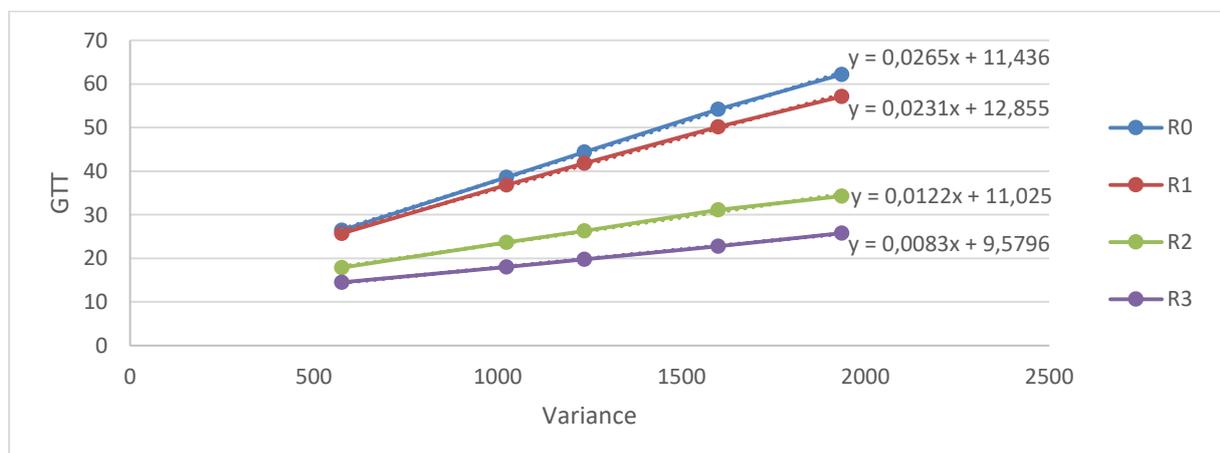


Figure 17 SFT for increasing variability

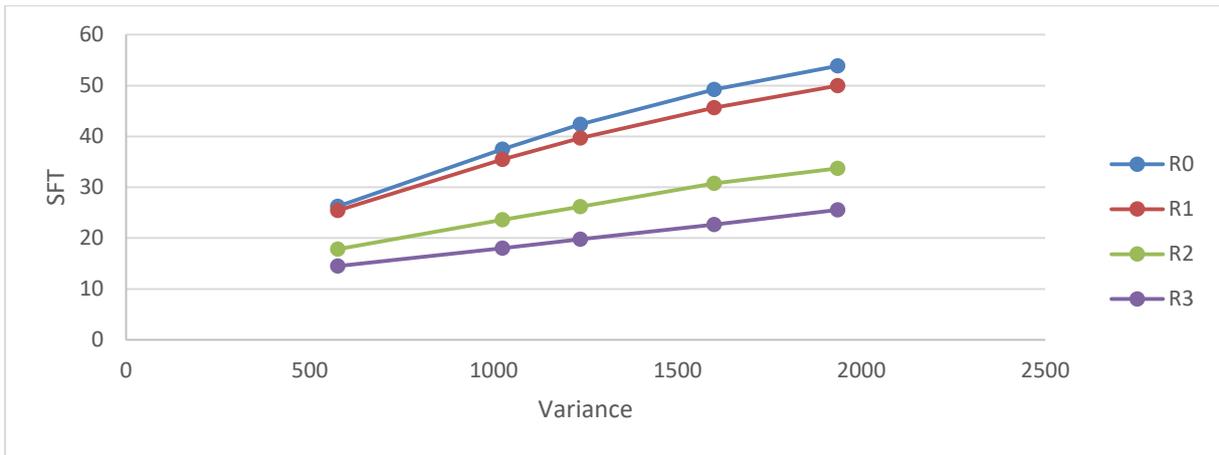


Figure 18 Tardiness for increasing variability

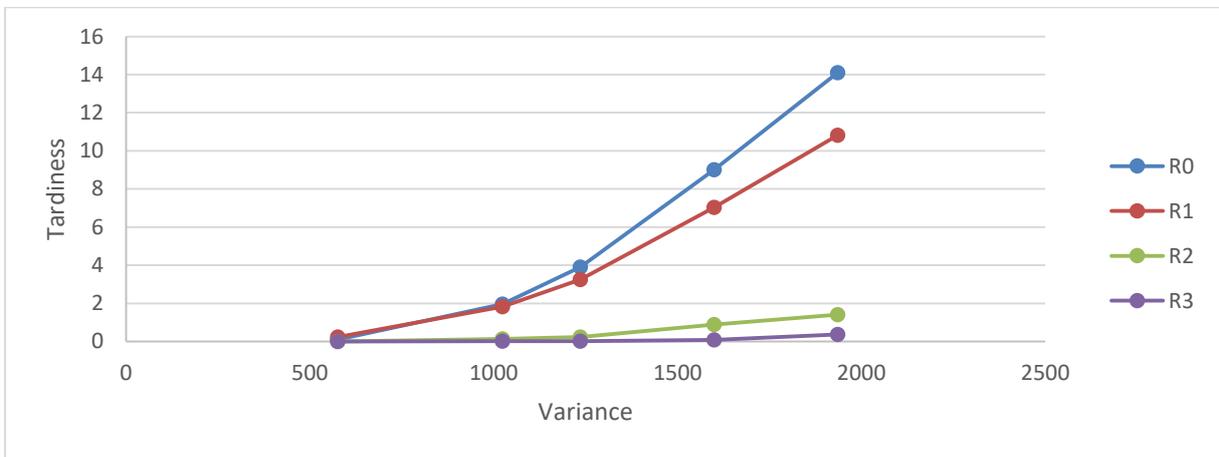
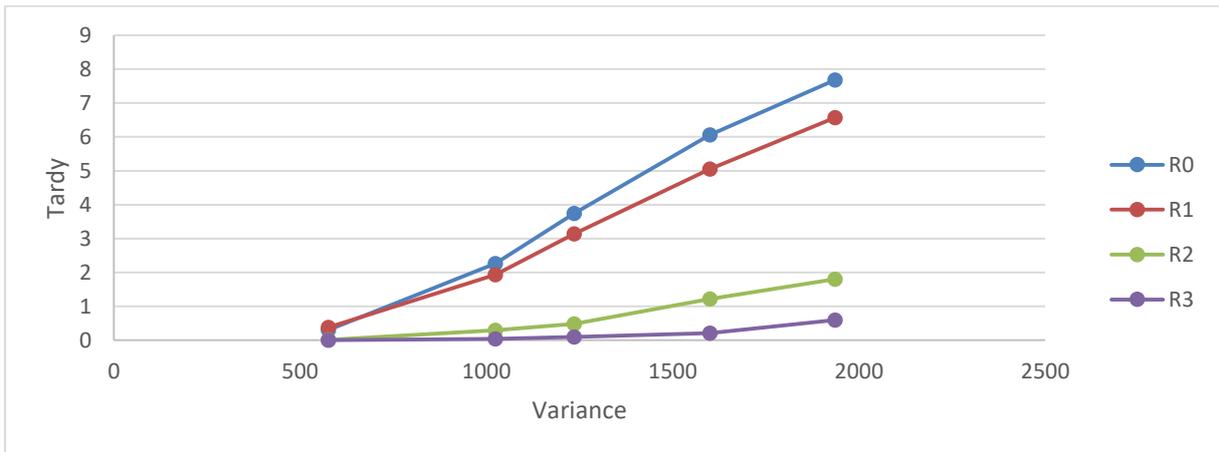


Figure 19 Tardy for increasing variability



5.2. RESEARCH QUESTION 2

How do different levels of operators cross-training and efficiency affect shop performance?
What is the impact of considering a heterogeneous flexibility pattern for operator cross training?

5.2.1. Homogeneous efficiency

Figure 20 compares the GTT performance of R1, R2, R3 with increasing operator efficiency ($GTT(R_i, eff_j)$ is plotted). All the curves are decreasing functions of the efficiency meaning that it positively impacts performances, as expected. Figure 21 show the comparison of GTT performance between decentralized rules (R3) and centralized rules (RS02, RS033, RS04). Even in decentralized rules the incremental improvement is confirmed. Increasing efficiency level, the centralized rules outperform R3.

Figure 20 GTT trend with increasing efficiency for decentralized rules

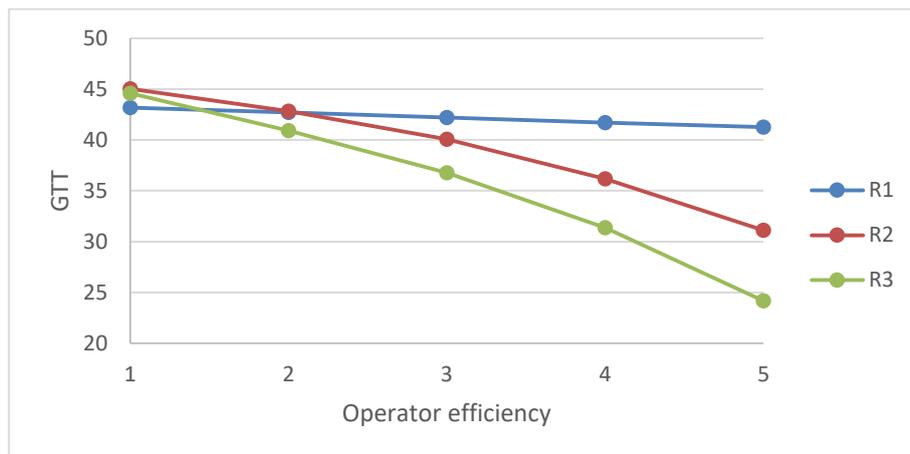


Figure 21 GTT trend with increasing efficiency for centralized rules

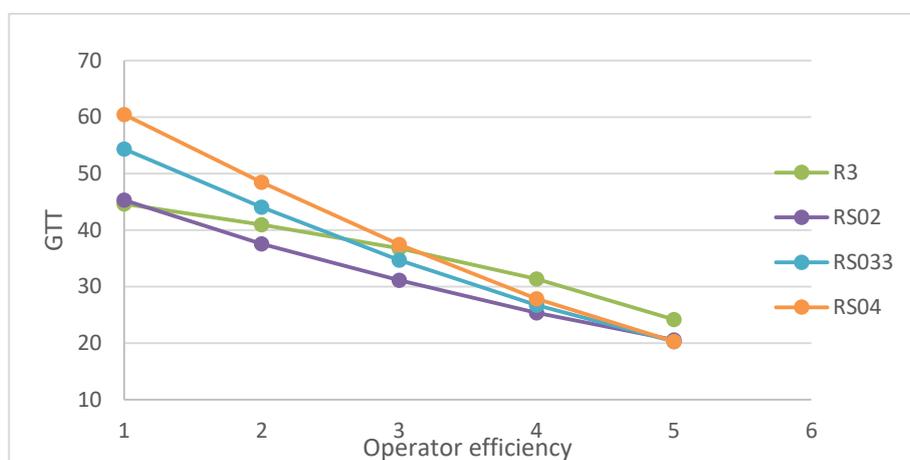
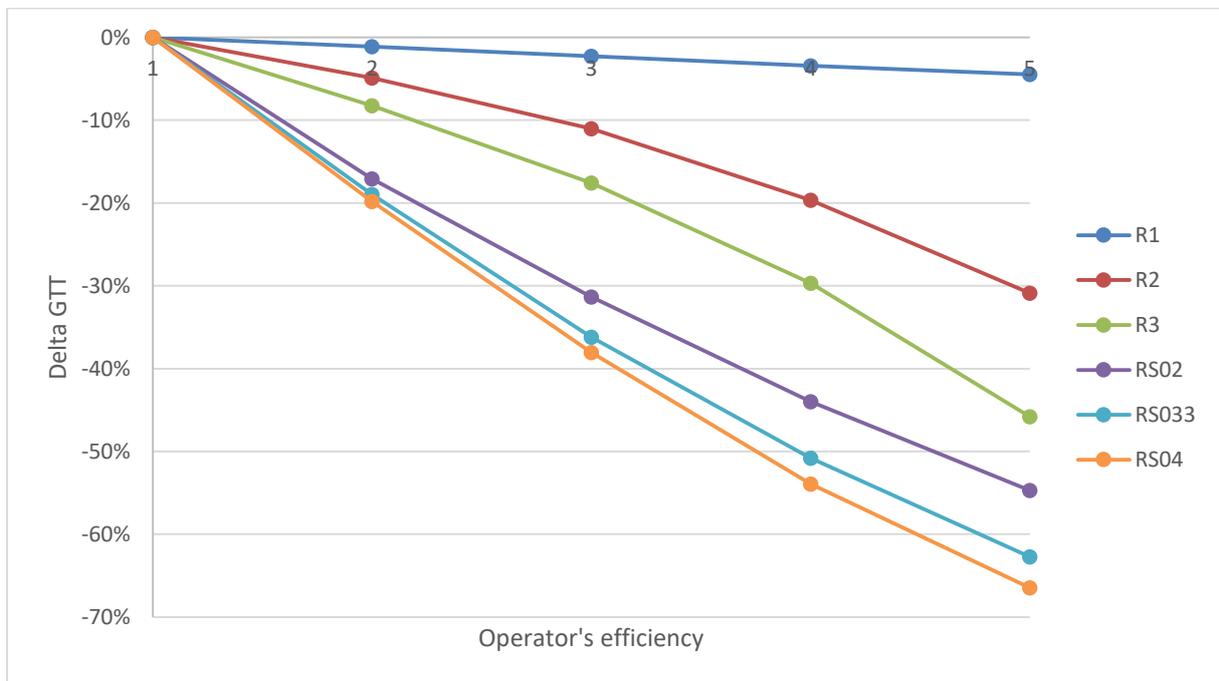


Figure 22 shows the percentage improvement of GTT respect to lowest level of efficiency of the correspondent rule. It is evident, as already mentioned, that efficiency is very beneficial and that for decentralized rules, the higher the flexibility, the larger the impact on GTT (In this case $\frac{GTT(Ri.eff_j)-GTT(Ri.eff_1)}{GTT(Ri.eff_1)}$ is plotted). However, it is interesting to assess to what extent it is advantageous to increase flexibility: for R1 after improving for the second level of efficiency, the results are steady, no matter how good the efficiency is. For R2 and R3 and all the centralized rules instead the benefit increases. There is a significant difference between the degree of improvement for decentralized rules which reaches at most 45%, while centralized rules can reach from 50% to 63% improvement with highest efficiency.

Figure 22 GTT percentage improvement with increasing efficiency

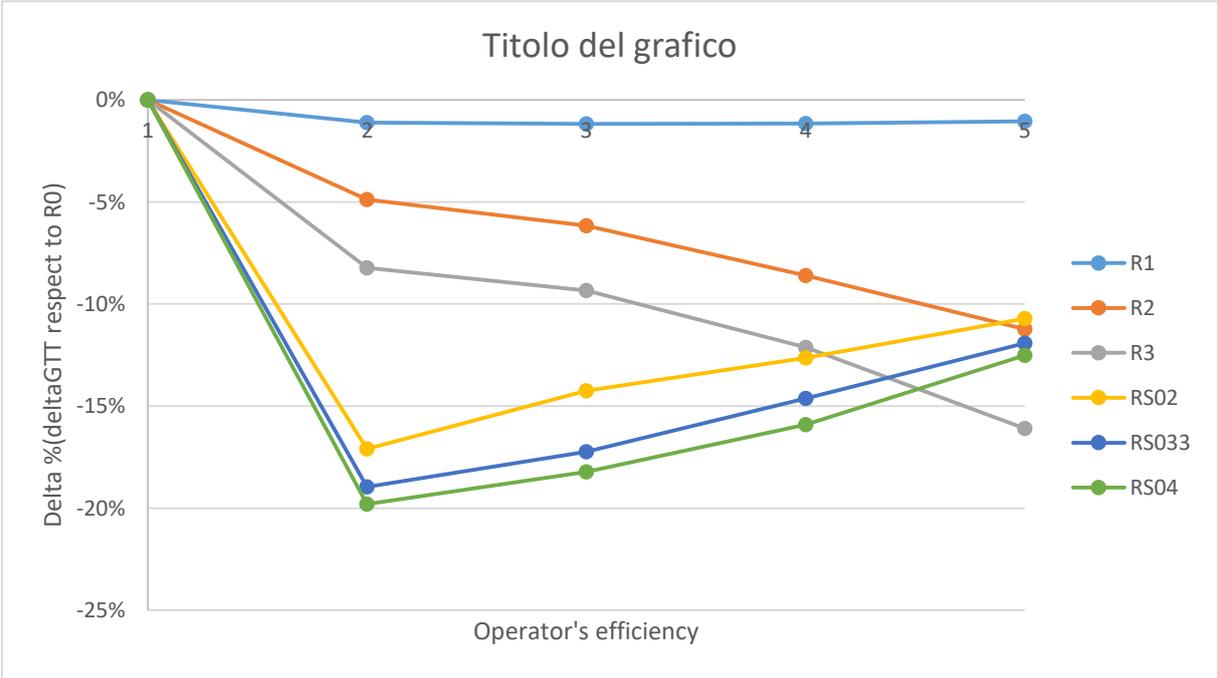


In Figure 23, the marginal improvement expressed as the difference between the percentage improvement of one efficiency respect to the previous one is plotted.

$$\left\{ \begin{array}{l} \frac{(GTT(Ri.eff_1)-GTT(Ri.eff_j))}{GTT(Ri.eff_1)} - \frac{(GTT(Ri.eff_1)-GTT(Ri.eff_{j-1}))}{GTT(Ri.eff_1)} = \frac{(GTT(Ri.eff_{j-1})-GTT(Ri.eff_j))}{GTT(Ri.eff_1)}, \\ \text{if } j= 2,3,4,5 \\ 0, \\ \text{if } j=1 \end{array} \right.$$

R1 shows a slightly increasing trend, which implies as already hinted that the marginal benefit of increasing efficiency becomes lower and lower. A decreasing trend as the one observed for R2 and R3 implies that for every marginal improvement of flexibility, the benefit is greater and greater: there are increasing marginal gains. For what regards centralised rule, the largest improvement is given with the first efficiency increase, which means that with very low effort, significant gains can be obtained. When it comes to a larger efficiency increase, the performance improvement is not as brilliant, though it remains similar to centralized rules' improvement.

Figure 23 Delta variation of GTT



The following Figures 24, 25, 26, show the other performance improvements due to increasing efficiency. SFT results are very similar to GTT. For what regards Tardy and Tardiness, the improvement is not very significant for R1. For R2 and R3 efficiency is a significant driver for performance improvement as it is for all centralized rules. Tardy and Tardiness reduction can reach up to 90% with highest efficiency.

Figure 24 SFT variation with increasing efficiency

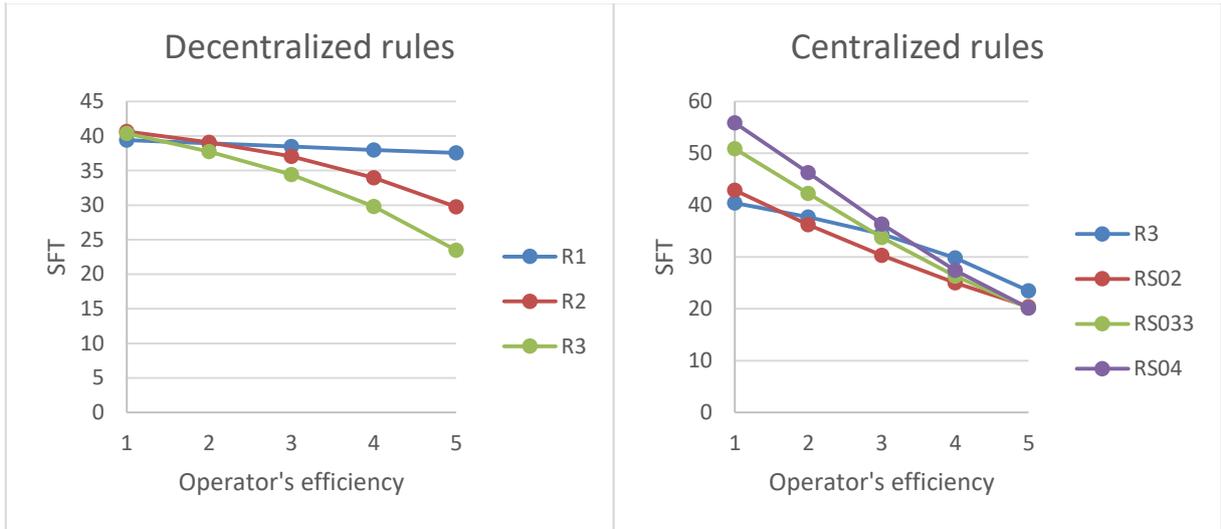


Figure 25 Tardiness variation with increasing efficiency

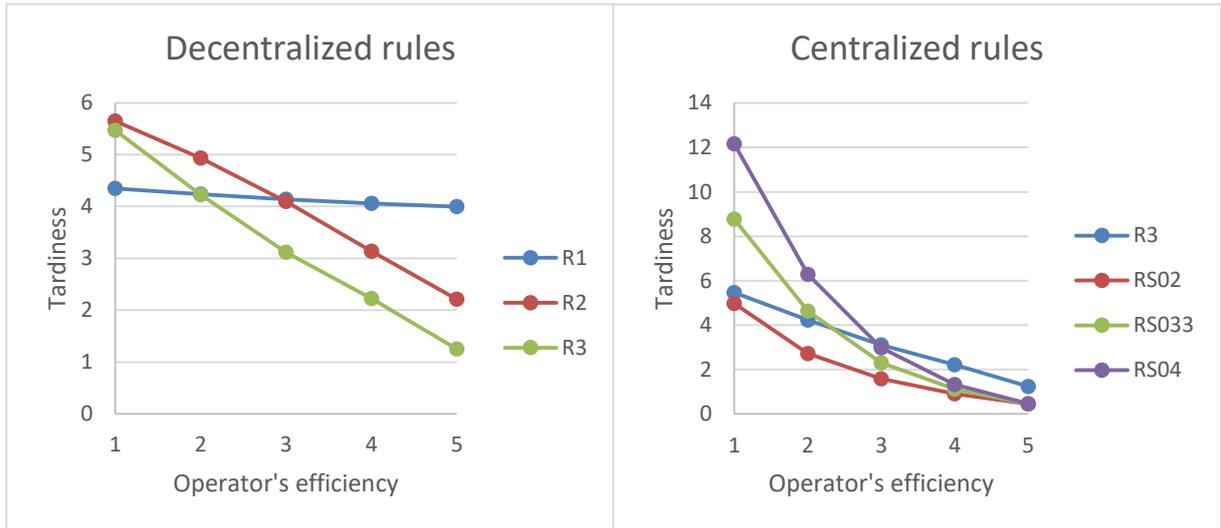
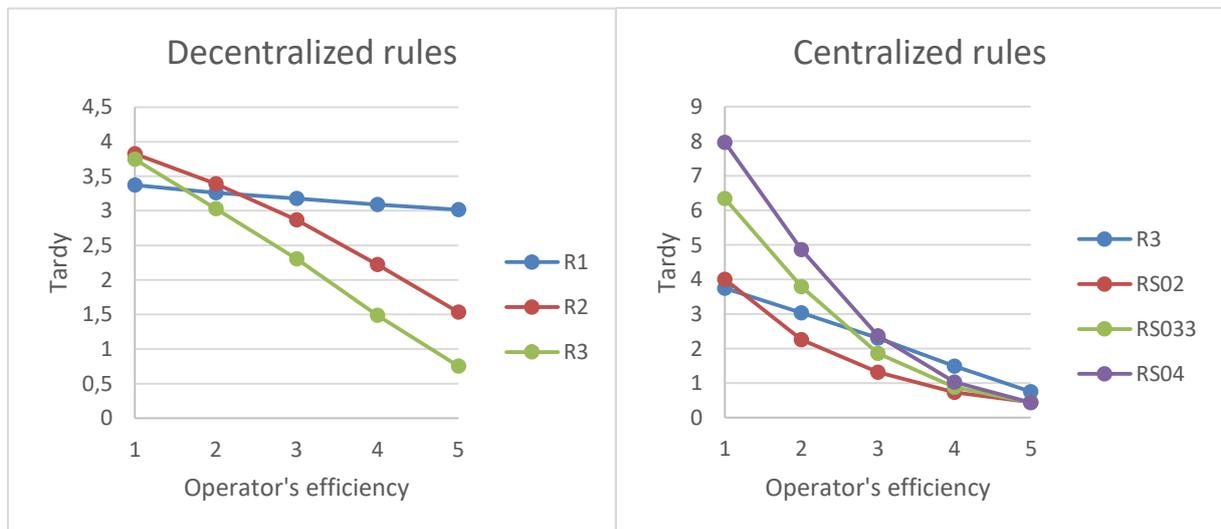


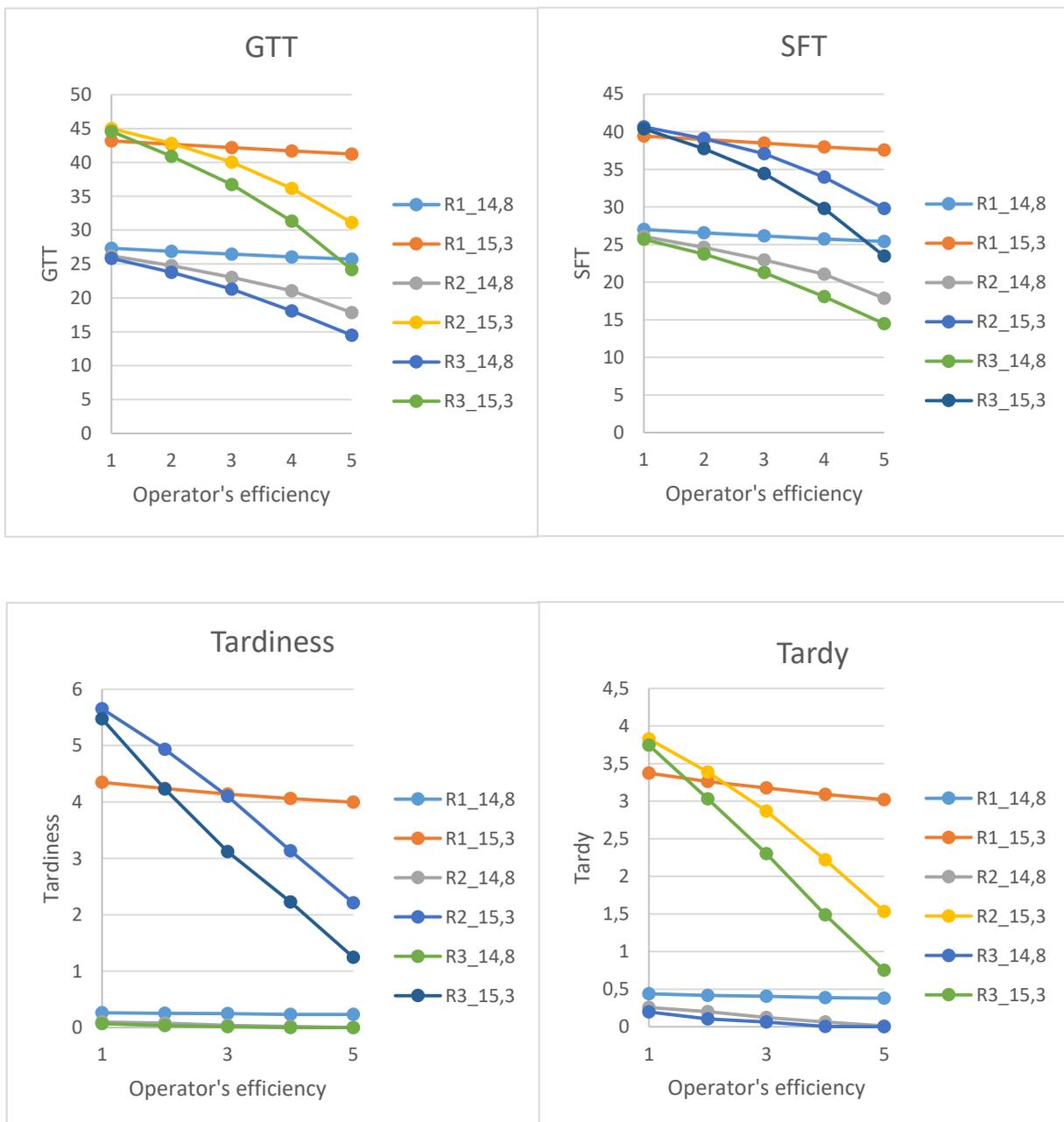
Figure 26 Tardy variation with increasing efficiency



Increasing saturation

In this paragraph, the effect of increasing efficiency in different saturation contexts is investigated. The graphs show that for GTT, SFT the performance improvement brought by the increase in efficiency, are very similar in high and low saturation context. For tardy and Tardiness, the impact of improving efficiency becomes very relevant for R2 and R3, it reaches 80% reduction for R3 in best efficiency conditions.

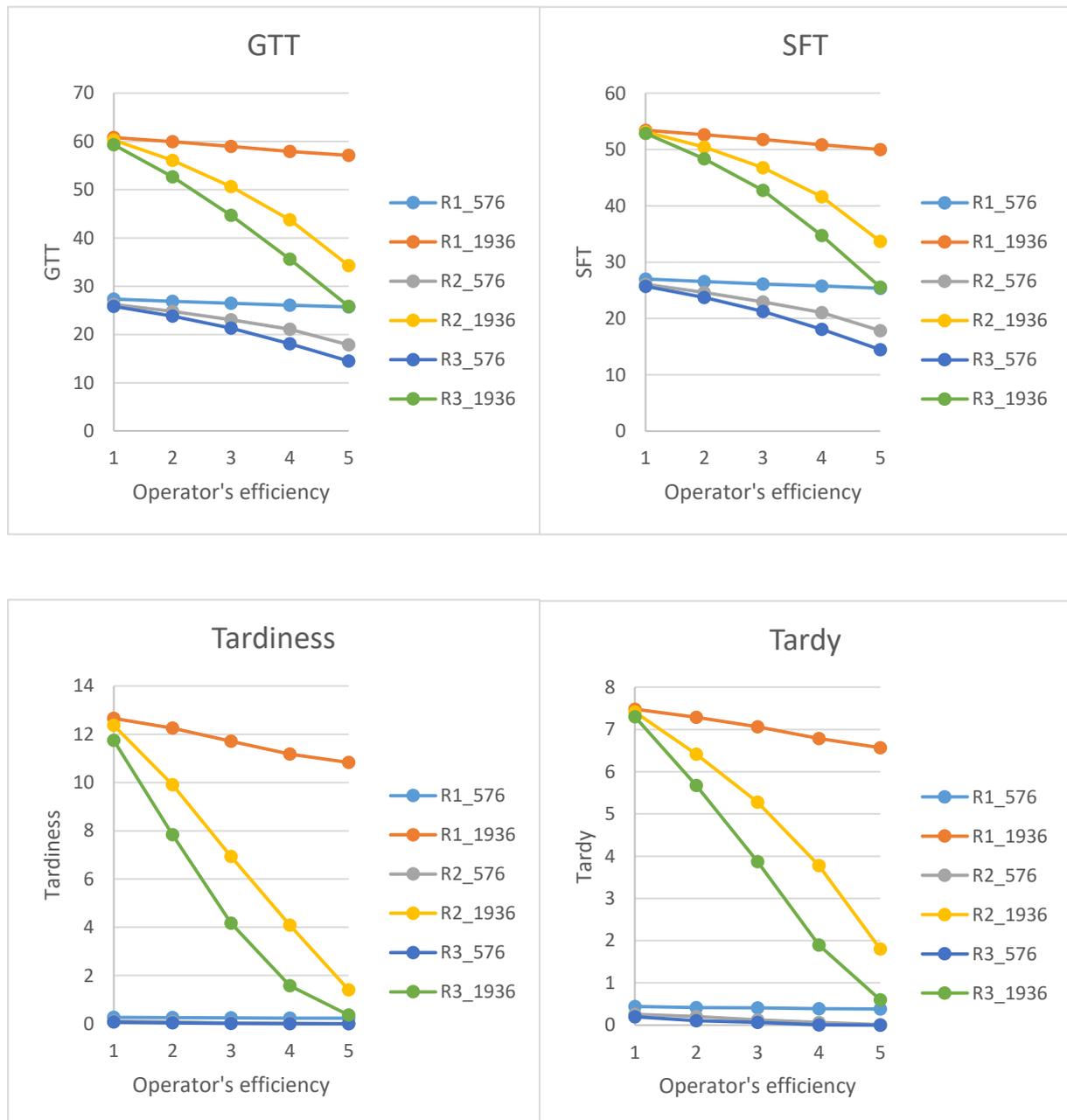
Figure 27 Shop performance with increasing saturation



Increasing variability

The results of increasing variability to shop performances are very similar to those occurred for an increase of saturation. The improvement caused by improving efficiency in high variability context are very similar to low variability ones for GTT and SFT. Tardiness and Tardy, instead, are very impacted by an improvement in efficiency starting from R2. In this case data shown in the graphs refer to simulations performed with low saturation.

Figure 28 Shop performance with increasing variability



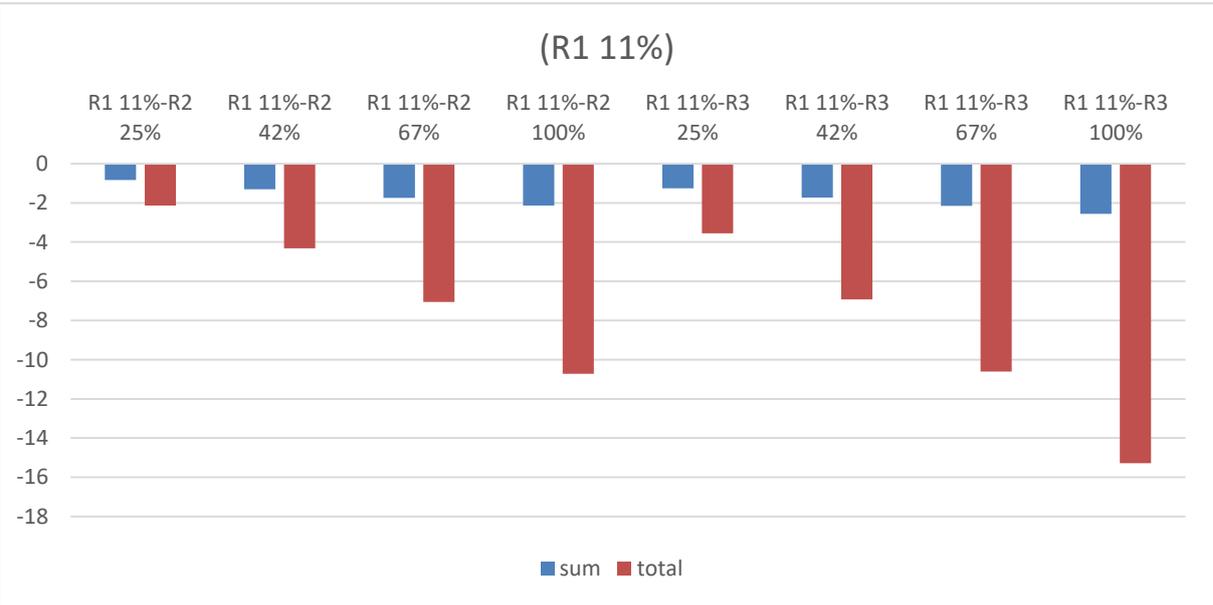
5.2.2. Synergy of worker flexibility and efficiency

The analysis on flexibility and efficiency has clarified the impact of these drivers on shop performances. It is interesting to evaluate the effect of the combination of the two. In order to tackle this issue, the benefit of improving these parameters is calculated. What we want to demonstrate is that the effect of the combination of efficiency and flexibility is higher than the sum of the effects of applying an efficiency and flexibility improvement separately. In the first matrix the benefit of shifting only in efficiency (vertically) or only in flexibility (horizontally) is calculated respect to (R1 11%). In the second matrix, instead, the benefit of shifting both in flexibility and efficiency is calculated again respect to (R1 11%). As Figure 29 shows the effect of efficiency and flexibility is much higher than the sum of the effects considered separately: there is a synergy between the two variables.

Table 6 GTT improvement respect to R1 11%

		RULE					RULE		
		R1	R2	R3			R1	R2	R3
EFFICIENCY	11%	0	-0,36	-0,78	EFFICIENCY	11%			
	25%	-0,47				25%		-2,12	-3,54
	42%	-0,94				42%		-4,31	-6,92
	67%	-1,37				67%		-7,05	-10,60
	100%	-1,78				100%		-10,71	-15,27

Figure 29 Sum of the GTT reduction caused by flexibility and efficiency improvement separately taken compared with GTT reduction for flexibility and efficiency together. The starting point is R1 11%



The same approach was used with changing starting point: (R2 11%), (R2 42%), (R2 67%). The synergy is confirmed.

Figure 30 Sum of the GTT reduction caused by flexibility and efficiency improvement separately taken compared with GTT reduction for flexibility and efficiency together. The starting point is R2 11%

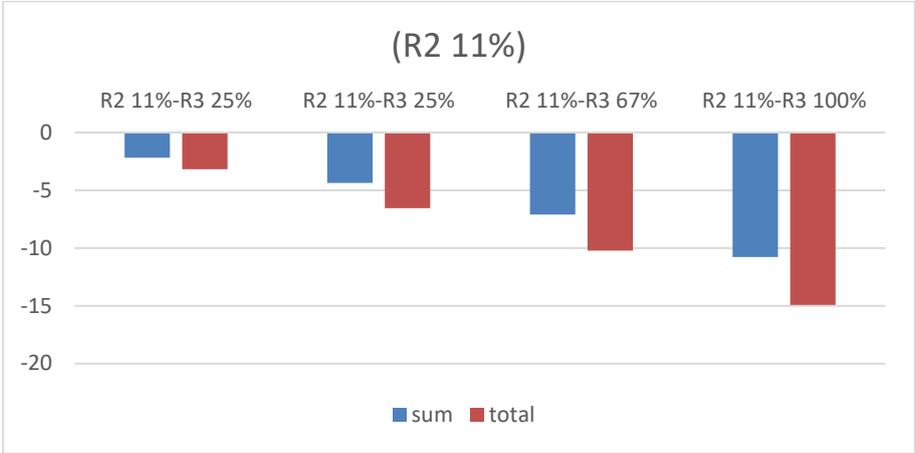


Figure 31 Sum of the GTT reduction caused by flexibility and efficiency improvement separately taken compared with GTT reduction for flexibility and efficiency together. The starting point is R2 42%

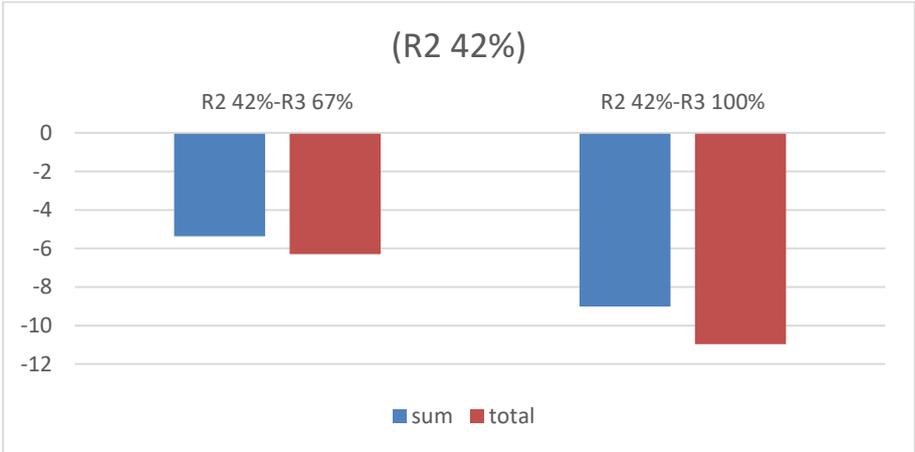
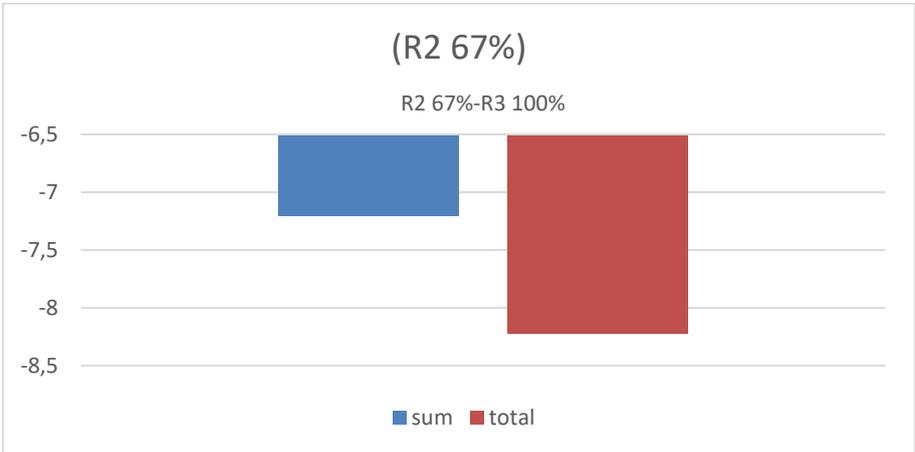


Figure 32 Sum of the GTT reduction caused by flexibility and efficiency improvement separately taken compared with GTT reduction for flexibility and efficiency together. The starting point is R2 67%



5.2.3. Heterogeneous efficiency

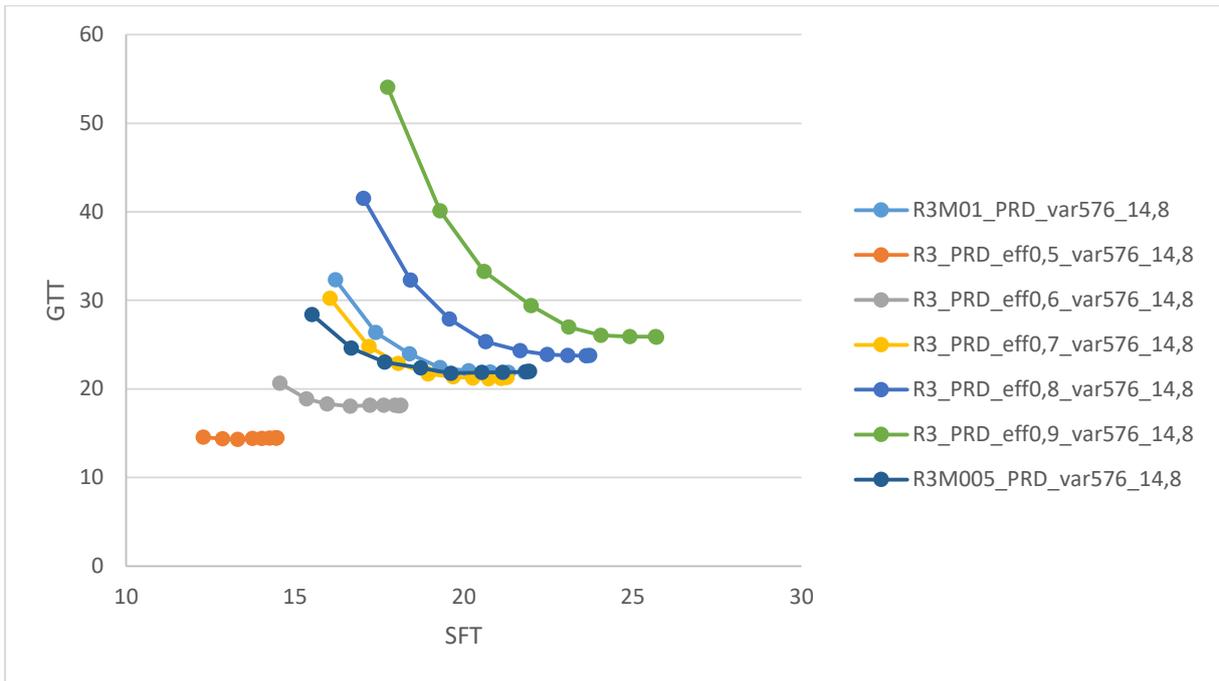
The following matrices in Figure 33 show the efficiency pattern of two tested simulation model for heterogeneous flexibility which means that the operator has different efficiency depending on the stations he works on. R3M01 implies an average reduction of 30% of the processing time while the average reduction with R3M005 is 40% of the processing time.

Figure 33 Heterogeneous efficiency matrices. Coefficients for processing time reduction.

		R3M01					R3M005				
		WORK STATION					WORK STATION				
		1	2	3	4	5	1	2	3	4	5
WORKER	1		0,6	0,7	0,8	0,9		0,55	0,60	0,65	0,70
	2	0,6		0,6	0,7	0,8	0,55		0,55	0,60	0,65
	3	0,7	0,6		0,6	0,7	0,60	0,55		0,55	0,60
	4	0,8	0,7	0,6		0,6	0,65	0,60	0,55		0,55
	5	0,9	0,8	0,7	0,6		0,70	0,65	0,60	0,55	

Figure 34 shows a GTT-SFT comparison of different efficiency patterns to assess the effectiveness of the heterogeneous efficiency model respect the most commonly used in WLC literature homogeneous efficiency pattern. The plot shows that R3M01 performs slightly worse than R3_eff0,7 (same average processing time reduction of R3M01). Instead R3M005 is significantly worse than R3_eff0,6 (same average processing time reduction of R3M005) and even worse than R3_eff0,7 for high workload norms.

Figure 34 Comparison between heterogeneous and homogeneous efficiency pattern



The same is observable for Tardy and Tardiness.

Figure 35 Tardiness comparison of homogeneous and heterogeneous efficiency patterns

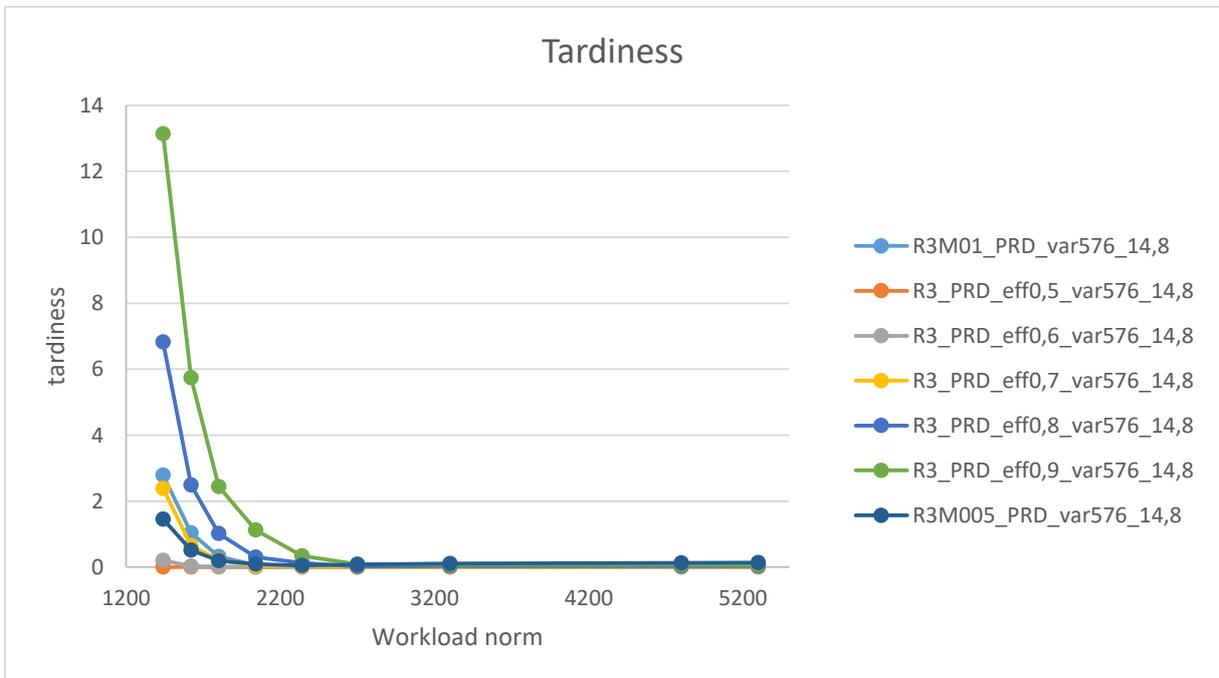
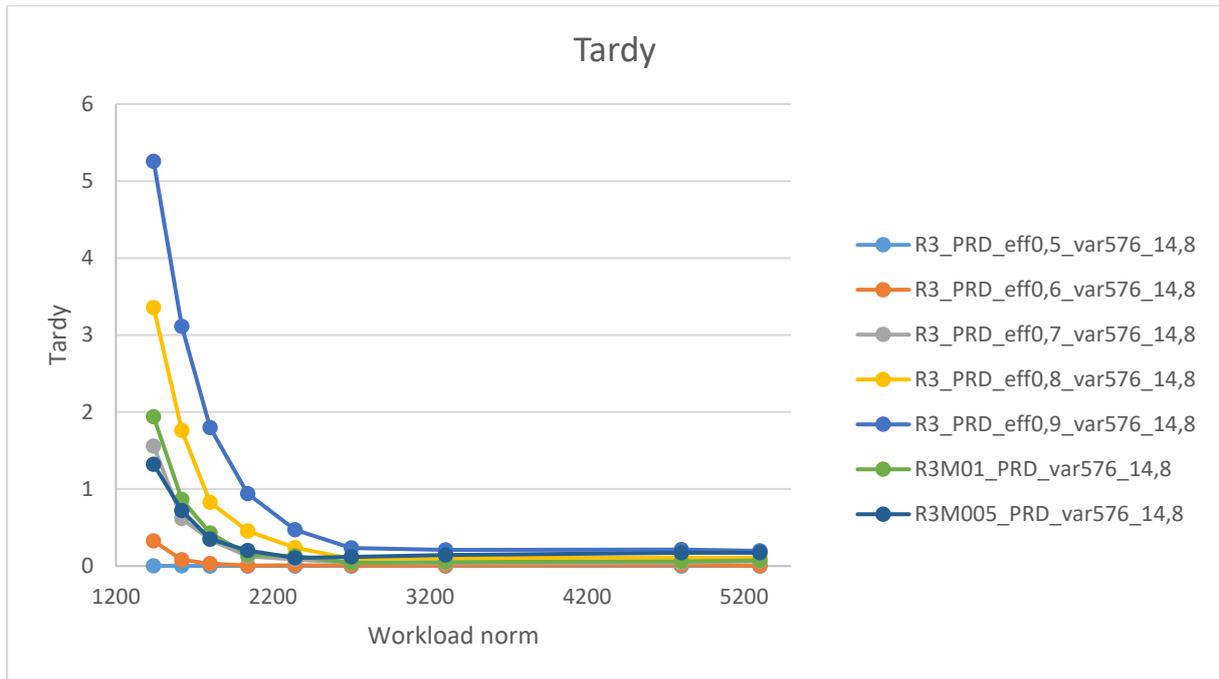


Figure 36 Tardy comparison of homogeneous and heterogeneous efficiency patterns



The impact of efficiency on GTT and SFT is not influenced by high saturation or variability environment. As Figures 37, 38 show, the GTT and SFT performance deterioration caused by heterogeneous efficiency is similar independently from the saturation or variability level.

Figure 37 GTT-SFT comparison of homogeneous and heterogeneous efficiency patterns with increasing saturation

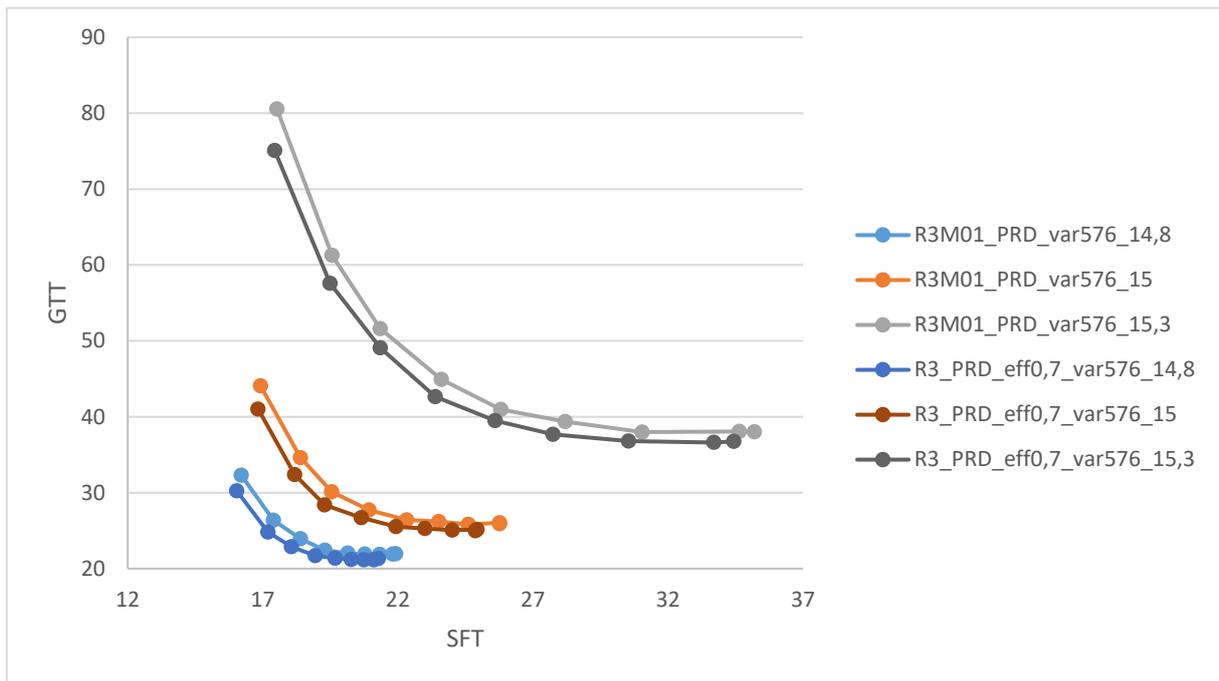
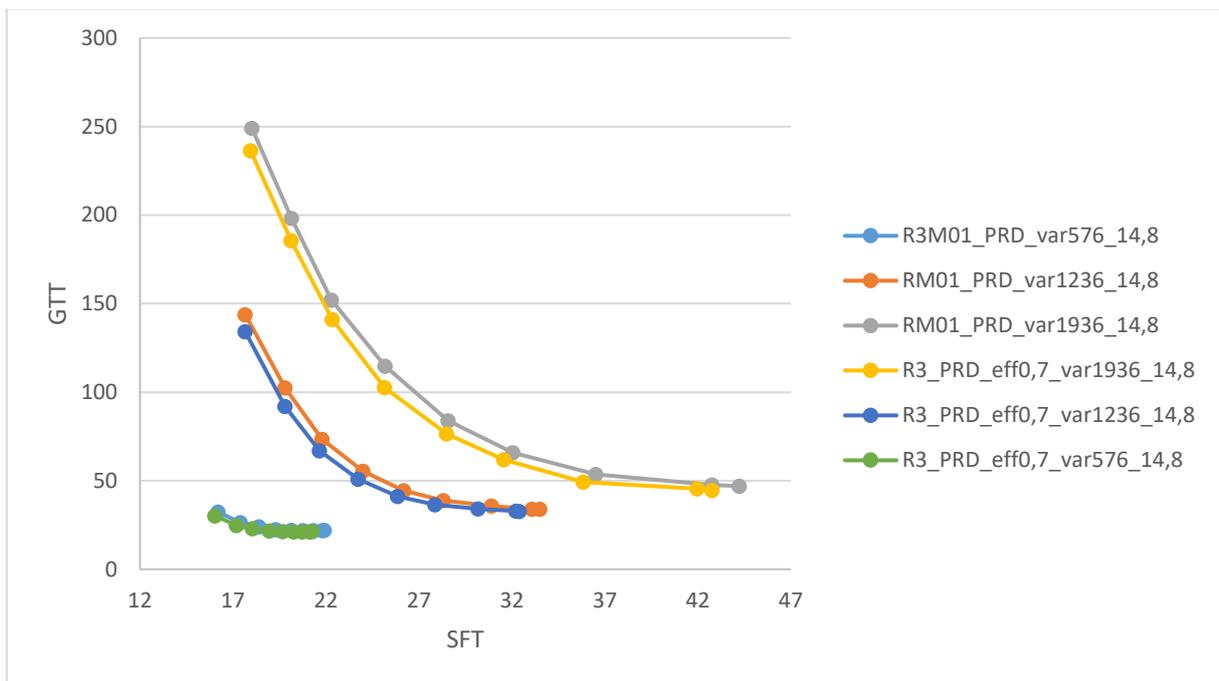


Figure 38 GTT-SFT comparison of homogeneous and heterogeneous efficiency patterns with increasing variability



5.3. RESEARCH QUESTION 3

Which “When” rule, decentralised rule transferring operators when idle or centralised rule allowing the transfer on the basis of the queue length, is most performing? And how do shop performances react to different load threshold when a decentralised “When” rule is selected?

Table 7 Centralized and decentralized “When” rules summary

RULE	FLEXIBILITY LEVEL	WHEN RULE	TRANSFER CRITERION
R3	5	Decentralised	Operator idle
RS02	5	Centralised	$\min workload \leq 0,2 * \max workload$
RS033	5	Centralised	$\min workload \leq (1/3) * \max workload$
RS04	5	Centralised	$\min workload \leq 0,4 * \max workload$

The third research question aims at the definition of system behaviour when different when rules are implemented.

At first the analysis will focus on the relation between R3 and one of the three rules implementing a centralised when rule, RS033. Then in the following section the impact of different levels of workload threshold triggering operator’s transfer will be studied.

5.3.1 Centralised vs decentralised when rule

Figures 39 and 40 show the comparison of flow time performance for the centralised rule RS033 and the decentralised rule R3 with saturation 14,8 (Figure 39) and 15,3 average orders per day (Figure 40). Gross throughput time results are plotted against the relative shop floor time. Increasing levels of SFT correspond to increasing levels of workload norm implemented in the ORR mechanism. Therefore, nine combinations for each rule are plotted defining the corresponding curves. With low saturation and maximum level of flexibility (flexibility = 5), the system achieves almost the same performance in terms of GTT and SFT independently to the when rule implemented. In Figure 35 the curves are horizontal and almost overlapping. This means that for every workload norm and for both when rules the average time between the order entry in the pre shop pool and the departure of the finished job from the shop floor is constant. Moreover, the range which shop floor time varies in is very limited (between 12 and 15 hours). Data prove that in this low saturation environment, with flexibility equals to five, the selection of one or the other workload norms and one of the two when rules to regulate

operators transfers to other stations yields almost identical results. A different situation is presented in Figure 40 where saturation is 15,3 average orders per day. In this case a significant gap between when rules effects is clear. RS033 outperforms R3 in terms of GTT for every workload norm selected, and average shop floor time is slightly better when greater levels of workload norm are selected.

Figure 39 Centralized and decentralized rules comparison of GTT-SFT. Low saturation.

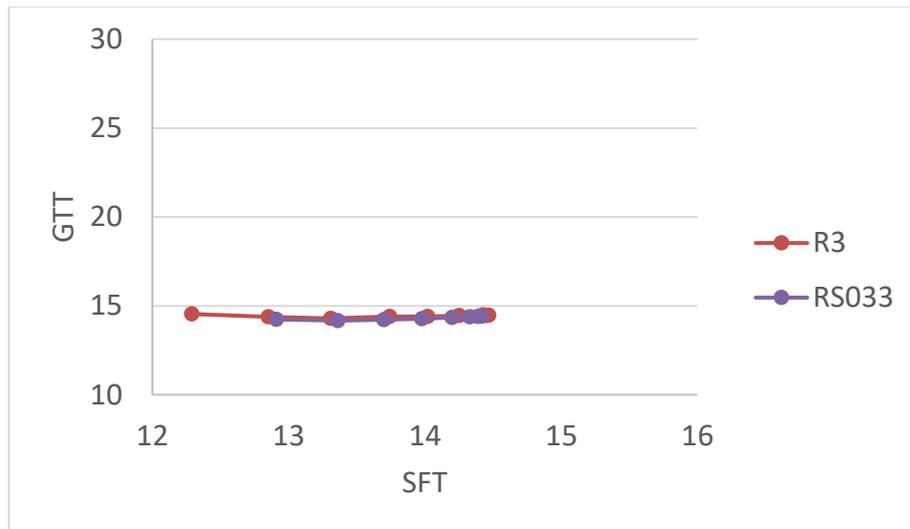
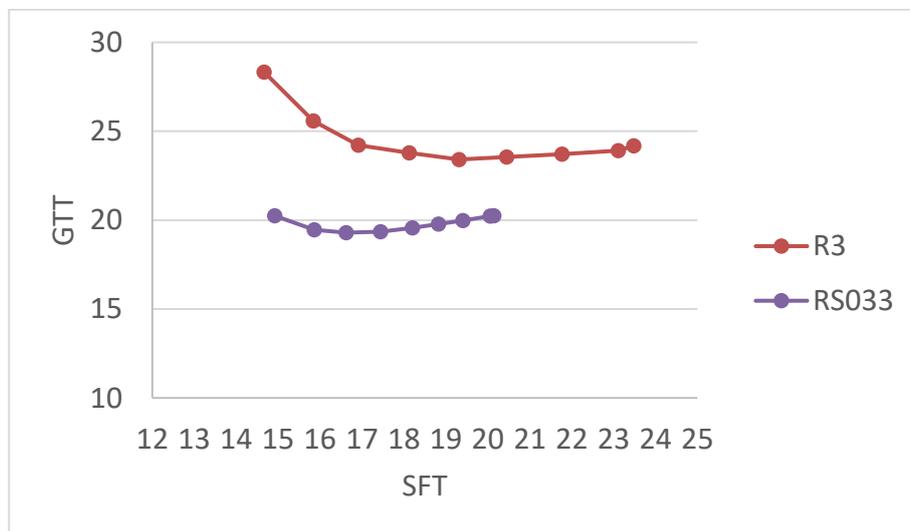


Figure 40 Centralized and decentralized rules comparison of GTT-SFT. High saturation.



In high variability contexts flow time performances are significantly worsened. However, RS033 proved to be more effective than R3 in reducing GTT and SFT and the analysis made stays valid. The relative graphs are available in Appendix B.

As it was highlighted before, in low saturation context and maximum flexibility, the plant simulated has enough capacity to face the demand achieving great results with every

configuration adopted. Workers flexibility more easily enables the respect of due dates, too. For this reason, the average tardiness and the number of tardy orders are negligible with saturation 14,8. Higher demand leads to higher flow time in the system and the effectiveness of the system in delivering finished products within the due date of 7 days is affected, too. Figures 41 and 42 compare R3 and RS033 in average tardiness and tardy orders. Performance data are plotted against the nine workload norm levels. The patterns in the two graphs are equivalent and show that with the centralised when rule the system is more effective in respecting due dates than with the decentralised one. The value of tardiness reported are almost null for RS033, but still acceptable for R3. The same consideration can be done for the number of tardy orders. With greater processing time variability these performances are subject to an inevitable worsening. RS033 performs much better than R3 and the performance gap is more important in absolute value. The greater difference is observed in combination with the lowest levels of workload norm included in the study. The graphs concerning higher variability are available in Appendix C.

Figure 41 Centralized and decentralized rules comparison of Tardiness with high saturation

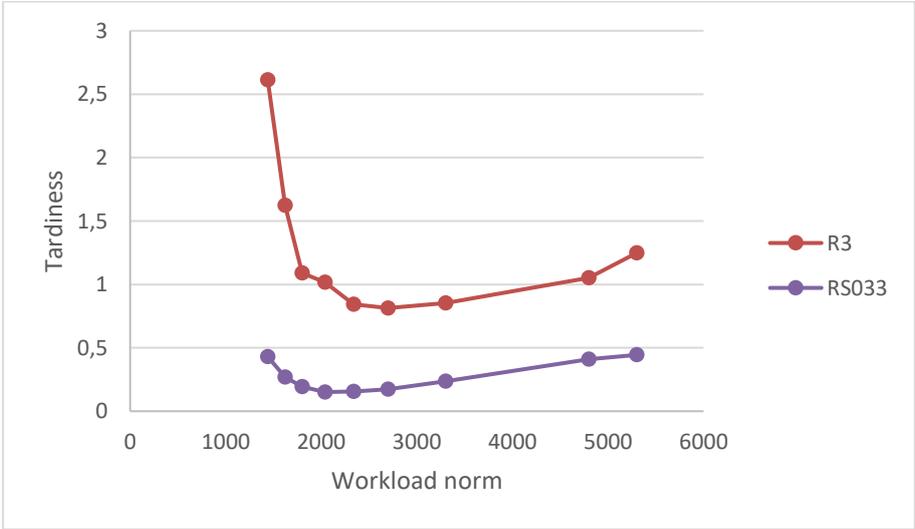
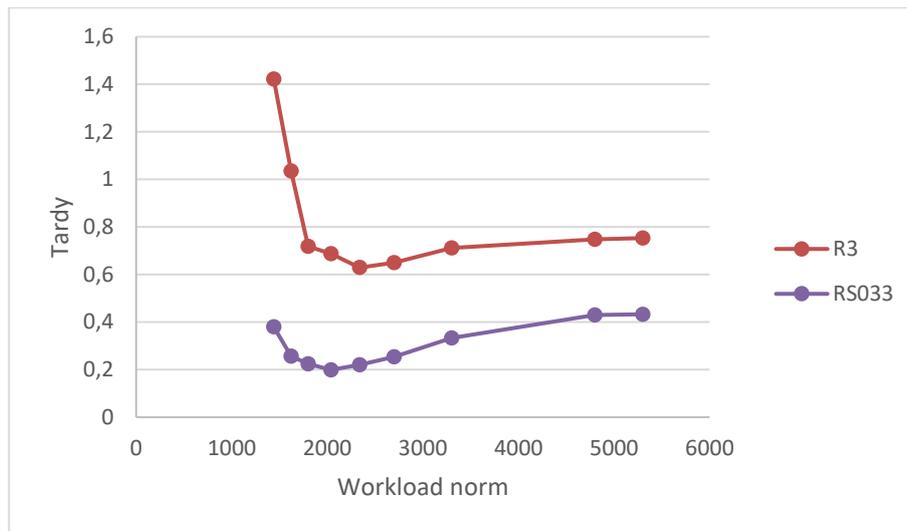


Figure 42 Centralized and decentralized rules comparison of Tardy with high saturation



Low operators' efficiency

The analysis carried out in the previous paragraph considered fully flexible operators able to work on each station with 100% efficiency. In this section the impact of a centralised and a decentralised rule will be compared when the productivity of workers is low when they are transferred among the shop floor. In Figure 43, 44 and 45 system performances with saturation 14,8 and the lowest workers' efficiency tested are shown. When transferred operators determines an order's processing time reduction of 10%. Figure 43 shows GTT against SFT, while tardiness and tardy orders are plotted against the workload norm as before.

The first observation is that the conclusions drafted in the previous paragraph in this new scenario are totally inverted. The centralised rule proved to be more effective than the decentralised one in every condition of workload norm, saturation and processing time variance. Now, with low workers' efficiency, the decentralised rule R3 outperforms RS033 in every performance criterion taken into consideration. With the centralised rule the occurrence of worker movements within the shop floor is higher than the number of transfers when the condition is the total lack of work in the queue at the workstation. When the training of operators in performing the work on other stations is lacking and his productivity produces a limited processing time reduction (in this case only 10%), transferring an operator causes his absence from the assigned workstation for a relatively long period of time (90% of

order’s processing time) and the impossibility to work on the jobs waiting in his queue. The benefits obtained are limited (10% processing time reduction) and may not exceed the delay of work at the operator’s station. If the movement is permitted even if there are jobs at the station (as it is with the decentralised when rule) this problem is more serious than if the worker can leave only when completely free of work. This is the reason why RS033 produces worse results than R3 with low operators’ efficiency.

A further conclusion that can be deduced from the graphs is that not only RS033 is worse than R3, but it is also worse than R0 which is the scenario where no cross-training is delivered to the workforce and no operators transfer is allowed. Even R3 yields almost identical results to R0. This suggests that, under a certain level of workers’ productivity, enabling operators to move from their position to help others is not beneficial with the decentralised when rule and it even worsens system performances when a centralised when rule is selected.

Figure 43 Centralized and decentralized rules comparison of GTT-SFT with low efficiency

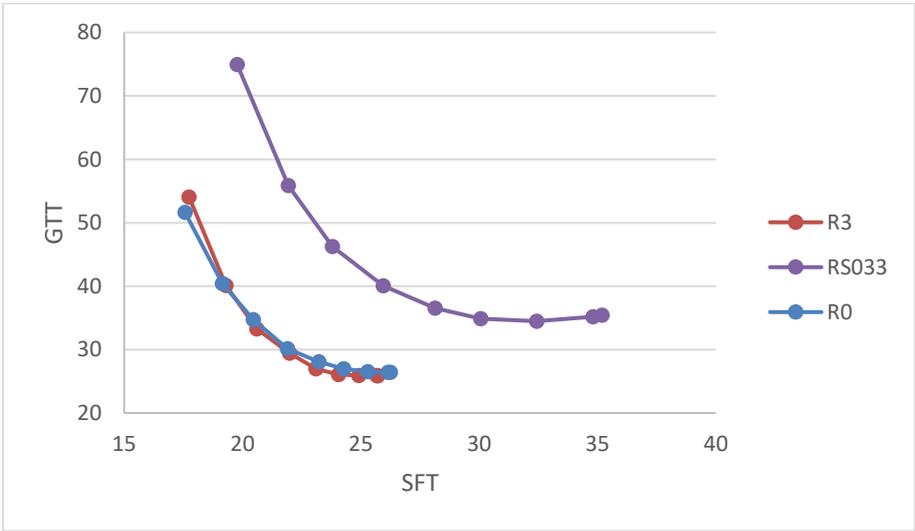


Figure 44 Centralized and decentralized rules comparison of Tardiness with low efficiency

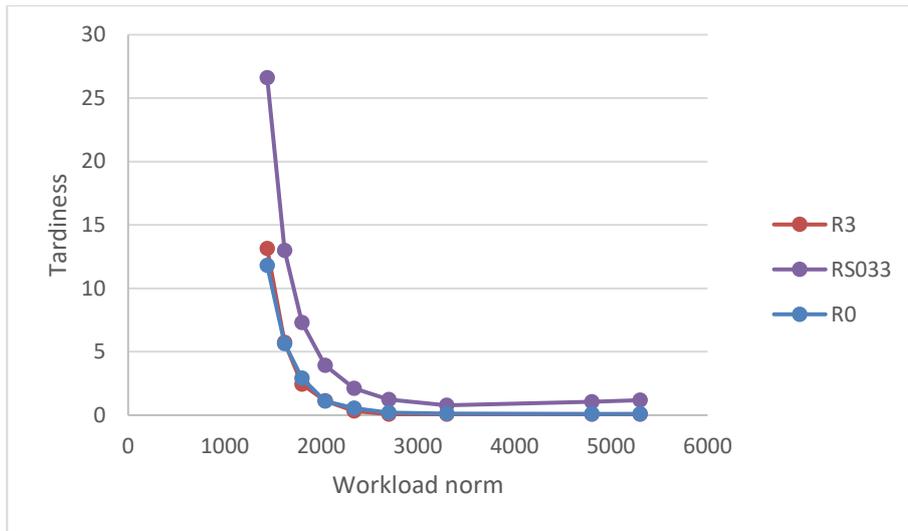
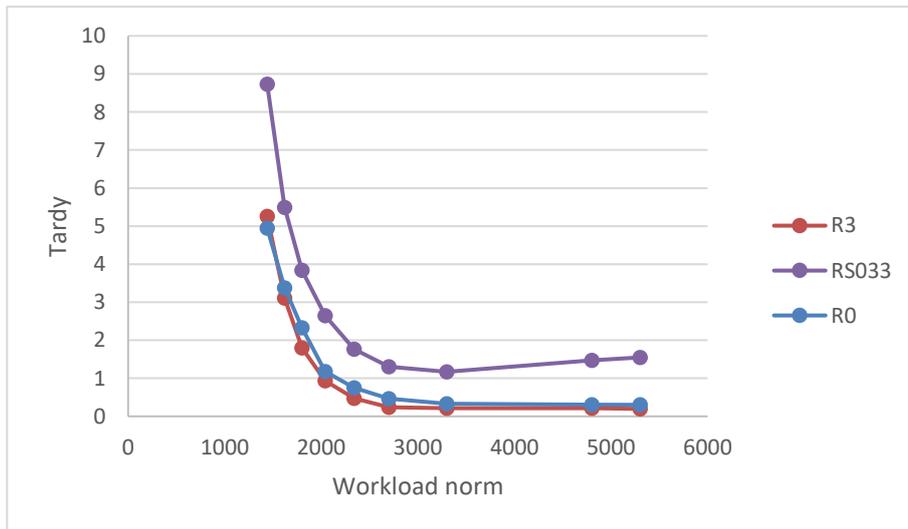


Figure 45 Centralized and decentralized rules comparison of Tardy with low efficiency



5.3.2. Workload threshold comparison

Considered the low differences in orders' flow time between R3 and RS033 highlighted before and the negligible values of tardiness and tardy orders per day with low system saturation 14,8, performances in contexts with maximum saturation 15,3 are presented for the comparison between the three workload threshold used. As it is possible to observe in the graphs of Figures 46, 47 and 48, the three threshold triggering operators' transfer do not produce significant performance differences. RS033 and RS04 behave almost identically and the two curves overlap in every graph and in correspondence of almost every value of workload norm, while RS02 shows slightly inferior results. However, the differences in

absolute value of GTT are minimal and the value of tardiness and tardy are so low that it makes no sense to state that results show significant performance gap between the three configuration studied.

In previous paragraphs the effect of processing time variability has been already presented. Increasing variance of the time that operators need to perform different jobs causes an increase in orders' flow time and a worsening in the ability to respect due dates. In this scenario, where a centralised when rule is implemented with three workload threshold, processing time variability has the same effect. GTT and SFT grow, average tardiness and number tardy orders increase. Moreover, the gap between RS02 and the other two rules RS033 and RS04 is amplified so that absolute values of performance differences are more significant. This confirms the consideration deduced by the curves with moderate processing time variance that a workload threshold of 0,2 is outperformed by the other two values of this variable ((1/3) and 0,4).

In the presence of workers that are cross-trained but have low efficiency (allowing a processing time reduction of only 10%), RS02 has less poor results than the others. But, as it was stated before, a fully flexible workforce with these characteristics does not bring benefits to the system and may even have negative effects on overall system results.

Figure 46 Centralized rules comparison of GTT-SFT with high saturation

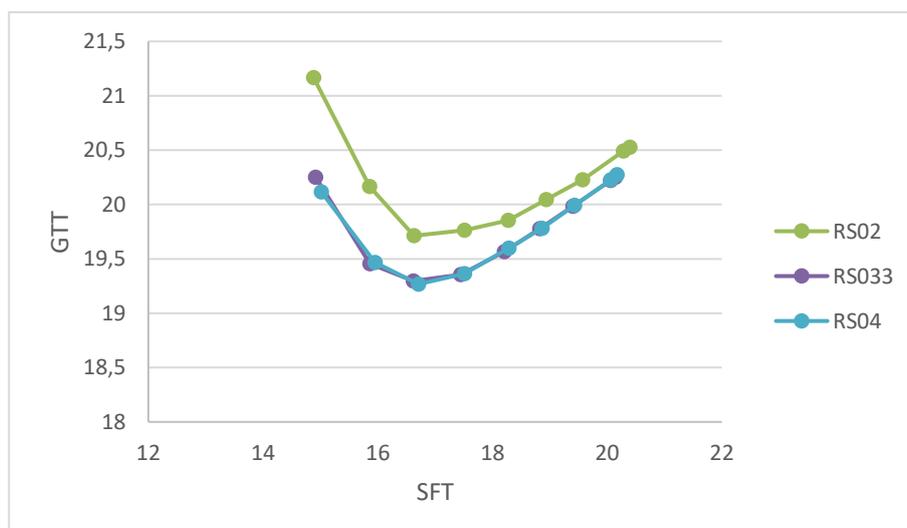


Figure 47 Centralized rules comparison of Tardiness with high saturation

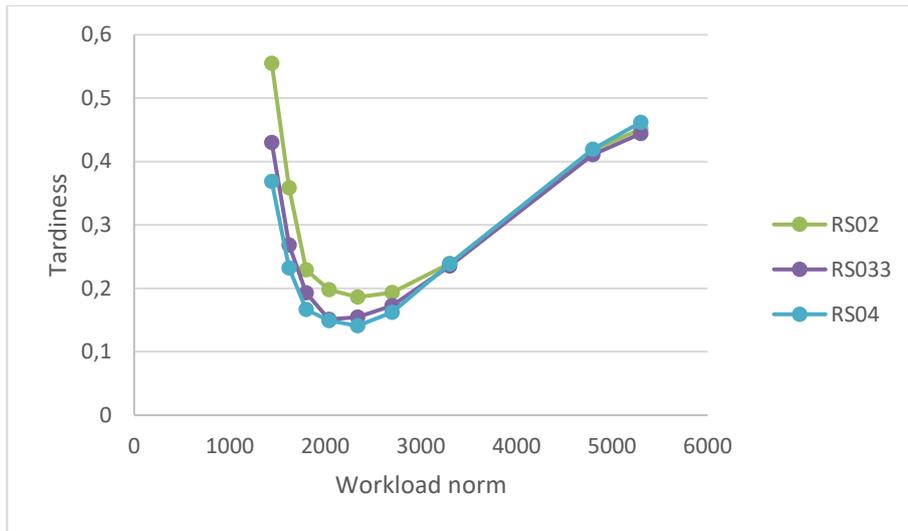
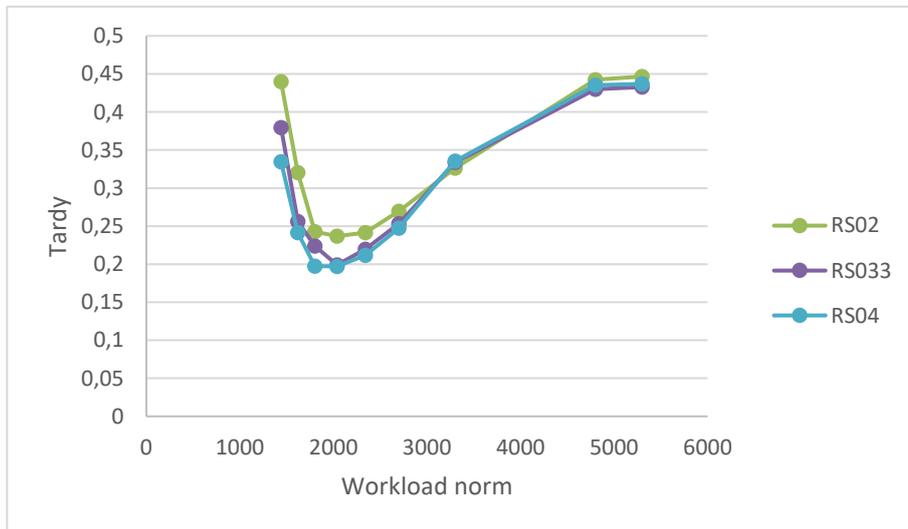


Figure 48 Centralized rules comparison of Tardy with high saturation

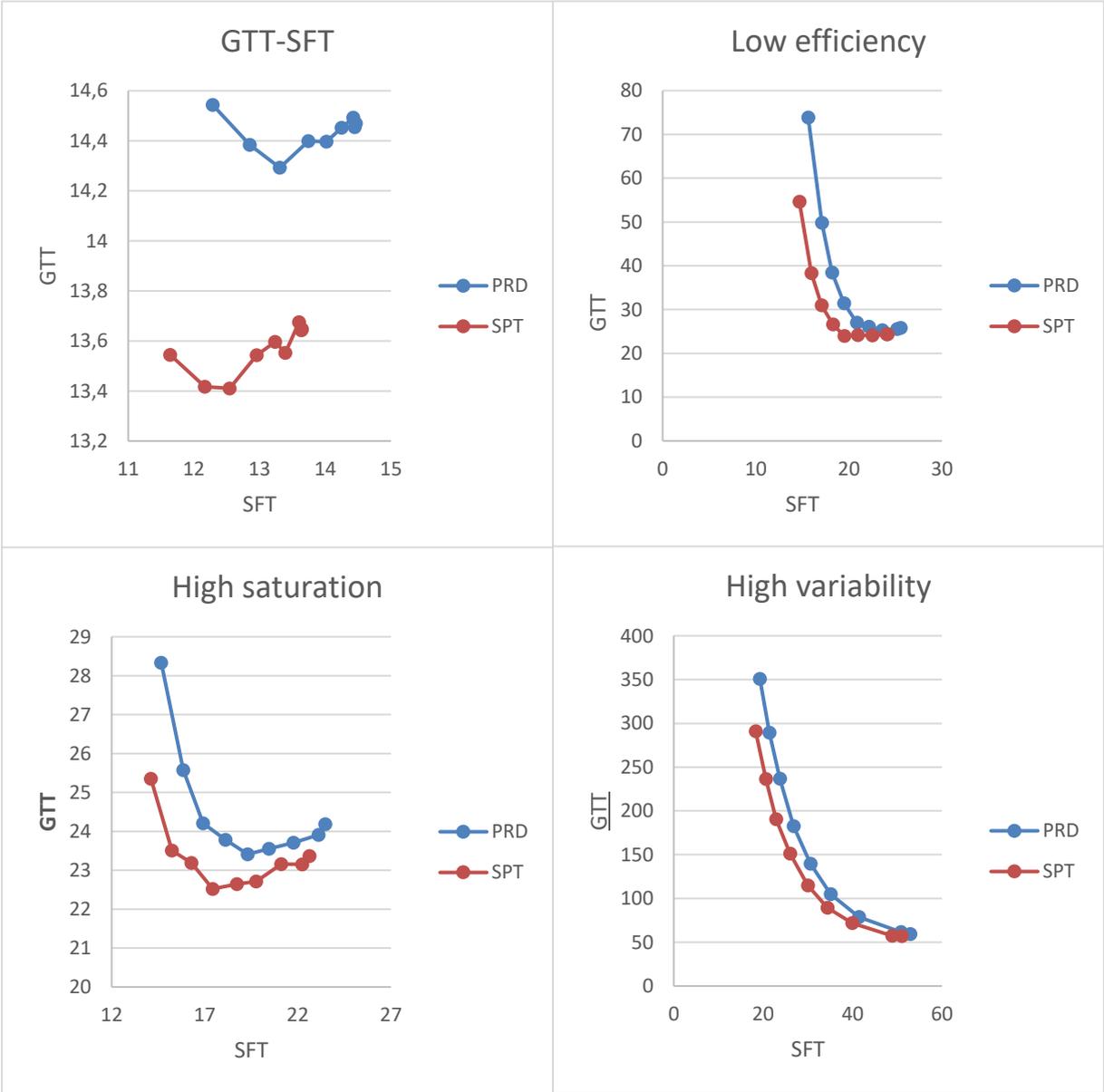


DISPATCHING RULES

The objective of this paragraph is to discuss the effect of applying a different dispatching rule by assessing performances with variable input parameters. The dispatching rule is responsible for the order in which jobs are reviewed for dispatching. Shortest Processing Time (SPT) orders jobs by increasing processing time, while Planned Release Date (PRD) prioritizes those orders that have an earlier date of release. Since the model assigns the same release date to orders arriving the same day, it corresponds to a first come first served logic.

As displayed by the Figures 49, 50, 51, 52 below, SPT always outperforms PRD in terms of GTT-SFT at any level of efficiency, variability and saturation. This confirms literature results (Jayamohan and Rajendran 2000). However, it must be noted that the percentage GTT reduction caused by the implementation of SPT rule is around 6% in standard conditions, while it reaches 26% for low efficiency and low workload norm, 17% with high variability, and 10 % with high saturation. Instead, the SFT improvement of one rule respect to another keeps fairly constant to 5% in any environment. This means that SPT has the capability of significantly reducing the average waiting time in the pre shop pool.

Figure 49 Comparison of PRD and SPT dispatching rules. GTT_SFT performances in different conditions



In standard conditions, both Tardy and Tardiness are very near to zero. By increasing variability, saturation or decreasing operators' efficiency, the Tardy and Tardiness performances are impacted for low workload norms. The graphs below show that SPT outperforms PRD. However, the difference is less and less significant when increasing workload norm.

Figure 50 Comparison of PRD and SPT dispatching rules. Tardiness and Tardy performances with low efficiency

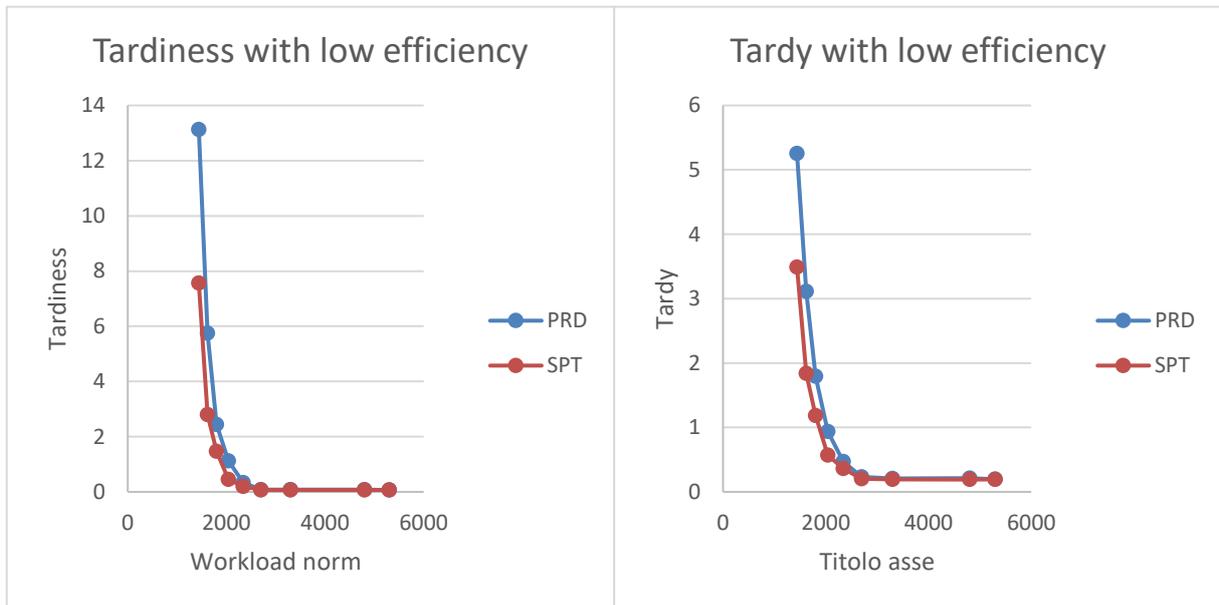


Figure 51 Comparison of PRD and SPT dispatching rules. Tardiness and Tardy performances with high variability

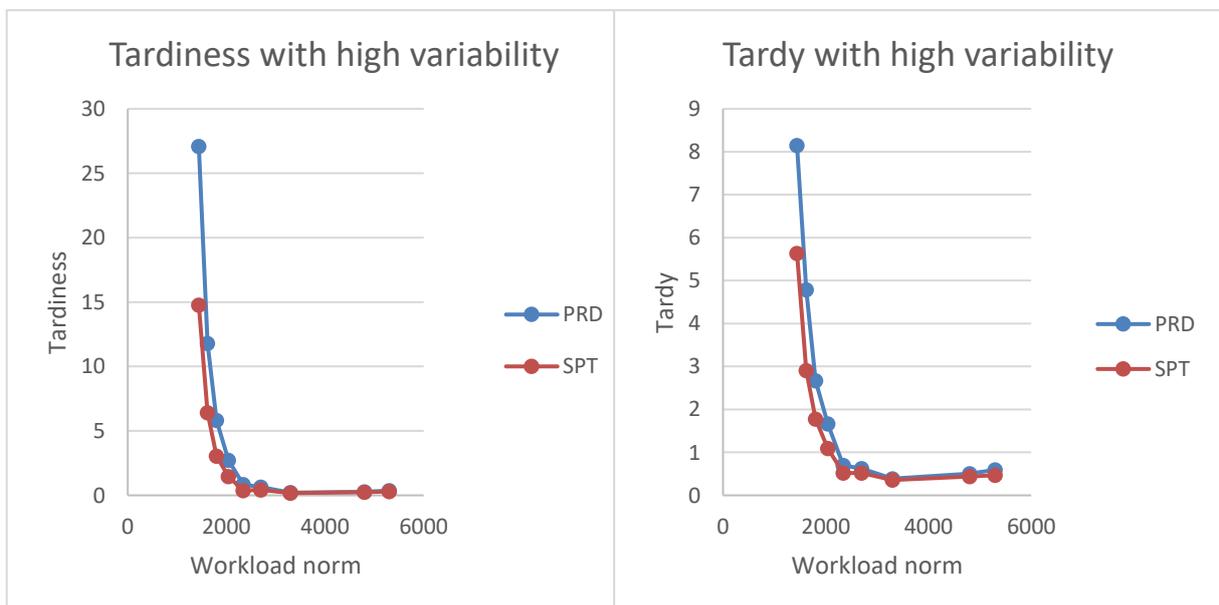
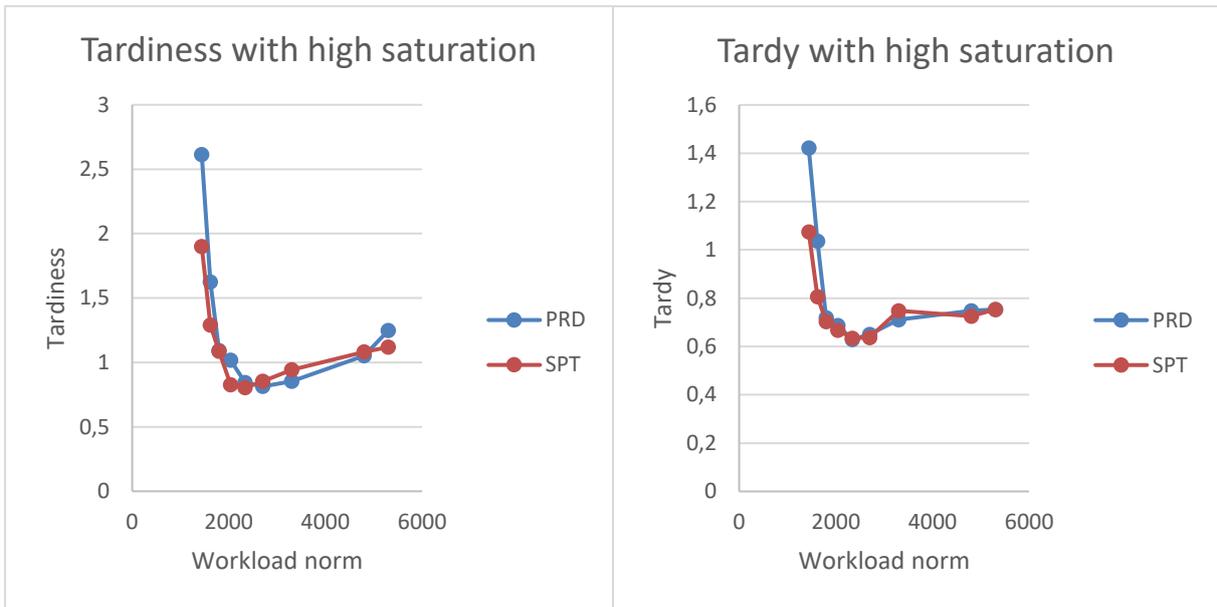


Figure 52 Comparison of PRD and SPT dispatching rules. Tardiness and Tardy performances with high saturation



6. DISCUSSION AND CONCLUSIONS

1. What is the impact of incremental flexibility on average orders gross throughput time (GTT), shop floor time (SFT), tardiness and number of tardy orders, when combined with an ORR method for input control?

The first research question aimed at the assessment of the effect that different levels of worker flexibility have on the job shop system's simulation model developed and used. In the model, an ORR mechanism regulates order's arrival and storage in the pre shop pool and the release to the shop floor according to the overall aggregate workload already in the shop at each work station. This solution is thought and implemented to reduce the occurrence of work load imbalances between different productive stages. Within the shop floor, a reactive solution to work load imbalances is represented by cross-trained workers' reallocation to other stations when their load is under a certain threshold or null. With these characteristics, the model allows to study the effect of a flexible workforce in a productive system that combines an input control mechanism with an output control solution. Literature studies of worker flexibility effects in similar system are few and lacking. One of the aims of the research was to fill this literature gap and clarify the contradictory results on incremental flexibility's benefits.

Data collected through the simulations confirm the beneficial effects of worker flexibility on each system performance evaluated in the study (GTT, SFT, average tardiness and number of tardy orders). As it was demonstrated by previous studies (Park 1989; Park 1991; Felan 2001; Givi 2015), the marginal improvement of incremental flexibility decreases after a certain flexibility level. Indeed, data show that shifting from flexibility level 2, meaning that operators are able to work on two different work stations, to 3 has the greater benefit in terms of gross throughput time reduction: -35% against a further -11% due to the effort in cross-training the workforce beyond that level until a fully flexible workforce (flexibility = 5). The same conclusions can be drafted by shop floor time performance and average tardiness and number of tardy orders. This result complies with those obtained by Givi (2015) while in Park (1989; 1991) and Felan (2001) the maximum benefit was recorded with the minimum level of flexibility (flexibility = 2). However, all previous study observed negligible marginal improvement after the most beneficial flexibility level. In this research, cross-training operators until the maximum flexibility, so that each worker can perform the job on each

work station, still has a significant, though lower, impact on shop performances. This result is due to the high system saturation levels used in the simulations. Even between the three tested saturation, the difference in the impact of the shift between flexibility 2 to 3 and 3 to 5 decreases as saturation increases.

A further relevant finding of this research is how worker flexibility affects shop performances when orders' processing time variability increases. Data show that increasing processing time variance has a negative impact on the results of the production system with every flexibility rule tested. However, the performance deterioration is much faster without worker flexibility and decreases with the increase of the level of operators' cross-training. In conclusion, incremental worker flexibility has beneficial effects on system's performance and worker reallocation among different station in the shop floor even smooths the impact of workload imbalances caused by high processing time variability.

2. How do different levels of operators cross-training and efficiency affect shop performance? What is the impact of considering a heterogeneous flexibility pattern for operator cross training?

Increasing operator efficiency for cross training is very beneficial for production performances. Gross throughput time can be reduced up to 50% with decentralized rules. Centralized rules outperform decentralized ones for high efficiency levels and their responsiveness to operators' efficiency improvement can reach 65% GTT reduction. The marginal improvement for centralized rules is at its highest for low efficiency and decreases gradually. Instead, the marginal improvement for decentralized rules steadily increases and reaches its highest value for the best efficiency level. The different behaviour of the two rules is caused by the larger number of helps caused by centralized rules: more operator reallocation with low efficiency are not very beneficial. They also risk to prevent operators to work on their own station at highest efficiency. As a result, the first efficiency improvement is very advantageous, while the following are less and less effective.

The same significant improvements are observed on other performances SFT, Tardiness and Tardy. It must be noted that for centralized rules, reaching the highest level of efficiency can bring to a near to zero level of Tardiness and Tardy orders. Following these results, managers

shall select a different reallocation rule on the basis of the actual proficiency of the operators, and be ready to redefine their decision after training workers.

It must also be noted that the advantage of improving efficiency is greater with specific internal or external conditions (saturation and processing time variability). Managers shall be aware that improving operators' efficiency is an effective lever to face these environments.

Moreover, the analysis on flexibility and efficiency has clarified the impact of these drivers on shop performances. What we have demonstrated is that the effect of the combination of efficiency and flexibility is higher than the sum of the effects of applying an efficiency and flexibility improvement separately.

The use of a heterogeneous flexibility matrix as operators' efficiency pattern slightly worsens all the performances of the job (respect to a homogeneous matrix with the same average efficiency). The reason is that the disadvantage brought by the operators with worse-than-average efficiency when they can't work on their own station because they shifted to another for a long time is not compensated by the advantage of those operator having better-than-average flexibility. However, the pattern is more realistic since it models the actual variability of operators' proficiency when working with different machines. Thereby reflecting different personal capabilities in facing different technological processes.

3. Which "When" rule, decentralised rule transferring operators when idle or centralised rule allowing the transfer on the basis of the queue length, is most performing? And how do shop performances react to different load threshold when a decentralised "When" rule is selected?

The third issue addressed in the study is the rule and criteria to control operator movement within the shop floor. Two when rules were tested which are the most common and used in the literature: centralised when rule and decentralised when rule. The former evaluates operator's eligibility for transfer by considering the load of work waiting to be processed in the queue in front of his station. The latter enables an operator to move to help another in processing an order only if no job is waiting at his station.

Both rules were tested with the maximum flexibility level (flexibility 5) and, with high efficiency of cross-trained workers, produced significant system's improvements. In low

saturation context, results are too close to spot a noteworthy difference. With the maximum saturation level, the centralised rule proved to yield better results in every system performance demonstrating greater effectiveness in improving both orders flow time and ability of the system to respect due dates. This conclusion fails when operators have a limited efficiency in performing the work on other stations. The centralised when rule even worsens systems performances respect to the standard situation in which workers cannot move to other stages, while the decentralised rule does not produce significant improvement.

The three workload thresholds tested had similar results. The lowest threshold (0,2) yielded slightly worse results, while (1/3) and 0,4 performs identically.

During the decision making process for the selection of the when rule regulating operator transfers, production managers have to consider the specific characteristics of the productive system's workforce. The exploitation of worker flexibility and of the benefits of a centralised when rule depends on operators' ability to perform different tasks. An adequate effort in cross-training is necessary to avoid the risk of implementing solutions that increase managerial complexity but prove to be ineffective or even have detrimental effects on production performance.

The comparison of the two dispatching rules PRD and SPT in this model, in which an input control method and an output control mechanism are implemented, concluded that order sequencing by increasing processing time yields better results than the standard rule PRD that, in this model, corresponds to a first come first served rule. This complies with previous literature findings (Jayamohan and Rajendran 2000).

LIMITATIONS AND FUTURE RESEARCH

This thesis contributes to the literature and the body of knowledge on workload control with a specific focus on output control through worker flexibility in a productive system in which order release to the shop floor is regulated by an ORR mechanism. For the aim of the study a simulation model including both these workload control methods has been used. In the model, operators are able to transfer from their assigned position to other stations producing

a processing time reduction of the order being processed at that station. No transfer delays and productivity losses due to the worker movements among the shop floor are considered. In productive systems where the dimensions of the plant and the distance between departments and workstations determines long transfer delays this issue should be included in the definition of the actual benefit of reallocating operators. Worker productivity is constant along the entire simulation period independently on the fatigue that may be collected throughout the working day. Despite it was not the main focus of the research, this other aspect can be included in the model. Maintaining the focus on cross-trained operator efficiency the main conclusions of this thesis were collected by considering an equal level of efficiency for every operator. This limits the validity of conclusions since it lacks of realism: operators usually have different productivity in performing different types of job. They may be realistically more productive in similar job, while less productive when required to work out of their main area of expertise. The introduction of the heterogeneous efficiency concept is a significant step to fill the gap, but the analysis and the practical implication of different operator proficiencies can be deepened and completed by further studies and researches. No cost for operator cross-training has been considered in order to evaluate from an economical perspective the actual benefits brought to the system by a flexible workforce. Finally, worker flexibility has been studied in combination with one specific ORR mechanism. The effects can be confirmed and studied with other ORR mechanism to have more general and comprehensive conclusions. A further step in workload control research can include and combine input control with output control in the release mechanism implemented by production planning.

7. REFERENCES

Albin, S., 1982. On Poisson approximations for superposition arrival processes in queues. *Management Science*, Volume 28, p. 126–137.

Baker, K. R., 1974. *Introduction to sequencing and scheduling*. New York: John Wiley & Sons.

Baker, K. R., 1984. The effects of input control on the performance of a simple scheduling model. Volume 4, p. 99-112.

Baykasoğlu, A. & Gçken, M., 2010. A simulation based approach to analyze the effects of job release on the performance of a multi-stage job-shop with processing flexibility. *International Journal of Production Research*, 10 February.

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

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

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

Bertrand, J. & Van Ooijen, H., 2002. Workload based order release and productivity: a missing link. *Production Planning and Control*, 13(7), p. 665–678.

Bertrand, J. & Wortmann, J., 1981. *Production Control and Information Systems for Component-Manufacturing Shops*. Elsevier.

Bobrowski, P., 1989. Implementing a loading heuristic in a discrete release job shop. *International Journal of Production Research*, Volume 27, p. 1935-1948.132

Bertrand, J. & Van Ooijen, H., 2002. Workload based order release and productivity: a missing link. *Production Planning and Control*, 13(7), p. 665–678.

Bertrand, J. & Wortmann, J., 1981. *Production Control and Information Systems for Component-Manufacturing Shops*. Elsevier.

Bobrowski P. M., Park P. S., 1993. An evaluation of labor assignment rules when workers are not perfectly interchangeable. *Journal of Operations Management*, 257-268

J. A. C. Bokhorst , J. Slomp & G. J. C. Gaalman (2004) On the who-rule in Dual Resource Constrained (DRC) manufacturing systems, *International Journal of Production Research*, 42:23, 5049-5074,

J.A.C. Bokhorst & G.J.C. Gaalman (2009) Cross-training workers in Dual Resource Constrained systems with heterogeneous processing times, *International Journal of Production Research*, 47:22, 6333-6356,

Breithaupt, J., Land, M. & Nyhuis, P., 2002. The workload control concept: theory and practical extensions of load oriented order release. *Production Planning and Control*, 13(7), p. 625–638.

Brusco M. J., Johns T. R., 1998. Staffing a Multiskilled Workforce with Varying Levels of Productivity: An Analysis of Cross-training Policies, *Decision science*, 29, 2

Cigolini, R. & Portioli-Staudacher, A., 2002. An experimental investigation on workload limiting methods with ORR policies in a job shop environment. *Prod. Plan. Control*, 13(7), p. 602–613.

J. T. Felan, Philipoom P.R., 1993. Labour flexibility and staffing levels in a dual-resource constrained job shop

J. T. Felan & Timothy D. Fry (2001) Multi-level heterogeneous worker flexibility in a Dual Resource Constrained (DRC) job-shop, *International Journal of Production Research*, 39:14, 3041-3059,

Fry, T. & Smith, A., 1987. A procedure for implementing input/output control: A case study. *Production and Inventory Management Journal*, 28(4), p. 50-52.

Fry T. D., Kher H. V., Malhotra M. K., 1995. Managing worker flexibility and attrition in dual resource constrained job shops. *International Journal of Production*, vol 33

Germes, R. & Riezebos, J., 2010. Workload balancing capability of pull systems in MTO production. *International Journal of Production Research*, 48(8), p. 2345–2360.

Glasse, C. R. & Resende, M. G., 1988. Closed-loop job release control for VLSI circuit manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, Volume 1, p. 36-46.

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

Hendry, L. & Kingsman, B., 1989. Production planning systems and their applicability to make to order companies. *European Journal of Operational Research*, Volume 40, p. 1-15.

Hendry, L., Land, M., Stevenson, M. & Gaalman, G., 2008. Investigating implementation issues for workload control (WLC): A comparative case study analysis. *International Journal of Production Economics*, Volume 112, p. 452–469.

Hendry, L. & Wong, S., 1994. Alternative order release mechanisms: a comparison by simulation. *International Journal of Production Research*, 32(12), p. 2827–2842.

Henrich, P., 2005. Applicability aspects of workload control in job shop production. Ph.D. Thesis, University of Groningen, The Netherlands, Labyrint Publications, Ridderkerk, The Netherlands, ISBN:9053350470.

Henrich, P., Land, M. & Gaalman, G., 2004. Exploring applicability of the workload control concept. *International Journal of Production Economics*, Volume 90, p. 187-98.

Henrich, P., Land, M. & Gaalman, G., 2005. Grouping machines for effective workload control. *International journal of production economy*.

M.S. Jayamohan, C. Rajendran 2000. New dispatching rules for shop scheduling: a step forward. *International journal of production*.

Kher H.V., Malhotra M.K., 1994. Acquiring and operationalizing worker flexibility in dual resource constrained job shop with worker transfer delays and learning losses

H. V. Kher, M. K. Malhotra, P. R. Philipoom, T. D. Fry, 1999. Modeling simultaneous worker learning and forgetting in dual resource constrained systems. *European Journal of Operational Research*

Kher H.V., 2000. Examination of flexibility acquisition policies in dual resource constrained job shops with simultaneous worker learning and forgetting effects. *Journal of the operational research society*

H. V. Kher & T. D. Fry (2001) Labour flexibility and assignment policies in a job shop having incommensurable objectives, *International Journal of Production Research*, 39:11, 2295-2311

Malhotra M.K., Fry T. D., Kher H. V., 1993. The Impact of Learning and Labor Attrition on Worker Flexibility in Dual Resource Constrained Job Shops

Kingsman, B., 2000. Modeling input–output workload control for dynamic capacity

planning in production planning systems. *International Journal of Production Economics*, 68(1), p. 73–93.

B. G. Kingsman & Linda Hendry (2002). The relative contributions of input and output controls on the performance of a workload control system in Make-To-Order companies

Kingsman, B., Hendry, L., Mercer, A. & De Souza, A., 1996. Responding to customer enquiries in make-to-order companies: problems and solutions. *International Journal of Production Economics*, Volume 46-47, p. 219-31.

Kingsman, B. & Mercer, A., 1997. Strike rate matrices for integrating marketing and production during the tendering process in make-to-order subcontractors. *International Transactions in Operational Research*, 4(1), p. 251-7.

Land, M., 2006. Parameters and sensitivity in workload control. *International Journal of Production Economics*, 104(2), p. 625–638.

Land, M. & Gaalman, G., 1996. Workload control concepts in job shops: a critical assessment. *International Journal of Production Economics*, p. 535–548.

Land, M.J., Stevenson, M., Thürer, M., Gaalman, G.J.C., 2015. Jobshop control: in search of the key to delivery improvements. *Int.J.Prod.Econ.* 168, 257–266.

Lu, H., Huang, G. & Yang, H., 2010. Integrating order review/release dispatching rules for assembly job shop scheduling using a simulation approach. *International Journal of Production Research*.

MacCarthy, B., 2006. Organisational, systems and human issues in production planning, scheduling and control. In: *Handbook of Production Scheduling*, International Series in Operations Research and Management Science, New York: s.n., p. 59-90.

Melnyk, S. & Ragatz, G., 1989. Order review/release systems: research issues and perspectives. *International Journal of Production Research*, p. 1081-1096.

M. R. Moreira, R. Alves, 2006A new input-output control order release mechanism: how workload control improves manufacturing operations in a job shop

Mosca, R., Giribone, P. & Guglielmo, G., 1982. Optimal length in O.R. simulation experiment of large scale production system.. Proceedings of IASTED International Symposium on Applied Modeling and Simulation, p. 78–81

Oosterman, B., Land, M. & Gaalman, G., 2000. Influence of shop characteristics on workload control. International Journal of Production Economics, 68(1), p. 107–119.

Park, C., Song, J., Kim, J. & Kim, I., 1999. Delivery date decision support system for the large scale make to order manufacturing companies: a Korean electric motor company case. Production Planning & Control, 10(6), p. 585-9.

Park P. S., Bobrowski P.M., 1989. Job R&me and Labor flexibility in a Dual Resource Constrained Job Shop. Journal of operation management Vol8, N 3

Park P.S., 1991. The examination of worker cross-training in a dual resource constrained job shop. European Journal of operation research 51.

Perona, M. & Portioli, A., 1996. An enhanced loading model for the probabilistic workload control under workload unbalancement. Production Planning and Control, 7(1), p. 68-78.

Perona, M. & Portioli, A., 1998. The impact of parameters setting in load oriented manufacturing control. International Journal of Production Economics, Volume 55, p. 133–142.

Philipoom, P., Malhotra, M. & Jensen, J., 1993. An evaluation of capacity sensitive order review and release procedures in job shops. Decision Sciences, Volume 24, p. 1109-1133.

Portioli, A., 1991. Proposal and evaluation of new load oriented algorithms for shortterm production planning in a job shop environment (in Italian). Milano: PhD Thesis,

Politecnico di Milano, Italy.

Portioli-Staudacher, A. & Tantardini, M., 2012. A lean-based ORR system for nonrepetitive manufacturing. *International Journal of Production Research*, 50(12), p. 3257-3273.

Ragatz, G. & Mabert, V., 1988. An evaluation of order release mechanisms in a job shop environment. *Decision Sciences*, Volume 19, p. 167-189.

Riezebos, J., 2010. Design of POLCA material control systems. *International Journal of Production Research*, 48(5), p. 1455-1477.

Sabuncuoglu, I. & Karapinar, H., 1999. Analysis of order review/release problems in production systems. *International Journal of Production Economics*, 62(3), p. 259–279.

Sabuncuoglu, I. & Karapinar, H., 2000. A load-based and due date- oriented approach to order review/release in job shops. *Decision Sciences*, 31(2), p. 413–447.

M. Sammarco, F. Fruggiero, W.P. Neumann & A. Lambiase (2014) Agent-based modelling of movement rules in DRC systems for volume flexibility: human factors and technical performance, *International Journal of Production Research*, 52:3, 633-650,

Silva, C., Stevenson, M. & Thürer, M., 2015. A case study of the successful implementation of workload control A practitioner-led approach. *Journal of Manufacturing Technology Management*, Volume 26, p. 280-296.

Stevenson, M. & Hendry, L. C., 2006. Aggregate load-oriented workload control: A review and a re-classification of a key approach. *International Journal of Production Economics*, 104(2), p. 676-693.

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

M. Stevenson, L. C. Hendry, 2006, Aggregate load-oriented workload control: A review and a re-classification of a key approach

Tatsiopoulou, I., 1997. An order release reference model as a link between production management and shop floor control software. *Computers in Industry*, Volume 33, p. 335–344.

M. Thürer, M. Stevenson, M. J. Land, Fredendall, 2012. Workload control and order release. *Production and Operations Management* 0(0), pp. 1–15, © 2012 Production and Operations Management Society

M. Thürer, M. Stevenson, M. J. Land, Fredendall, Melnyk 2014. Lean control for make to order companies: integrating customer enquiry management and order release

Thürer, M. et al., 2015. Concerning Workload Control and Order Release: The Pre-Shop Pool Sequencing Decision. *Production and Operations Management*, 24(7), p. 1179-1192.

M. Thürer, M. Stevenson, M. J. Land, 2016, On the integration of input and output control: Workload Control order release

Mark Treleven (1989) A Review of the Dual Resource Constrained System Research, *IIE Transactions*, 21:3, 279-287,

Wein, L., 1988. Scheduling semiconductor wafer fabrication. *IEEE Transaction on Semiconductor Manufacturing*, 1(3), p. 115–130.

Welch, P., 1983. The statistical analysis of simulation results.. *The Computer Performance Modeling Handbook*. Academic Press, p. 268–328.

Wiendahl, H., 1995. Load Oriented Manufacturing Control. Berlin, Springer.

Witte, P., Kirchhof, N. & Meseth, T., 2008. Simulation based evaluation of the Workload

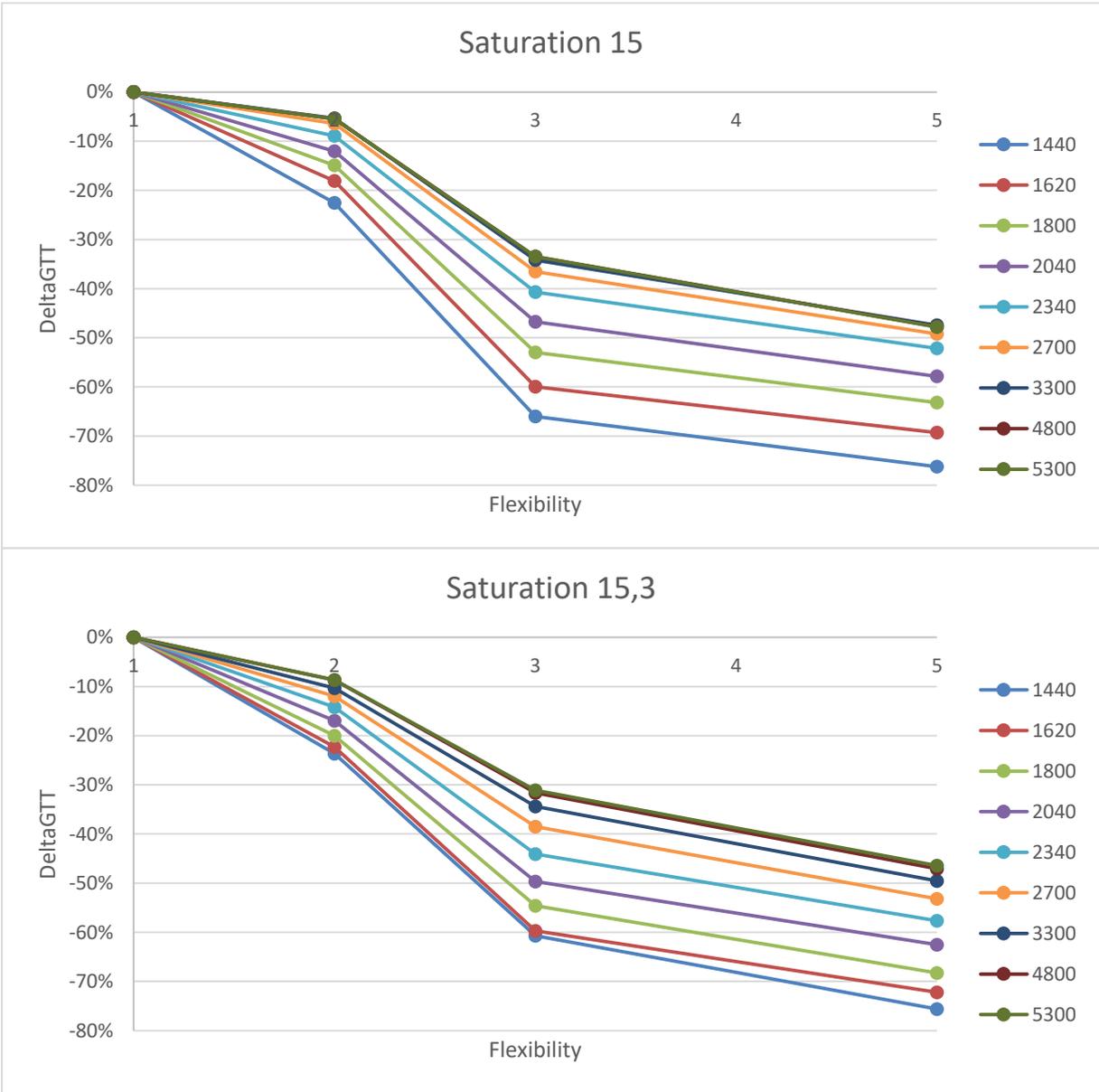
Control concept for a company of the automobile industry. Osnabruck, WSC 2008, p. 1856-1862.

Yuan Huang (2017) Information architecture for effective Workload Control: an insight from a successful implementation, *Production Planning & Control*, 28:5, 351-366

8. APPENDIX

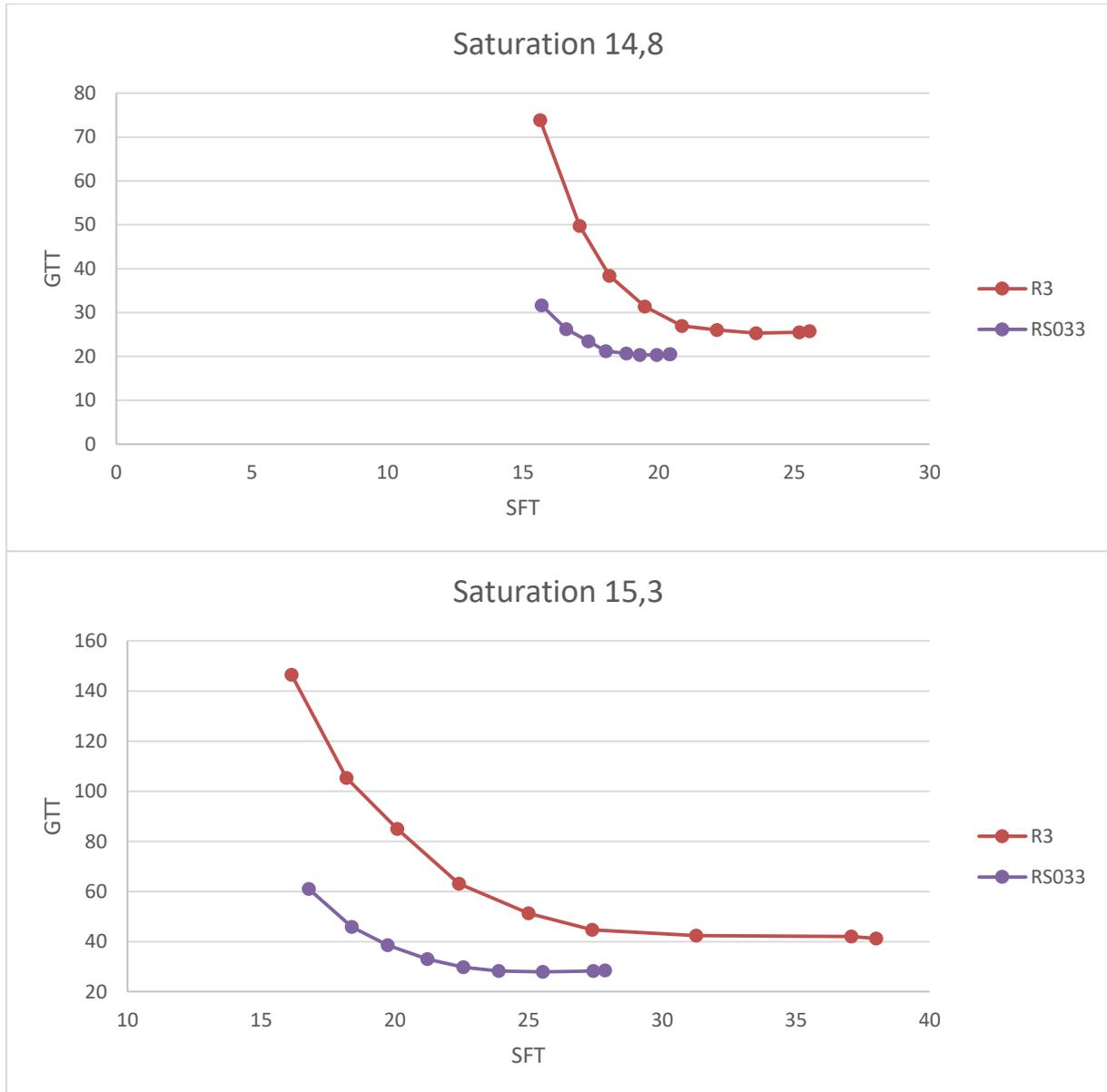
APPENDIX A

Percentage GTT variation for every workload norm due to increasing level of worker flexibility respect to R0



APPENDIX B

Centralized and decentralized rules comparison of GTT-SFT in high variability context (var 1936). Low saturation.



APPENDIX C

Centralized and decentralized rules comparison of tardiness and tardy orders with high saturation in high variability contexts.

