PERFORMANCE MANAGEMENT

IN MANUFACTURING

A method for short term

performance evaluation and diagnosis

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van
de Rector Magnificus, prof.dr. J.H. van Lint,
voor een commissie aangewezen door het College
van Dekanen in het openbaar te verdedigen op
dinsdag 12 maart 1996 om 16.00 uur

door

Paulus Petrus Marinus Stoop

geboren te Breda
Thesis Technische Universiteit Eindhoven. - With ref. - With summary in Dutch.  
ISBN 90-386-0175-1  
NUGI 689  
Subject headings: performance management / production control.

© 1996, P.P.M. Stoop, Veldhoven
PERFORMANCE MANAGEMENT

IN MANUFACTURING

A method for short term

performance evaluation and diagnosis
ACKNOWLEDGMENTS

This thesis is the final product of a research project sponsored by the Netherlands Organization for Scientific Research (NWO) and conducted in the department of Operations Planning and Control of the Faculty of Technology Management of Eindhoven University of Technology. Many people have provided support in various stages of the project. I would like to thank them all and I realize that the following list is incomplete.

I acknowledge Will Bertrand for his valuable input in the project. The many in-depth discussions have improved the level of this thesis without any doubt and his quick feedback on the various deliverables during the project kept the project going. Corné Dirne proved to be a constructive discussion partner, and his help with the development of the simulation model has been very valuable. I thank Jacques Theeuwes and Henk van der Veeken for their useful comments in the regular steering group meetings. Peter Sander and Ton de Kok are acknowledged for their contributions to chapters 4 and 5. The members of the department of Operations Planning and Control provided an amicable and stimulating atmosphere. I enjoyed the great soccer and travel stories of my room mate Harrie de Haas. Jannet van Zante, Jos Verrijdt and Gert Regterschot are acknowledged for their support with mathematical issues.

The building of the software tool for the two production departments of the aircraft manufacturer has been done by Koos Huibers, who did his job very accurately. The aircraft manufacturer enabled me to use and implement the theoretically developed insights in a practical environment for more than two years. I especially appreciate the pleasant cooperation with Hans Gies Broesterhuizen. Marc de Leeuw, Marjan Sprenkels and Kasper Wilderboer I acknowledge for the support during the project in practice. The discussions with the shift leaders (Lemi Akin, Rob van Doorn, Willem Vlaming and Wout Bosman) and the schedulers (Peter Velleman and Erik Neerings) were very valuable, and they also gave many suggestions for improving the software instrument.

Finally, I wish to thank my family and friends for their interest and support. Most important I wish to thank Marion. Last but not least I would like to thank Denise who from the moment of her birth already seemed to understand that I should enjoy a good night's rest during the last months of the completion of the research.

Paul Stoop
Veldhoven, December 1995
# Table of Contents

Chapter 1  Introduction  
1.1 Motivation for the research  
1.2 Problem description and research objective  
1.3 Outline  

Chapter 2  Basic concepts  
2.1 Production unit control and goods flow control  
2.2 Production units  
2.3 Decision making in production units  
2.4 Performance management in production units  

Chapter 3  Performance measurement and evaluation  
3.1 Performance measurement and performance measures  
3.2 Performance measurement systems  
3.3 Performance evaluation and performance targets  
3.4 Performance models  
3.5 Frequency of performance measurement and evaluation  
3.6 Summary and preview  

Chapter 4  Prediction model  
4.1 Model variables and performance measures  
4.2 Throughput prediction rule  
4.3 Production order completion time prediction rule  
4.4 Estimation errors of model variables and model extensions  

Chapter 5  Simulation study  
5.1 Requirements  
5.2 Experimental design  
5.3 Prediction rules  
5.4 Simulation results  
5.5 Discussion of results
CHAPTER 1

INTRODUCTION

1.1 Motivation for the research

In many manufacturing organizations performance measurement has become a standard procedure nowadays. In the past most attention was directed at financial measurements. It was generally believed that organizations could only survive when their financial results (such as return on investments or net profits) were good, or at least better than the average in the line of business. Consequently, costs and efficiency were the factors that gained the most concern from the top management. Management accounting systems were developed that supported this financial focus and are still in use today (e.g. Kaplan, 1983, Kaplan, 1990).

This financial focus on the performance of an organization remained important, but got less attention as a result of two developments:

- The rise of the Japanese manufacturing technologies and philosophies.
- The change from a sellers market to a buyers market.

First, the Japanese set a trend to manufacture products with a high quality. This focus on quality not only saved manufacturing costs but it appeared that the quality aspect became an order winning aspect too. Waste reduction, zero defects, and total quality control became issues organizations could not neglect anymore to remain competitive. Besides the increasing focus towards quality, also timeliness became an important success factor for companies. Not only the reduction of order flow times was regarded as important in order to establish fast delivery times, but also delivery reliability became an important issue. The Just-In-Time philosophy arose and has been embedded in many organizations today. An overview of these techniques and philosophies is given by Schonberger (1982).

The second reason for the decreasing emphasis on costs was that the market started to change from a sellers market to a buyers market. The customer's wishes gradually got the upper hand over the manufacturer's strategies to produce as little as possible product variants in the most efficient way. Innovation, effectiveness and flexibility became the new buzz words. New products were required to meet the customer's demands and these products had to be
developed and put on the market fastly to gain a market share as large as possible. The effectiveness of the manufacturing process became more important than the efficiency. On top of that, more flexibility was required to follow the changes in demand with respect to the volume of the products, i.e. volume flexibility, as well as the variety in the types of products, i.e. mix flexibility (e.g. Bolwijn and Kumpe, 1990).

Clearly, both developments have financial consequences for an organization. The result of these two developments therefore was that organizations had to focus on costs, quality and timeliness simultaneously rather than costs only in order to be or to remain competitive. For that reason, management had also to develop new performance measures that were non-financial by nature. Examples of these performance measures are: product quality, delivery time, delivery reliability, and product flexibility. Actually, these kinds of performance measures were not new. The use of many of these performance measures were common already at the operational level for monitoring and controlling the goods flows. At this operational level, the use of financial measures has always been very limited, because these measures were not "recognized" as relevant; physical measures have always played the first fiddle because these kinds of measures were easy to use for planning and control purposes.

Despite the fact that operations management already used the "new" performance measures, the changes in the world indeed had an impact on the operational level. Because the well-known measures at the operational level were now regarded as important by higher management levels too, the operational level became the source for performance improvements in terms of quality and time. While performance measurements still were executed on the total organizational level to judge how well the organization was doing compared to the competitors (for example by competition analysis), at the same time performance measurements were required for internal use because the performance of individual plants, departments or functions in an organization became subject for improvement.

At the operational level, the manufacturing system usually is split up into some sub-systems. In each sub-system, a well-defined part of the total manufacturing process takes place. For the moment we will call such a sub-system a production department. A more detailed definition will be given in Chapter 2. The tuning of the different processes in the different production departments is done by a goods flow control function in the organization, whereas each production department should plan and control the performance of its own physical transformation activities. Often, at the production department level we can observe a short term focus of the production department management with respect to the planning and control
of the activities. There are several reasons for this short term focus. First, an inherent characteristic of operational processes is that it always requires ad hoc decisions to overcome all kinds of disturbances like machine-breakdowns, tool unavailability, and product quality problems. Second, the measurement reports at this level appear frequently (e.g. each shift, daily or weekly) which lead to a decision behavior that is adapted to this reporting frequency. An advantage of this frequent measurement of the results over short time periods is that it leads to a fast feedback of the realized performance. According to feedback theory, the shorter the feedback loops, the more effective it is to detect, to foresee and to overcome any shortcomings in the performance in time (e.g. Algera, 1990). A good example of the adaptation of human behavior to the measurement reporting frequency is the so-called hockey-stick phenomenon (e.g. Chase and Aquilano, 1989), where people make an extra effort at the end of a measurement period to meet specific performance targets, like the throughput. A last reason why the focus at production department level is usually directed at the short term is that due to the customer demands on shorter delivery times quicker responses are required. For the production department level this means that more flexibility is needed with respect to, for example, capacity (re)allocations, the use of alternative routings and machine setups. The increase in the required flexibility will influence the decision behavior that will be more focused on the short term.

To relate the subject of this research to the current trends towards shorter delivery times, higher productivity and higher delivery reliability, we will restrict the performance of production departments to the performance with respect to the throughput and timeliness aspect. The two other important aspects of the total performance, i.e. costs and quality, will get less attention in this research; only when evident relationships with throughput and timeliness exist, these aspects will be taken into account. For example, if an important part of the produced products is rejected, this will not be considered as a quality issue only, but it also will be related to the impact on the throughput and the risk of a bad delivery reliability.

Performance measurement at the production department level is needed to monitor the actual performance or trends in the performance in order to foresee unwanted situations or to react on undesirable situations. The results of the performance measurements however, only give insight where the actual performance is worse than expected; they do not give insight in why the actual performance differs from the expected performance nor do they tell how one can improve the actual performance. An answer to the question why there exists a performance deviation should give enough clues for finding ways to improve it. In practice, often qualitative explanations are given that might explain the performance deviations. Although

*Introduction*
these explanations may sound very reasonable, they usually are very subjective and based on intuitive feelings. Seldom an objective quantification of the impact of each of the causes on the performance can be given. Besides the subjective explanations, often also the determination of the expected performance is subjective by nature. The expected performance or performance target not always is considered to be realistic and it does not reflect the maximum achievable performance. The lack of an appropriate method that can determine objective performance targets for performance evaluation purposes and that can quantitatively explain the achieved performance forms the main motive for this study.

1.2 Problem description and research objective

Despite its main drawback of not giving insight how the performance can be improved, manufacturing companies still evaluate the performance on costs, quality and timeliness by comparing their own performance with the performance on these aspects achieved by competitors in the same branch of trade. However, more and more organizations begin to recognize that such a way of competition analysis can not simply be transposed to lower organizational levels such as a production department. For instance, comparing the productivity of two production departments of different organizations is a difficult task, because all kinds of factors, which are often difficult to quantify, play a role in the resulting performance at the end of a performance reporting period. Examples of these factors are product complexity, task complexity, educational level of the production personnel, and motivation. Also, organizational choices about for example the lay-out of a production department and the degree of automated production play an important role. So, because of the many possible differences between two or more production departments, comparing their performances on different aspects is impossible, and consequently unfair. Therefore, it should be clear that each production department actually needs its own performance targets to compare with the actual performance.

Although the performance measurements give insight in the actual performance that is achieved, in practice often two questions arise:
• How good is the actual performance of the production department?
• Where and how can the performance be improved?

The first question addresses the issue that one doesn't know what the maximum performance is that can be attained. Usually, there are periods in which the actual performance on a specific aspect is substantially above the performance target, and there are periods in which
the actual performance will be significantly lower than the performance target. In this way, the performance target represents a sort of average performance that should be realized over a relatively long period of time instead of a more realistic value that gives insight in the maximum performance level that can be realized within a certain shorter period of time. The danger of using a performance target that not corresponds with the level that can be realized is that in situations where the performance target already has been achieved one may slow down the work activities, whereas in situations where one is far away from the performance target, one may not work at maximum speed because the target is regarded as unattainable whatever the effort is. Further, if it is assumed that the performance target should be equal to the average performance of the past, then there is also the danger that the current levels of waste are built in.

The second question where and how the performance can be improved in fact is derived from the first question. This question relates to the problem that one does not know which factors cause the deviations between the actual performance and the performance targets or, in other words, which factors should be handled in order to realize the maximum possible performance. Usually, one is able to name some factors that play an important role in the realization of the performance, but due to the many complex relationships between the considered factors, it is not always clear if a named factor really is the factor that causes the performance deviation. In most situations only qualitative remarks about factors influencing the production department's performance can be made. For example, a machine breakdown for several days clearly will have a negative influence on the throughput at the workcenter the machine belongs to. However, the influence of this breakdown on the throughput level at other machines or workcenters is more difficult to determine. Even more difficult will be the determination of the effect of this machine breakdown on the delivery reliability of the production department.

The above questions address in fact the issue of continuous improvement, where all kinds of waste in and around the manufacturing processes are considered to be a source of improvement. From the two questions it will be clear that performance measurement alone does not suffice to have an objective view on the achieved performance and to find ways for performance improvements. Performance measurement will just be the starting point of the ongoing process directed at continuous improvement. Next, related to the second question, an objective performance evaluation should take place. The objectivity of the performance evaluation can be realized by having objective performance targets. Finally, a diagnosis is needed to explain deviations between the actual performance and the performance target. A method by which performance measurement, performance evaluation, and performance

Introduction
diagnosis are related, will facilitate the finding of sources for performance improvements. This method by which performance measurement, performance evaluation and performance diagnosis is used to find ways to improve the performance we will call the process of performance management.

The goal of the research is to develop a method that can be used to evaluate and diagnose the performance of complex production departments in the short term. The results of the performance evaluation and diagnosis should be expressed as much as possible in a quantitative way. This will help operations management to get a more detailed and quantitative rather than qualitative insight in the processes that take place at the production department level. The method should give insight in the maximum performance that can be achieved in a certain period of time given the specific characteristics of the production department. Instead of using performance targets coming from "everywhere", internal performance targets that are only directed at the production department considered are needed to evaluate a production department's performance. The relationship of the production department's performance on the performance of other production departments in the chain of the manufacturing process will also be taken into account. The goods flow control function will thereby play an important role, because this function is responsible for the tuning of the inputs and outputs between different production departments.

1.3 Outline

The objective of this research is to develop an instrument for the purpose of performance evaluation and diagnosis in the short term at the production department level of manufacturing organizations. In Chapter 2 some basic concepts will be introduced that form the starting points for this research. This will include a more detailed definition of the process of performance management, which will largely determine the outline of this thesis. Because each of the phases in the performance management process requires its own approach, these phases will be dealt with separately in the chapters of this thesis. This can be seen in Figure 1.1 where an overview is given of the main subjects of this research.

In Chapter 3 we will discuss the subjects of performance measurement and performance evaluation. First, a literature overview will be given about performance measurement and performance measures. Second, it will be discussed how the different performance measures can be related to each other in a coherent performance measurement system. Third, a definition will be given of performance evaluation. Also an overview will be given of how
Figure 1.1 Relationship between the main subjects of this thesis.

performance targets can be set. Fourth, a short overview will be given of performance evaluation models, and it will be made clear why these models are inappropriate for generating performance targets to be used for the short term. The last section in Chapter 3 deals with some issues related to the frequency of the measurement and evaluation of the performance. Because existing performance models are inappropriate to generate realistic performance targets for production departments in the short term, we developed a new model. The development of this model will be discussed in Chapter 4. The main characteristic of this model is that it is state-dependent which means that relevant information about the actual state of the production department is used. The quality of the state-dependent prediction rules that are part of the state-dependent model is tested by simulation experiments in Chapter 5. As measure for the quality of the predictions, the standard deviation of the prediction error will be used. For a number of different situations (characterized by the number of workcenters in a production department and the average utilization rate) the quality of the state-dependent predictions will be compared with some state-independent predictions. Statistical tests are used to evaluate whether the state-dependent prediction rules outperform the state-independent prediction rules. Then, in Chapter 6, the diagnosis of the performance will be discussed. After a literature review about performance diagnosis, we will put forward our definition of this term. Because the performance measurement and performance evaluation are quantitative by
nature, also the performance diagnosis will be restricted to the quantitative part of the performance analyses. Then, a framework will be given that shows how performance measurement, performance evaluation and performance diagnosis are related to each other and how the state-dependent prediction model should be used in this framework. The extent to which the achieved performance can be diagnosed with our method will be discussed in the last section of this chapter. In Chapter 7 we describe an example of the implementation and use of the performance evaluation and performance diagnosis method in two machining production departments. In this chapter most attention will be paid to the way the people involved actually use the method and the results that were obtained by applying the method. Based on the results of the simulation experiments and the empirical results, a more general discussion will be held in Chapter 8 about the trade-off that has to be made between the required efforts and expected benefits related to the implementation and use of the performance evaluation and diagnosis method in a practical situation. The chapter will end with the naming of some conditions and guidelines to be fulfilled to make the implementation of the method as smoothly as possible. Finally, conclusions and recommendations for further research are presented in Chapter 9.
CHAPTER 2

BASIC CONCEPTS

The aim of this chapter is to define the starting points for this research. In Section 2.1 we will define production unit control and its relationships with goods flow control. The decisions that are part of the production unit control will also be discussed in this section. In Section 2.2 we will describe which elements can be distinguished within a production unit and how they interact. In Section 2.3, the starting points for decision making at the production unit level will be described. Finally, we will discuss which steps should be taken to manage the performance of production units.

2.1 Production unit control and goods flow control

In the introduction (Chapter 1) we split up a production system into production departments which are responsible for a part of the production process. In the remainder of this thesis, a more restricted definition of the term production department will be used. Therefore, we introduce the concept of production unit. A production unit is defined as "a production department which on short term is self-contained with respect to the use of its resources, and which is responsible for the production of a specific set of products (the production unit end-items) from a specific set of materials and components (the production unit start-items)" (Bertrand et al., 1990a). A production unit can range from a (large) machine to a large production department and thus is not necessarily equal to an organizational production department. In many production organizations the total production control is split up into production unit control and goods flow control. Production unit control refers to the extent to which a production unit can control its production with respect to objectives like quality and timeliness. To coordinate the inputs and outputs of the various production units in a production system goods flow control is used (see Figure 2.1). Goods flow control coordinates the production levels of the production units and the materials supply to the various decoupling points (i.e. inventory points to secure uncertainty with respect to time and volume of supply and demand or to decouple the manufacturing of subassemblies to a plan and the assembly of sales items to customer demand) between the manufacturing phases. The release of orders to the production units is thereby the coordination mechanism. The ultimate
objective of goods flow control is to deliver the right products in the right quality on the right time at the right place and at minimum costs, which we will refer to as the overall delivery performance to the final customer. In this research we focus on production unit control and its interactions with goods flow control.

Figure 2.1 Relationship between production units and goods flow control.

With respect to production unit control, 6 decision functions are of importance (see Figure 2.2, taken from Bertrand et al., 1990b). At the plant level parameters are set with regard to the total production control, such as lot sizes, order flow times and delivery reliability (e.g. Durlinger, 1985 and 1986, Karmarkar, 1987). Here trade-offs are being made for a longer time horizon in such a way that the organizational goals will be met. At the next level, the production system level (a production system comprises of a number of production units), goods flow control takes care of the capacity planning to level out demand and capacity fluctuations. Also decisions regarding order acceptance and delivery date setting are taken at this level, taking the capacity plans as a starting point. The next decision function, the order release function, is a decision function that forms the interface between the production system level and the production unit level. At the release moment often there is a possibility to change some of the parameters for individual production orders (such as a lot size) based on specific information from the production unit to establish smoother goods flows within the production unit. When the orders are released, the production units are responsible for the progress within the unit. To control the order progress, production units take three types of
decisions: detailed order scheduling, capacity allocation and order sequencing. These decisions will be discussed in more detail in the next sections of this chapter.

<table>
<thead>
<tr>
<th>Plant level</th>
<th>Parameter setting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Capacity planning</td>
</tr>
<tr>
<td></td>
<td>Utilization planning</td>
</tr>
<tr>
<td>Production system level</td>
<td>Order acceptance</td>
</tr>
<tr>
<td></td>
<td>Due date setting</td>
</tr>
<tr>
<td></td>
<td>Production order release</td>
</tr>
<tr>
<td>Production unit level</td>
<td>Detailed production order scheduling</td>
</tr>
<tr>
<td></td>
<td>Capacity allocation</td>
</tr>
<tr>
<td></td>
<td>Production order sequencing</td>
</tr>
</tbody>
</table>

**Figure 2.2** The decision functions related to production unit control.

### 2.2 Production units

In this research we will consider complex, job-shop like production units. The physical elements of such a production unit are the so-called workcenters. Each workcenter consists of one or more machines that have the same functionality. Generally, the physical layout is set in such a way that the main stream of the production orders flow from workcenter to workcenter with a minimum of distance for internal transportation. Decisions regarding the organization of production units will not be taken into account in this research; the production unit's organization is assumed to be given.
In a production unit several capacity resources can be distinguished:
- operators;
- machines;
- tools.

Usually these capacity resources are scarce (i.e. they are attuned to the average demand and not to the maximum demand within a certain period of time) in order to minimize costs given the objective of the overall delivery performance as stated by goods flow control. To increase the production unit's flexibility with respect to the capacity, the following measures can be taken:

- Operator-capacity measures such as creating operator-flexibility so that they can work at two or more workcenters (e.g. Fryer, 1974), interchanging capacity between production units, working overtime, and temporarily use of operators; an overview of operator-capacity measures is given by Faißt (1992).
- Creating machine-flexibility by defining alternative production order routings.
- Subcontracting of work.

All these measures influence the utilization of the production unit, defined as the ratio of the utilized capacity and the available capacity. In this thesis we assume the operator and machine flexibility as given. The determination of the number of tools and the allocation of tools to machines will not be considered in this research.

A production order consists of one or more products which have identical characteristics. The number of products is the lot size of the production order. A production order can be created from a set of customer orders. Production efficiency considerations may lead to the creation of these sets. The complexity of production unit control in job-shop-like production units is caused mainly by the typical characteristics of production orders entering the production units. First, there is a large variety in routings of production orders. This variety refers to both the number of workcenters in the routing and to the sequence of workcenters in the routing. Second, there is a large variety in processing times at workcenters.

The behavior of a production unit as part of the goods flow can be described by the so called operational characteristics (Bertrand et al., 1990a). These characteristics are used by goods flow control to coordinate the goods flow between production units. Examples of operational characteristics are production order flow times and capacity restrictions. The detailed events that occur within the production units are not considered by goods flow control. The problems that arise within the production units should be handled by the production units themselves. It is assumed that there is enough flexibility in the production units to overcome these problems.
in the progress of production orders if the release function of production orders takes into account the operational characteristics.

2.3 Decision making in production units

Generally, the process of decision making can be subdivided into three stages (Simon, 1960):

- Searching the environment for conditions calling for a decision - the intelligence activity.
- Inventing, developing, and analyzing possible courses of action - the design activity.
- Selecting a particular course of action from those available - the choice activity.

Simon models the decision maker as a so-called "administrative man", someone who looks for a course of action that is satisfactory or "good enough". In this way relatively simple rules of thumb can be used instead of searching for all the possible alternatives, which makes the decision maker's world much simpler. The decisions themselves can be programmed or non-programmed. These are not mutually exclusive but rather make up a continuum stretching from highly programmed decisions at the one end to highly unprogrammed decisions at the other. Highly programmed decisions have a repetitive and routine character and can be found mainly at the lowest organizational level. Decisions are unprogrammed when they are new and unstructured and often are made at the top organizational level.

In our opinion, decision making in production units is located between the two extreme types of decisions as defined by Simon. In Section 2.1 we already defined the three types of decisions that take place at the production unit level: detailed production order scheduling, capacity allocation, and production order sequencing. These decisions will be discussed more extensively in the remainder of this section.

In complex production systems, usually goods flow control already has made up a schedule by generating due dates based on standard flow times per production unit or production system. Sometimes also operation due dates for the production orders are generated. Within individual production units more detailed scheduling may be required, especially if no operation due dates have been set. It may be the case that certain operations can be done only at specific moments in time or in a specific sequence. This detailed schedule thus serves the goals of the production unit control, for example by trying to maximize the efficiency within the production unit. Here, specific and actual production unit information is used to optimize the production schedules. Usually this detailed information is not available for goods flow control. If the objectives of the schedules and all the rules to compose a schedule of the
production orders available are defined, the detailed scheduling decision can be regarded as a programmed decision. If there are many exceptions from the scheduling rules which are difficult to formalize, more ad hoc decisions are required which gives this decision function a more unprogrammed character.

The capacity allocation decision is a decision that usually is taken on a daily basis. However, due to all kinds of disturbances that may take place within the production unit (such as machine breakdowns and the release of rush orders) it is possible that reallocation decisions have to be made to ascertain a smooth production order flow. Having defined the criteria for making the first allocation decision, this decision can be considered as a programmed decision. Usually, reallocation decisions are ad hoc decisions based on the specific disturbances that occur and the judgment about the impact of the disturbances on the production order flows. The faster the reallocation can be made, the less influence a specific disturbance will have on the performance. The effects of the speed in which a reallocation decision is made on the performance with regard to production order flow times and delivery reliability of the production unit, is described by Fryer (1974). Because a reallocation decision is based on the specific cause for the decision and on the subjective judgement of the expected impact, the reallocation decision can be considered as an unprogrammed decision.

The sequencing decision is the decision to select a production order out of a particular queue at a workcenter. In many production units of the type we take as a starting point some kind of a priority rule is used. This priority rule can be based on a particular characteristic of the operation (e.g. SPT, Shortest Processing Time), a planning characteristic of the production order (e.g. ODD, Operation Due Date), or the sequence of arrival of the production orders (e.g. FCFS, First Come First Served). An overview of priority rules and their characteristics is given by Baker (1975), and more recently, Ramasesh (1990). Ideally, this type of decision can be regarded as a standard or routine decision. However, in practice the priority rules will hardly be followed exactly, mainly due to efficiency considerations made by the persons that select the production orders. This observation shifts the sequence decision more into the direction of an unprogrammed decision.

In many situations, a specific decision may have a positive influence on the performance of some aspect, but at the same time a negative influence on the performance of another aspect. For example, the decrease of lot-sizes may result in shorter production order flow times, but it also results in higher setup costs. Another example is that the decision to work overtime to achieve a better delivery reliability. Besides possible social conflicts that may arise, this decision results in extra costs. These examples show that the decision making process in the
production unit is very complex. Generally, one can make a decision to optimize a certain goal, thereby neglecting possible negative influences on other goals, or one can try to make a compromise to partly satisfy several goals. This also indicates that the decision making process at the production unit level is more concentrated on the unprogrammed side of the continuum of types of decisions than the programmed side.

In order to make satisfying decisions, a decision maker requires measurements. The measurements give decision makers insight in what the results are on certain performance measures of the combined effects of the near-past occurrences and the related decisions. According to Mason and Swanson (1981) decision makers seldom suffer from the lack of measurement data; usually substantial data are available and if additional data are needed they may frequently be obtained through research and analysis. The problem for decision makers usually is:

- to secure the data that are most relevant to the current decision;
- to grasp more fully the meaning of the data at hand;
- to derive the appropriate implications from the available data to relate them to alternative courses of actions.

At the production unit level often all kinds of standardized measurements are available. The frequency of these measurements is relatively high. This is due to the characteristics of the decision making processes at this level. The three types of decisions at the production unit level are often called operational decisions. Important characteristics of these operational decisions are that the time horizon is relatively short and that they are relatively detailed compared with strategic and tactical decisions (e.g. Waters, 1991). This implies that the focus within the production units generally is directed at the short term performance. To support the short term focussed decision making, fast feedback, i.e. a high measurement frequency, is required to timely react on disturbances (e.g. Algera, 1990).

2.4 Performance management in production units

When speaking about the performance of a production unit, one usually refers to the performance on one or more aspects. Chase and Aquilano (1989), for example, give the following list:

- Volume of output.
- Cost (materials, labor, delivery, scrap, etc.).
- Utilization (equipment, labor).
Quality and product reliability.
On-time delivery.
Investment (return on assets).
Flexibility (volume and product change).

Although there is some overlap in the above list and some important performance aspects seem to be missing (e.g. order flow time), the variety in the list makes clear that it is impossible to speak about the performance of a production unit. We therefore make a distinction into organizational related performance aspects and order related performance aspects. The organizational related aspects refer to (long term) organizational choices about for example the product mix, environmental policies and social factors. Order related performance aspects refer to aspects directly related to the characteristics of production orders, i.e. quality, costs, and timeliness. These order related performance aspects are used for internal control purposes. In this research we focus on the order related performance aspects, which are related to the performance of production units in the short term; the organizational related performance aspects are assumed to be given.

The management of the production unit performance is regarded as an ongoing process focussing on ways to improve the actual performance. In this way, this process can be considered as a working-out of the well-known Deming-cycle which originally was developed for quality control purposes. In this cycle, four points of actions are executed, once an activity is found unsatisfactory:

- **Plan** - make plans to improve a given aspect of the performance.
- **Do** - carry out these plans.
- **Check** - compare the actual results with the desired performance.
- **Act** - consolidate the measures that proved successful.

This cycle should be executed again and again, because improvement will always be possible. The latter is based on the philosophy of continuous improvement through elimination of all kinds of waste that especially occur in operations (e.g. Robinson, 1991).

Based on the philosophy of continuous improvement and the Deming-cycle, we developed a performance management process that consists of the following steps (see Figure 2.3):

- **Performance measurement** - the performance on the defined performance measures is reported to all whom it concerns. For this performance measurement, data are required of the organization of the production unit, the way the operations are performed, and the policies.
- **Performance evaluation** - the reported performance is compared with performance targets.
Performance diagnosis - the achieved performance is justified, which will also explain possible performance deviations found at the performance evaluation.

Execute improvement actions - the causes for performance deviations determine what kind of improvement actions are required.

As is shown in Figure 2.3, two kinds of actions are possible. The first kind of actions are related directly to changes or improvements with respect to the production unit itself. That is, the organization, the way operations are executed, and the way policies are defined are subject for change. The second kind of actions are the adjustments of the performance targets. The results of these actions can be evaluated when the next performance measurements are executed, which closes the loop of the performance management process. On this performance management process the outline of this thesis is based.
In Chapter 2 we restricted the research to the performance of complex production units with respect to aspects related to production control. We also made an explicit distinction between performance measurement, performance evaluation, and performance diagnosis in the performance management process. In this chapter we will discuss the performance measurement and evaluation phases (see Figure 3.1).

Further, attention will be paid to the performance targets that are used with the performance evaluation. It will be shown that there is a need for more realistic performance targets at the production unit level for the short term. In section 3.1 we discuss why performance measurement is useful, which specific performance measures can be used at the production unit level, and how these performance measures can be obtained. In section 3.2, attention will be paid to how the performance measures should be chosen and defined in order to get a coherent performance measurement system. In section 3.3, the subject of performance evaluation will be discussed, and how performance targets can be derived. In section 3.4 an overview will be given of performance evaluation models. Some issues related to the frequency of performance measurement and evaluation will be discussed in Section 3.5. In Section 3.6 a summary of this chapter will be given and a short preview of the contents of the subsequent chapters.
3.1 Performance measurement and performance measures

In every type of organization performance measurements are executed and reported. Technically, measurement is the process of obtaining symbols to represent the properties of objects, events, or states (Mason and Swanson, 1981). These symbols should have the same relevant relationship to each other as do the things they represent. A very broad definition of the function of measurement is given by Churchman (1981). He states that the measurement's function is "to develop a method for generating a class of information that will be useful in a wide variety of problems and situations". This class of information will be used to base (managerial) decisions on. Mason and Swanson (1981) let this class of information respond to either as attention directing ("What problems shall I look into?"), or as problem solving ("What course of action is better?"), or as scorecard keeping ("How well am I doing?"). The problem solving function will not be addressed in this chapter, because we consider problem solving as a method that should be used in the performance diagnosis phase instead of the performance measurement or evaluation phase. However, some authors believe that solutions automatically follow from the measurements. For example, Novitsky (1986) states that performance measurement is done to get continuous improvement, without saying how this process of improvement can be initiated and continued. Edson (1988) mentions that continual monitoring of key performance criteria leads to corrective actions for improvements, but he also does not make clear how the right corrective actions can be found. So, performance measurement does not automatically give an answer to the question how good the actual performance is, neither does it give suggestions for where performance improvements are possible. Subsequent steps will be necessary to overcome these "shortcomings" of performance measurement, as can also be deduced from Figure 3.1. Performance measurement thus can be regarded as a starting point for further analyses. Therefore, we consider the performance measurement's function as directing someone's attention. An example of a performance measurement system directed at the monitoring of operational processes is the performance evaluation and diagnosis system developed by Wiendahl et al. (1994). The Balanced Scorecard, developed by Kaplan and Norton (1992), is another example of such a system, although it is more directed at the organizational performance as a whole instead of a production unit's performance.

Having explained why measurements are useful, we now have to answer the question which performance measures are important. In the past, mainly financial performance measures were of interest to management, such as return on investment and net profit. In the 1970s and 1980s, the Japanese industry came up with a new focus on manufacturing; quality and time became competitive factors (e.g. Schonberger, 1982). As a result, customers became aware
of their influence and they were not anymore focusing on low product prices only, but they also demanded timely deliveries and high quality products. For western manufacturers this change from a buyers to a sellers market meant that the company's performance could not only be depicted by financial performance measures only, although the financial performance is ultimately leading for an organization's existence. Besides the financial performance measures, also performance measures for the quality as well as the time aspect were needed. However, the relation with or the translation to financial measures, still remains important, because the traditional accounting systems can only deal with the financial performance. The disadvantages of the current financial performance measurement systems with respect to the manufacturing and market changes are very well described by Kaplan (1983). Also Anderson et al. (1989) discuss how this "performance gap" between financial and non-financial measurements occurred and they provide a measuring model to bridge this gap. Other solutions to overcome some of the problems involved with the current accounting systems are, for example, Activity Based Costing (e.g. Cooper et al., 1992) and Production and Inventory Control Cash Flow Analysis (Corbey, 1995).

At the production unit level, non-financial performance measures have always dominated the financial ones, because only non-financial measures are relevant for planning and controlling purposes. For instance, the total number of products to be made in a certain planning period is a more useful measure for an operations manager and his subordinates than the same measure expressed in monetary terms, i.e. the planned sales for that period. Globerson and Riggs (1989) put it this way: "although financial measures capture the monetary consequences of operational performance, they are too broad to help managers cope with such daily decisions as resource allocation and job design". Although the non-financial measures are not new for the operational level, they have become of greater importance because of the increasing interest of higher management levels in the non-financial performance. This increasing interest of operations management as well as top management in the non-financial performance can also be found in the literature about performance measurement. Ferdows and De Meyer (1990) for example, show a change in interest from performance measures directed at financial aspects towards measures that can be used for production planning and control. Also Kenny and Dunk's (1989) survey about which measures to use, indicates that non-financial performance measures are becoming more important. A survey in Dutch manufacturing firms, finally, also illustrates the importance of non-financial measures compared with the financial measures (Ministerie Economische Zaken, 1994).

So far, we have explained that the main part of performance measures for the operational level consists of non-financial measures. But the question still remains which performance
measures should be used. In general, there are two streams in the literature dealing with performance measures at the production unit level. The first part of the literature discusses specific performance measures that may be useful for a manufacturing organization or production unit (e.g. Van Elswijk, 1990, Gelders, 1990, Duijker, 1990, Cross and Lynch, 1990). Usually, in these kinds of papers only the measures are given with an explanation why they can be useful. But the definite choice of the measures is left to the reader. This choice may be very difficult, because of the amount of performance measures listed. Some authors avoid lists of performance measures by presenting only the measures that are most useful to their opinion. For example, Maskell (1991) presents the performance measures that he considers to be the relevant ones for world class manufacturers. Another way to make the choice of the performance measures easier is the use of classifications of performance measures. The structured way the measures are provided in the classifications is helpful to decide which measures can be used for a specific part of the production process or organization. The following classifications can be found in the performance measurement literature:

- Input and output measures. Input measures are for internal use (e.g. capacity utilization), whereas the external measures give information about how the customers judge the performance (e.g. delivery reliability). More about this classification is given by Fortuin (1988).

- Hierarchically based measures. In this classification it is assumed that performance measures are linked together hierarchically, which results in different measures for each organizational level. For example, the average production order flow time at the production system level is the result of the production order flow times within several production units. Cross and Lynch (1990) describe this classification in more detail.

- Effectiveness and efficiency measures. This distinction can be found in McKaige (1989), and Azzone et al. (1991). Effectiveness performance measures are related to the extent to which certain goals are obtained, while efficiency performance measures deal with the way how the goal has been obtained, i.e. they relate the outputs of the processes to the inputs. Generally, effectiveness measures have a higher priority than efficiency measures, because being efficient does not guarantee survival.

- Financial and non-financial measures. This distinction has been discussed earlier in this section. For production unit control, this distinction is of little importance, because mainly non-financial performance measures are being used.

It should be noted that an individual performance measure can be placed into each classification. For example, the delivery reliability of a production unit is an output measure, hierarchically based at the production unit level, referring to the effectiveness of the production unit, and a typical non-financial measure. Further, it should be noted that the
existence of these classifications does not intend to create a value judgement about the usefulness of a specific performance measure; the performance measures in the difference classes should be considered as complementary rather than conflicting.

The second part of the literature deals with the development of performance measures (e.g. Globerson, 1985, Wisner and Fawcett, 1991). The development steps described in the literature are about the same. The first step usually is the brainstorming phase, where a group of people generates a list of candidate performance measures. Then, these candidate measures will be grouped, and redundant measures will be deleted. Finally, each chosen performance measure is defined. It should be noted that these development steps are not specifically directed at the production unit level. Further, it should be avoided to develop a too large number of performance measures, because the function of attention directing of the measurements will lose its effect and the possibility of redundancy between performance measures will be larger. This requirement towards the development of performance measures and other requirements towards individual performance measures are discussed by Keegan et al. (1989), NEVEM (1989), and Maskell (1991).

In deciding which specific performance measures are appropriate for a specific situation, the first part of the literature may give a global insight in what kinds of measures may be relevant. One can choose a number of performance measures from the lists and start to implement them. Although this is a quick way of getting performance measures, this method is not recommended for several reasons. First, the chosen performance measures won't be directed at the specific characteristics of the considered production unit. Second, this approach can be critical with respect to the acceptance of the performance measures, because there will be a chance of lacking commitment to the performance measures that were chosen and their exact definitions. Third, the set of chosen performance measures may lack structure in the sense that there is little or no consistency between performance measures mutually. Because of these disadvantages, the development process of performance measures is recommended.

Although performance measures with regard to production control for different production shops will differ, some aggregate performance measures with respect to production unit control will be applicable to every production unit, namely delivery reliability, average production order flow time, work in process, and capacity utilization. These are not only the global performance measures that are reported in the literature about performance measurement in operations, but they are also used in the performance evaluation models that will be described in Section 3.4. For these reasons, and because of generalization considerations, these performance measures will be used as the basic performance measures.
in our discussions. These performance measures can be regarded as detailed measures of the general aspects quality, costs, and timeliness as described in Chapter 2. Clearly, delivery reliability and average production order flow time are related to the timeliness aspect. The amount of work in process can be translated easily to the costs aspect. Capacity utilization, particularly machine utilization, can also be related to the cost aspect by management accounting systems to attribute writings off to cost prizes of products. The quality aspect will only be considered in an indirect way; if there are product quality problems, extra production orders are required which has an impact on the timeliness performance measures (e.g. a lower delivery reliability) as well as the capacity utilization rates.

3.2 Performance measurement systems

Having chosen appropriate performance measures does not automatically mean that one has developed a performance measurement system. The total set of chosen measures may lack structure, especially when the measures are just picked from the available lists in the literature. In developing a performance measurement system, relationships between performance measures play an important role. There are direct hierarchical relationships and more hidden relationships in the sense that some measures influence each other in a more indirect way.

The hierarchical relationships between different performance measures, originally result from translating the strategic goals of the organizations to individual measures at each organizational level (e.g. Globerson and Riggs, 1989, Cross and Lynch, 1990). Wisner and Fawcett (1991) state that a performance measurement system should be the link between organizational strategy and operational decisions. The second kind of relationships appears when trade-offs have to be made. The trade-off between costs and another aspect, such as time or quality, is a well-known one. For example, management can decide to decrease the number of rejected products by implementing more quality checks in production which increases quality costs. The decrease in scrap can be translated into monetary terms and so compared with the costs involved with extra quality checks. However, this comparison is not complete, because there are also positive side-effects with the decrease of scrap, such as shorter production order flow times and a higher delivery reliability. Because these effects have another measurement dimension (i.e. time and a percentage respectively) a trade-off is difficult to make. The above example makes clear that no optimal decision is possible (only a satisfying decision) when trade-offs are made between performance measures with different measurement dimensions. Therefore, an explicit statement is necessary about priority setting.
between different performance measures. For the operational level, St. John and Young (1992) discuss the subject of priority setting and trade-offs, illustrated by a survey among operations managers that shows which operational measures have the highest priority. For specific performance measures with regard to production control such as batch size, order flow time and capacity utilization, Durlinger (1985, 1986) has developed a model to help goods flow control managers and production unit managers to make satisfying decisions. More about relationships between specific performance measures for production unit control can be found in Karmarkar (1989), and Rao (1992).

In the development steps to choose and define performance measures, described in Section 3.1, the most important measures may have been chosen, and relationships and trade-offs might have been discussed to some extent. However, no check has been made that far to see whether the system is complete as well as coherent. Flapper et al. (1996) present a method by which one is able to evaluate the consistency of the chosen performance measures. Further they introduce so-called parent-child relationships between performance measures in order to judge whether all chosen performance measures are related to each other.

For the remainder of this research, we assume that a performance measurement system for production unit control and goods flow control is available. This means that we assume that enough data are available yet or that the required data for performance evaluation and diagnosis can be relatively easily and quickly obtained from the data available.

3.3 Performance evaluation and performance targets

An unambiguous definition of the term performance evaluation cannot be found in the literature. Suri et al. (1993) define performance evaluation as "a methodology (including techniques and tools) for determining the performance measures that can be expected to result from a given set of decisions". From this definition, two characteristics emerge as being important. First, it is stated that the evaluation is directed at the expected performance, which indicates that performance evaluation deals with the future performance. Second, the authors speak about a given set of decisions. Together with the expectations about the performance, we can conclude that this set of decisions can be changed. This means that performance evaluation can function as a what-if instrument in order to determine the best set of decisions.

According to the ELA-Terminology, evaluation can be defined as: "the assessment of a situation possibly in comparison with plans and/or standards previously set as a target" (ELA, Performance measurement and evaluation 25
From Figure 3.1, it can be concluded that this definition of performance evaluation be used better than Suri et al.'s definition, because we compare the actual performance with performance targets. The obtaining of the performance targets we also include into the performance evaluation process. How these performance targets can be obtained will be discussed next.

For each performance measure a performance target should be available, otherwise the scorecard keeping function of the performance evaluation, as described in Section 3.1, cannot be fulfilled. There are several ways to set performance targets. They can be based on:

- intuition;
- strategic targets;
- competition analysis;
- benchmarking;
- agreements;
- average, past performance;
- quantitative performance models.

Taking the thought of continuous improvement as a starting point, as was described in Chapter 2, performance targets should meet at least three criteria. First of all, performance targets should be realistic for the period of time they are set for. This means that the performance target should be based on the length of the period for which the performance is measured, which we will call the measurement period. Second, and related to the first condition, performance targets should be challenging (e.g. Latham and Locke, 1991). If there is no drive for improvement, people not easily will think about ways to improve their performance. Third, performance targets should be as objective as possible to avoid discussions about the performance targets themselves; at the performance evaluation, the actual performance should be subject to discussion and not the performance targets.

**Performance targets based on intuition**

In situations where the actual performance is not known, it is very difficult to set realistic and challenging performance targets for a certain period of time. Performance targets can only be based on the intuition of the person or team that sets the target. This method to set performance targets has the danger that the targets are not realistic for the defined period of time and, more importantly, not objective. To avoid this problem, one can better monitor the actual performance for a couple of measurement periods in order to get an idea of realistic performance targets: for the first couple of measurement periods no performance targets should be set.
Strategic targets as performance targets
When the actual performance is known, management can set performance targets that must be realized within a long period of time. In this case, we speak about strategic performance targets. Performance targets are primarily challenging; they are set at a significant higher level than the actual performance. The extent to which strategic performance targets are realistic for the long term is difficult to determine. However, it is recognized that the actual performance cannot improve significantly over a short period of time. Therefore, the performance evaluation in the short term should not take strategic performance targets as performance targets for short measurements periods, because strategic performance targets are not realistic for short term evaluation purposes.

Performance targets based on competition analysis
A third method to obtain performance targets is by applying the traditional competition analysis, where the actual performance of a production unit is compared with the performance of similar units of the own organization or of competitive organizations. However, one should be careful with looking at the performance of competitors, because one may very easily compare apples with oranges, because strategic organizational choices can be different as well as the organizational structure. For these reasons, we think that competition analysis can only be useful in setting strategic performance targets or long term goals.

Benchmarking as method to set performance targets
Benchmarking goes far beyond traditional competition analysis (Karlöf and Östblom, 1994). Here, one looks for organizations that perform their operations with outstandingly good results. This "best practice" organization not necessarily should produce the same kinds of products or should have the same type of organizational structure. The way in which these outstanding organizations obtain their high performance is then tried to be copied to the own organization. Usually, the focus and structure of the organization and the skills of its leaders and members are subject for change. In Chapter 2 we already stated that choices about the organization of a production units will not be considered in this research, because this is an issue for the long term. Therefore, benchmarking will not be regarded as an appropriate method to set realistic performance targets for the short term.

Performance targets established by agreements
In many situations the performance is established by different organizational members, each having his own function. Basically, each of the people involved is able to have influence on the future performance. Therefore, the people involved should discuss at what level the performance targets will be set, until total agreement has been reached. In fact, trade-offs are
being made between organization members to find satisfying performance targets. An example of this method where it is also tried to obtain challenging performance targets, is ProMES (Productivity Measurement and Enhancement Method) developed by Pritchard et al. (1988). ProMES mainly has been used at lower organizational levels so far. To set performance targets for the short term this way requires discussions every measurement period, which may be considered as too frequently.

**Performance targets obtained via past performance**

A performance target can be obtained from the performance of the near history. For example, the average performance over the last month's delivery reliability may be used for performance target for the near future. It thereby is assumed that the past performance is a good reflection of the performance for the near future measurement periods. If the performance in subsequent measurement periods varies a lot, then the performance targets for near future measurement periods may be far from realistic. Also, there is no relationship with the future possibilities or a strive for improvement, which undermines the challenging criterium.

**Performance targets generated by quantitative models**

The final method is for getting performance targets is using quantitative models. The modelling phase as well as the quality of the inputs of the model are most crucial here. If the modelling has been done correctly and the input of the model is reliable, this method leads to very realistic performance targets, because it takes the specific situation the model refers to as the starting point. Because no people are directly involved with the target setting by a model, this method of target setting is very objective.

Weighing the pros and cons against each other with respect to the three criteria, performance target setting by quantitative models seems to be the best way. Compared to the other methods, this method of target setting is the most objective method, because no people are directly involved with the generating of the performance targets. Further, such a model can determine the maximum achievable performance, given the specific characteristics and organization of a production unit. This maximum achievable performance supports both the challenging criterium and the criterium of realism for the short term. Examples of quantitative models will be described in the next section. As can be noted, the function of the quantitative model is predictive by nature; it should generate, or predict, the performance targets for the short term.
3.4 Performance models

For production planning and control, numerous models are available. However, the use of these models in practice often raises considerable resistance, and also the success can be very limited (Fortuin et al., 1993). According to Bertrand et al. (1990a) this opposition and lack of confidence in models primarily results from the way models are used. Because each model has a different supporting function, each level of decision making needs its own special type of model. Therefore, they distinguish three types of models, each one related to a different level of decision making:

- Operational models that support operational decisions. Usually these models are embedded in rolling plans. Examples are scheduling models.
- Performance models to support tactical decisions by estimating the performance of controlled systems. Often, parameter setting is done by sensitivity analyses.
- Structure supporting models to support strategic decisions.

We will focus our attention to the operational models, because these models have to deal with the dynamics of production units in the short term.

A rather complete overview of performance models directed at production unit control and goods flow control is given by Suri et al. (1993). They make a distinction into:

- Static (allocation) models.
- Aggregate dynamic models.
- Detailed dynamic models.

Static models simply add up the total amount of work allocated to each resource, and estimate the performance from these totals. An example of this type of model is the rough cut capacity planning module in MRP-II systems. Aggregate dynamic models take into account the dynamics, interactions, and uncertainties in a system in an aggregate way. Often, the estimated performance is only the average performance when the system is in steady-state (a system is in steady-state when the probability distribution of all possible states remains constant). In detailed dynamic models, manufacturing systems can be modelled in considerable detail. These models can be deterministic or stochastic. Examples of deterministic detailed dynamic models are scheduling models. Detailed stochastic models include simulation models.

At first glance, the aggregate and detailed dynamic models are the most appropriate models for our research goals. Examples of aggregate models directed at the performance of production units are the Queueing Network Analyzer (Whitt, 1983), Manuplan II (or the...
commercial software version MPX) (Suri, 1989), Trade-Off-Module (Durlinger, 1985, 1986) and Mean Value Analysis (Reiser and Lavenberg, 1980). All these models take queueing theory as a starting point for their estimations and analyses of the performance. Typical performance measures for these models are: average waiting times at workcenters, average order flow times, and average number of orders in the system. In short, these models are directed at the average performance assuming that the system is in steady-state. However, in production units in practice the assumptions underlying the steady-state models (e.g. equally distributed demand pattern in subsequent measurement periods, equally distributed order processing times in subsequent measurement periods) rarely match reality. As a consequence, these models are not appropriate for operational, daily decisions. Further, all kinds of daily disturbances may occur in production units (e.g. Stoop and Wiers, 1995) that make it impossible to achieve the average performance in every measurement period. Lin et al. (1992) put it as follows: "due to the lack of viable modelling techniques, abnormal situations at the operational level in a dynamic environment cannot be handled by shop floor production control systems that are only applicable to steady-state performance".

Even if the processes at the production unit would be stable, the use of a steady-state or average performance as performance target is not realistic for production unit management. Over a longer period of time the average actual performance should be equal to the average performance in steady-state, but this usually will not be the case for each measurement period. Normally, the actual performance will fluctuate over time. So there will be periods in which the actual performance is better than the average performance in steady-state, but there will also be periods in which the actual performance is worse than this average. Especially in the latter cases, there may be a danger that production unit management makes nervous decisions trying to reach the average performance in steady-state. To avoid this nervousness, that also result in extra costs, realistic performance targets are necessary for the short term. Only then, production unit management can make the best decisions. And, we should not forget that the long term performance results from astute short term decisions (Globerson & Riggs, 1989). In practice, a short term focus in production units can often be observed. Daily problems caused by all kinds of disturbances should be solved and their is little time left to think about the medium/long term performance of the production unit. Because current performance models are not directed at the short term performance, we will develop such a model to generate short term performance targets that are realistic for short measurement periods. The development of this model directed at the short term will be described in Chapter 4.
3.5 Frequency of performance measurement and evaluation

Each measurement is involved with costs and takes (computational) time. So, from a financial perspective, the number of measurements should be as small as possible. On the other hand, measurements are necessary for the performance evaluation and performance diagnosis. The more frequent the measurements are being made, the better one is able to pay attention to the real problems that occur or to foresee certain problems. This is illustrated in Figure 3.2. Measurement frequency is defined as the number of measurements per time unit. If a frequency of 0.1 (i.e. one measurement per 10 time units) would have been chosen, the peak around time unit 5 would not have been noticed. But in case of a measurement frequency of 1, the peak would have been noticed, and even on time.

![Figure 3.2 Example of performance measurements with a different frequency.](image)

Further, for explaining the observed problems, a short measurement period is also recommended. People are not only able to memorize the problem and possible causes for the problem, but also some important data that can be used as evidence may have been lost due to storage capacity of data. Thus, the best explanations can be given very short after the reporting of the actual performance. Pritchard et al. (1988) for example, have shown that short term feedback on the performance is very important from a psychological point of view. On the other hand, from a controlling point of view, short measurement periods may lead to nervous decision behavior. If one observes a sudden deviation from the performance target, one should be careful in taking corrective measures. It is better to base the corrective actions on trends rather than just one observation. It also is preferable to execute the corrective actions in a smooth way.
Generally, it can be said that the length of measurement period should be a function of the probability that the state of the system changes. Actually, if the length of a measurement period is set in such a way that the probability of a change in the state of the system is zero, then the measurements are of little use because they will not initiate decision-making nor performance evaluations. The definition of the state will depend on the type of decision to be made. For example, when the average production order flow time in a production unit is 4 weeks and the average work in process is 100 production orders, a daily measurement of the average order flow time is of little use for evaluating the progress of individual production orders. However, for capacity allocation purposes, a daily report about the work in process at each workcenter, can be very useful. For production order scheduling in general, Steudel and Wu (1977) give a rough rule of thumb that the length of a measurement period to determine the work in process should be approximately ten times the average processing time of a production order. Steudel and Wu base this rule of thumb on the analysis of the relationship between the length of a measurement period, the utilization rate and the average production order processing time.

3.6 Summary and preview

Speaking about the performance of a production unit, actually the combined performance with respect to the delivery reliability, average production order flow time, work in process, and capacity utilization are referred. Performance evaluation is defined as the process of comparing the actual performance with performance targets. We thereby assume that a performance measurement system is available that reports the actual performance with a certain frequency. To avoid discussions about the performance targets themselves, performance targets should be objective. From this point of view the use of quantitative models for performance target setting is appropriate. The obtainment of performance targets for production units will be done by using a prediction model directed at the short term performance, and will be discussed in the next chapter. The place of the performance prediction model in the performance management process is depicted in Figure 3.3. A new activity, i.e. performance prediction, has been added, which uses the performance prediction model to generate performance targets. The performance prediction is executed with the same frequency as the frequency of the performance measurements. In this way, the performance targets relate to the same period of time the performance measurements are based on.

The result of the evaluation phase is the judgement whether the actual performance deviates from the performance targets or not. This comparison, however, does not explain how the
Actual performance has been realized. For this explanation, a performance diagnosis is needed. By this performance diagnosis it is determined whether the maximum achievable performance has been realized. In Chapter 6 we will elaborate on the subject of performance diagnosis. As can be seen in Figure 3.3, an eventual change in the performance targets (which is a possible action that results from the performance diagnosis) should be made by a change in the performance prediction model. This will also be described in Chapter 6.
CHAPTER 4

PREDICTION MODEL

In Chapter 3 we concluded that for performance evaluation purposes at the production unit level the available models do not suffice. First, the available models mainly are directed at the average performance in a steady-state situation, whereas the actual situation in complex production units will never be in a steady-state due to continuous changes in the environment. Second, the focus in production units is directed at the short term to solve daily problems caused by all kinds of daily disturbances; people in the production units hardly have any time to think about the medium/long term performance of the production unit. Those two reasons form the motives for developing a model that is directed at the short term performance.

We assume that the performance of a production unit in the short term will be determined largely by the actual state of the production unit. The prediction model thus will be a state-dependent model. It is this characteristic which makes the performance prediction model different from the performance models described in Section 3.4. The state-dependent performance prediction model will be used to obtain realistic performance targets in the short term.

In Section 4.1, we will first describe the variables that we take into account in the state-dependent model. This set of variables cover the actual state of a production unit. Also in this section, the specific performance measures to be predicted will be defined. In Section 4.2 and 4.3, a general description of the prediction rules for the chosen performance measures will be given. Specific interpretations of these prediction rules are dependent on the situations they are developed for. Examples of these specific interpretations will be given in Chapter 5, where the quality of the state-dependent prediction rules are tested compared with some state-independent prediction rules, and in Chapter 7, where some specific state-dependent prediction rules were implemented and used in two production units in practice. A discussion about the prediction errors resulting from the state-dependent prediction rules will be held in Section 4.4. Also the possibilities for extensions to the state-dependent prediction rules will be discussed in this section.
4.1 Model variables and performance measures

The basis for our state-dependent predictions is formed by the length of the performance measurement periods. The predictions will be made at the beginning of a measurement period so that for each measurement period a performance target is available that can be used for the performance evaluation. In this way the time between two performance evaluations is coupled to the frequency of the performance measurements. The length of a measurement period will be denoted by $T$ and the measurement periods will be numbered $n, n+1, n+2$, etcetera (see Figure 4.1), where measurement period $n$ is the measurement period that is just beginning, starting at time $t$.

The following three variables are considered to have a significant influence on the short term performance of a production unit:

- The amount of work in process.
- The planned or expected amount of capacity available.
- The planned or expected amount of work supply.

The combination of these variables we will call the actual state of a production unit. We will start the development of the state-dependent prediction rules at workcenter level. Therefore, the variables we will discuss next refer to individual workcenters. At the end of this section the aggregation to production unit level will be made.

The amount of work in process at a workcenter is measured in hours and is equal to the sum of the processing times of the production orders that are waiting and not finished at the workcenter. The work in process for workcenter $i$ at the beginning of measurement period $n$ will be denoted by $I_i(n)$. Because this amount of work in process is directly observable, this will be called the actual work in process at workcenter $i$. The expected amount of the work in process at workcenter $i$ at the beginning of future measurement periods will be denoted by $\hat{I}_i(n+1), \hat{I}_i(n+2)$, etcetera.
The planned or expected net available capacity during measurement period \( n \), measured in hours. For a workcenter \( i \) this variable will be denoted by \( \hat{C}_i(n) \). The expected net available capacity at a workcenter is a function of the net available operator capacity that is expected to be allocated, and expected net available machine capacity:

\[
\hat{C}_i(n) = f(\hat{O}_i(n), \hat{A}_i(n))
\]  

(4.1)

where \( \hat{O}_i(n) \) is the expected net available operator capacity, and \( \hat{A}_i(n) \) the expected net available machine capacity, both at workcenter \( i \) during measurement period \( n \) and measured in hours. It will depend on the specific situation, whether the allocated operator capacity or the available machine capacity is the constraining factor with respect to the amount of work that can be processed at the workcenter. For example, when two operators have been allocated to a certain workcenter with three machines, and both operators are skilled enough to organize their work in such a way that the three machines can work all at once, the available machine capacity will determine the maximum throughput of the workcenter. However, when two other operators, who are not as skilled as the operators mentioned above, would have been allocated, the throughput would have been constrained by the available operator capacity. We assume that the user of the model will specify what resource is the most constraining.

The planned or expected amount of work supply during measurement period \( n \) at workcenter \( i \), measured in hours will be denoted by \( \hat{W}_i(n) \). The expected work supply of a workcenter during measurement period \( n \) consists of new orders that are expected to be released for production (the expected external work supply), \( \hat{W}_{\text{ext}}(n) \), and orders expected to come from workcenters within the production unit (the expected internal work supply), \( \hat{W}_{\text{int}}(n) \). In formula:

\[
\hat{W}_i(T) = \hat{W}_{i,\text{ext}}(T) + \hat{W}_{i,\text{int}}(T)
\]

(4.2)

In practical situations, the available capacity and expected work supply during a measurement period at a workcenter usually differ from the long term averages, denoted by \( \overline{C}_i \) and \( \overline{W}_i \) respectively. It may be obvious that for individual measurement periods in real life \( \hat{C}_i(n) \) and \( \hat{W}_i(n) \) will be better estimates than \( \overline{C}_i \) and \( \overline{W}_i \) because information about the actual and near-future situation is taken into account. The relationship between \( \hat{C}_i(n) \) and \( \overline{C}_i \) respectively \( \hat{W}_i(n) \) and \( \overline{W}_i \) can be formulated as follows:

\[
\lim_{x \to \infty} \sum_{n=1}^{x} \frac{\hat{C}_i(n)}{x} = \overline{C}_i
\]

(4.3)

*Prediction model*
The quantitative models described in Section 3.4 only use $C$ and $W$ as model variables because they are more directed at the performance on the medium/long term. Because we take into account some relevant state-information, we expect that our model will lead to lower variations in the performance prediction errors. This will be tested in the next chapter.

If there is no knowledge about the expected available capacity for the next measurement period, it is suggested to use the average available capacity as an estimate. The same holds for the expected external work supply; the average work supply with mean order characteristics can be used as an estimation for the external work supply. In this situation, the external work supply will have no variations per measurement period. The expected internal work supply will always be a state-dependent variable because it is calculated from the available data about the actual work in process and the expected available capacity for the next measurement period, as will be described in the Section 4.3.

The above model variables refer to the workcenter level. An aggregation to the production unit level can easily be made. This leads to the following relationships for the work in process, available capacity, and work supply respectively:

$$\hat{I}(n) = \sum_{i=1}^{m} \hat{I}_i(n)$$  \hspace{1cm} (4.5)

$$\hat{C}(n) = \sum_{i=1}^{m} \hat{C}_i(n)$$  \hspace{1cm} (4.6)

$$\hat{W}(n) = \sum_{i=1}^{m} \hat{W}_i(n)$$  \hspace{1cm} (4.7)

where $m$ is the total number of workcenters in a production unit, and $i$ the index denoting the workcenter number.
From our definitions of the model variables, one can conclude that these variables are "time-aggregated" in nature, in the way that the specific moments events occur within a prediction period are not taken into account. For example, if one has planned preventive maintenance on a certain machine in a workcenter, the expected duration of the maintenance is subtracted from the expected total available machine hours in the coming measurement period. The specific moment of the action, for example at the beginning of the measurement period, is not considered. There are several reasons for this simplification. First, the aggregation keeps the prediction rules simple and understandable, which especially is important for the acceptance of the prediction rules in practice. Second, the requirement of aggregate data will require less effort than the requirement of detailed data. Third, it is more difficult to predict when a certain event takes place than to predict the total impact of that type of event if it occurs; this relates to the robustness of the model. If no time-aggregation would have been used, knowledge about the distribution of specific event times would be a first requirement. However, each predicted event derived from this distribution for the next prediction period has a very little probability of occurrence, so the prediction of future events will never match the actual occurrences. For that reason the development of a discrete event simulation model is not considered here.

In Chapter 3 we shortly discussed the performance measures that are usually being used in performance models as well as in practice. These performance measures were: production order flow time, delivery reliability, throughput, and capacity utilization. Here, we restrict the performance measures to be predicted to the throughput of a production unit for the next measurement period (i.e. the sum of the processing times of production orders' operations that are completed) and the completion times of production orders currently in the production unit. From these two kinds of predictions, the other relevant performance measures can easily be derived. Comparing the predicted completion time of each production order with its start date in production, the expected production order flow time can be calculated. The comparison of the predicted completion time with a given production order's due date, determines the expected lateness of a production order. This lateness can be used in the delivery reliability calculations. The expected capacity utilization, finally, can be obtained by dividing the expected throughput by the expected available capacity.

The relationships between the above model variables and performance measures are depicted in Figure 4.2. The arrows show the direction of the influence of the model variables and performance measures on each other. From the figure, it can be seen that the prediction of the performance measures can be quite complicated. This is mainly due to the indirect interactions. For example, the operator capacity allocation decision usually is largely based

*Prediction model* 39
on the actual distribution of the work in process over the workcenters and to some extent on the expected work supply (e.g., Stoop, 1990). The capacity allocation decision will result in a certain throughput on each workcenter which in turn leads to a flow of production orders to other workcenters within the production unit (i.e., the internal work supply). But this expected internal work supply already was the basis for the capacity allocation decision, which makes the allocation decision look like a paradoxical decision. Due to this kind of interactions, it will be clear that detailed discrete event modelling is impossible in the kind of production units we defined as subject of research.

4.2 Throughput prediction rule

The throughput prediction for measurement period $n$ at workcenter $i$, denoted by $\hat{Y}_i(n)$, is determined by the actual work in process, $I_i(n)$, the expected available capacity in measurement period $n$, $\hat{C}_i(n)$, and the expected work supply in measurement period $n$, $\hat{W}_i(n)$. Analogue to the aggregation of the work in process, available capacity, and work supply from workcenter level to production unit level the production unit's expected throughput, $\hat{Y}(n)$, can be obtained by totalling the expected throughput of the $m$ individual workcenters in the production unit:

\[ \hat{Y}(n) = \sum_{i=1}^{m} \hat{Y}_i(n) \]
\[
\hat{Y}(n) = \sum_{i=1}^{m} \hat{Y}_i(n)
\]  
(4.8)

If the actual work in process at the beginning of a measurement period is higher than the expected available capacity for that period, then it is evident that the predicted throughput for that period is equal to the expected available capacity. In formula:

\[
\hat{Y}_i(n) = \hat{C}_i(n), \quad \text{if} \quad I_i(n) \geq \hat{C}_i(n)
\]  
(4.9)

In case the actual work in process is lower than the expected available capacity, a more complex calculation is necessary because we have to determine the fraction of the work supply that is expected to arrive within the next measurement period that will be processed. Assuming that the remaining capacity, which is the capacity that remains after having processed the actual work in process, will be used for processing (a part of) the work supply, this fraction will depend on the amount of remaining capacity; if the remaining capacity is high, a larger part of the work supply can be processed than if the remaining capacity is low. The remaining capacity varies each measurement period and actually depends on the amount of work in process. So actually, the fraction of the work supply that can be processed depends on the actual work in process. Let \( \alpha[I_i(n)] \) be the fraction of the work supply that on average will be processed in measurement period \( n \), given the actual work in process, then the predicted throughput can be formulated as follows:

\[
\hat{Y}_i(n) = \min[I_i(n) + \alpha[I_i(n)] \hat{W}_i(n), \hat{C}_i(n)], \quad \text{if} \quad I_i(n) < \hat{C}_i(n)
\]  
(4.10)

Clearly, in case \( I_i(n) \geq \hat{C}_i(n) \) there will be no remaining capacity to process (a part of) the work supply, which implies that \( \alpha[I_i(n)] \) is zero. The essence of the calculation of \( \alpha[I_i(n)] \) is illustrated in Figure 4.3. First, the expected available capacity is used to process the actual work in process, which will be finished at time \( t+I_i(n) \) assuming that there is one machine in that workcenter. (If there are more machines, say \( k \), then the expected time the actual work in process is finished is approximated by \( t+I_i(n)/k \). At this moment of time, at least a part of the expected total amount of work supply shall be arrived at the workcenter. This expected amount at that specific moment of time, \( \hat{I}_i(t+I_i(n)) \), can be estimated by:

\[
\hat{I}_i(t+I_i(n)) = \frac{I_i(n)}{\hat{C}_i(n)} \hat{W}_i(n)
\]  
(4.11)

Prediction model

41
Figure 4.3 Required variables to calculate $a[I_t(n)]$.

assuming that on the average work supply arrives about equally spread during measurement period n, which in the figure is depicted by the striped line marked with $\hat{W}_i(n)$. Next, an estimation should be made for the expected amount of work in process the end of measurement period n (i.e. time $t+T$), $\hat{I}_i(n+1)$. For this estimation several methods can be used. In Chapter 5 and Chapter 7 two examples will be described of such kinds of estimation methods. The basic idea of the estimation methods is that the estimated work in process goes in the direction of the average work in process. In the example of Figure 4.3 this "natural drift" of an arbitrary amount of work in process towards the average amount is illustrated for the situation that the average amount of work in process is larger than the expected amount of work in process at time $t+I_i(n)$. In the example is the difference between $\hat{W}_i(n)$ and $\hat{I}_i(n+1)$ the expected amount of the work supply that cannot be processed in measurement period n. Hence, the formula for the fraction of the work supply that on average will be processed is:

$$a[I_t(n)] = \frac{\hat{W}_i(n) - \hat{I}_i(n+1)}{\hat{W}_i(n)} \quad (4.12)$$

A specific example of the calculation of the $a[I_t(n)]$'s will be given in the next chapter.

4.3 Production order completion time prediction rule

The estimation of production order completion times will be directed at individual production orders. Just like for the throughput predictions, for these predictions we take into account
information about the actual work in process, the expected available capacity, and the expected work supply. The specific characteristics of the production orders (e.g. processing times, routings, planned due dates) that are part of the work in process thus are also used for the production order completion time predictions. The prediction of the completion time of a production order is split up into two parts:

- A prediction of the expected production order flow time at the workcenter the production order is currently waiting at.
- A prediction of the production order flow time at the remaining workcenters in the routing of the production order.

The summation of the above two expected production order flow times is the total expected flow time of the production order. The actual moment of time plus this total expected production order flow time will then be the expected production order completion time (if relevant, future non-working days should be taken into account).

The expected flow time of a production order at the workcenter the production order is currently waiting for can be determined by the summation of the following factors:

(a) The processing times (including their possible setup times) of production orders with a higher priority that are currently waiting in the same queue. This will be called the minimum waiting time.

(b) The processing times of production orders with a higher priority that are expected to arrive within the period of time calculated under (a), which will be called the expected extra waiting time.

(c) The processing time of the order at the current workcenter.

In formula:

\[
\hat{t}_j, i = \hat{M}_j, i + \hat{\epsilon}_j, i + \hat{\beta}_j, i
\]  

where

\( \hat{t}_j, i \) = expected flow time of production order \( j \) at workcenter \( i \)

\( \hat{M}_j, i \) = minimum waiting time of production order \( j \) at workcenter \( i \)

\( \hat{\epsilon}_j, i \) = expected extra waiting time of production order \( j \) at workcenter \( i \)

\( \hat{\beta}_j, i \) = pre-calculated or expected processing time (including setup time) of production order \( j \) at workcenter \( i \)

It should be noted that we explicitly consider the processing times as estimates. For the manufacturing of complex products this assumption will hold because work preparators will have difficulties to estimate the actual processing times, especially for newly developed products.
\( \hat{M}_{j,i} \) and \( \hat{E}_{j,i} \) are calculated as follows:

\[
\hat{M}_{j,i} = \sum_{x \in V_{j,i}} \hat{p}_{x,i} \tag{4.14}
\]

\[
\hat{E}_{j,i} = \sum_{y \in Z_{j,i}} \hat{p}_{y,i} \tag{4.15}
\]

where

\( V_{j,i} = \) the set of production orders at workcenter \( j \) with a higher priority than production order \( i \)

\( Z_{j,i} = \) set of production orders that are expected to arrive at workcenter \( j \) within the coming measurement period, with a higher priority than production order \( i \)

Because it is assumed that the sequencing rule is known as well as the production order characteristics that determine the priority by the sequencing, the calculation of expected minimum waiting time is straightforward.

The determination of \( Z_{j,i} \) is done as follows. For each workcenter - except for workcenter \( i \) - an estimation is made which production orders will be processed at these workcenters based on the comparison of the current queue length (in hours) at the specific workcenter and the available capacity (in hours) at this workcenter. Then, this set of candidate production orders is reduced by removing the production orders that do not have their subsequent operation at workcenter \( i \). From this remaining set, the processing times of production orders with a higher priority than production order \( j \) are summed to obtain the expected remaining waiting time. We note that we only consider production orders that are one operation before the considered workcenter. This is based on our assumption that the length of the measurement period has been defined in such a way that the processing of more than two operations of a production order during a measurement period occurs only rarely.

The second step in the prediction of a production order's completion time consists of the estimation of the flow time at the remaining workcenters in a production order's routing. The expected waiting times at the other workcenters depend largely on the amount of work in process at the moment the production order arrives. The expected flow time at the first workcenter, say \( k \), in the remaining routing of production order \( j \) can be calculated as follows:

\[
\hat{f}_{k,i} = \beta \hat{f}_{k}(t + \hat{f}_{j,i}) + \hat{p}_{k,j} \tag{4.16}
\]
where $\beta$ represents a factor that depends on the priority rule that is being used. In case First Come First Served is used, $\beta$ will be equal to one; the production order that arrives at workcenter $k$ at time $t+\hat{\beta}$ must wait as long as the sum of the processing times of production orders in the queue at that moment. In other cases, the determination of $\beta$ will be more complex because a specific character of the production order (e.g. its due date or processing time) should be related to the characteristics of the production orders the queue will probably consist of.

As can be observed from equation 4.16, the expected future amount of work in process is needed to determine the expected production order flow time at the workcenters in the remaining routing of a production order. The expected future work in process levels at a workcenter will be estimated for the beginning of future measurement periods only, i.e. at the beginning of measurement period $n+1$ (i.e. time $t+T$), measurement period $n+2$ (i.e. time $t+2T$), etcetera. If the expected amount of work in process is needed for a time moment that does not correspond with the time that a new measurement period begins, then interpolation can be used. The expected amount of work in process at measurement period $n+1$ is calculated from the result of the throughput prediction:

$$ I_j(n+1) = I_j(n) + \hat{W}_j(n) - \hat{Y}_j(n) \quad (4.17) $$

For the estimation of the amount of work in process at the beginning of measurement period $n+2$, measurement period $n+3$, etcetera, the characteristic that the amount of work in process naturally drifts towards the average work in process (as already described in Section 4.2) will be used. The essence of the method has been depicted in Figure 4.4.

![Figure 4.4 The essence of the estimation of future amounts of work in process.](image-url)
In Chapter 5 as well as Chapter 7, specific examples of methods are given that determine the expected amount of work in process in future measurement periods. As can be observed in Figure 4.4 this estimation should be done as long as the expected amount of work in process equals the average amount of work in process. Interpolation is used in case the expected arrival time of a production order is not equal to the time a new measurement period starts. For example, to estimate the amount of work in process at time \( t + 1.5T \), the average of \( \hat{I}_i(n+1) \) and \( \hat{I}_i(n+2) \) is taken. This is depicted in the figure with the dark bar.

Due to the many operational disturbances and delays, we only predict the amount of work at the workcenter because it is impossible to predict which specific orders will be at a workcenter at a specific moment.

The estimation of the flow time of production order \( i \) on the first two workcenters now can be calculated as follows:

\[
\hat{\theta}_i = \hat{\theta}_{j,i} + \hat{\theta}_{k,i}
\] (4.18)

When there are more workcenter left in the remaining routing, the above estimation method starts again for the next workcenter in the routing.

### 4.4 Estimation errors of model variables and model extensions

The definitions of the model variables in Section 4.1 already made clear that two relevant variables, i.e., capacity and work supply, are estimations. This implies that the actual values of these variables can differ largely from the estimated values. If this is the case, the estimation errors will also have an influence on the quality of the throughput predictions and the production order completion time predictions. Therefore, these estimation errors should not be neglected. The estimation errors can be modelled as follows:

\[
\theta_i(n) = \hat{C}_i(n) - C_i(n)
\] (4.19)

\[
\psi_i(n) = \hat{W}_i(n) - W_i(n)
\] (4.20)

where \( \theta_i(n) \) and \( \psi_i(n) \) are variables representing the estimation errors in capacity and work supply respectively. A negative value of these variables thus indicates that the estimation was too low, whereas a positive value shows that the estimated value was too high. Another variable that may influence the quality of the performance predictions is the difference
between the calculated or estimated processing times and the real processing times. Analogous to the definition of the deviations of capacity and work supply, we can define a variable, $e_{j,i}$, representing the estimation error of the processing time of production order $j$ at workcenter $i$:

$$e_{j,i} = \hat{p}_{j,i} - p_{j,i}$$ \hspace{1cm} (4.21)

It is important to know these estimation errors, because the quality of the performance predictions will importantly be determined by the quality of the estimations of these variables. Feedback of the estimation errors to the persons who make the estimations may also lead to improvement of these estimations. As we will see in Chapter 6, we will both need the estimated values as well as the actual values for the performance diagnosis.

The defined performance prediction rules are quite general so far. One should be aware that these rules should be adjusted to the specific situation they will be used for. For example, there are many different ways to release orders to a production unit. If all the work is released at the beginning of a measurement period, there will be a smaller probability of idleness on workcenters than if the work is released at several moments of time during the measurement period. Consequently, this will also have an impact on the defined performance measures. Examples of how such adjustments will be made, will be given in Chapter 5 where specific prediction rules are used in simulation experiments, and in Chapter 7 where performance prediction rules are developed and implemented in two production units in practice.

Besides the adjustments as discussed above that are necessary to match the model with reality, there may also be other relevant variables that should be considered for a specific production unit. The question which variables these are, depends on the specific characteristics of the production unit which is subject of study. It will depend on the extent to which specific characteristics influence the performance whether new variables should be included or not. For example, in production units where the amount of rework caused by product quality defects is relatively large, and the amount of hours spent on rework is not included in the throughput, it may be worthwhile to include this variable into the prediction rules. So, when $\hat{s}_{i}(n)$ is defined as the expected rework in measurement period $n$ at workcenter $i$, then equations 4.9 and 4.10 should be changed into:

$$\hat{y}_{i}(n) = \begin{cases} 
\hat{c}_{i}(n) - \hat{s}_{i}(n) & \text{if } l_{i}(n) \geq \hat{c}_{i}(n) \\
\min[(l_{i}(n) + \alpha[l_{i}(n)]\hat{w}_{i}(n),\hat{c}_{i}(n)] - \hat{s}_{i}(n) & \text{if } l_{i}(n) < \hat{c}_{i}(n)
\end{cases} \hspace{1cm} (4.22)$$

**Prediction model**
A last remark about the model is that the predictions of the model are point estimates; only the first moment of the predictions is considered. The disadvantage of point estimates is that there is no insight to which extent the real performance can deviate from the predicted performance. Therefore we need a range where the actual performance will probably be in. For example, suppose that the throughput prediction rule as described in Section 4.2 estimates a throughput for the next measurement period of 450 hours. However, due to differences between the expected and actual values of the available capacity and work supply, the actual performance can range from 390 hours to 510 hours. There are two ways to obtain this range. First, by including the distribution of estimation errors of model variables into the prediction rules. If the ranges for the individual model variables are known, the expected maximum and minimum values can be used to predict the maximum and minimum throughput respectively. If we assume that the variables representing the differences between the estimates and the actual values follow a specific distribution with known average and standard deviation, then these variables can be regarded as stochasts. For example, suppose that $\hat{g}_1(n)$ is normally distributed with mean zero and standard deviation 5, then the interval in which $\hat{C}_1(n)$ with a probability of 95% will be, can be defined as $[\hat{C}_1(n) - 1.96 \times 5, \hat{C}_1(n) + 1.96 \times 5]$. These values can then be entered into the throughput prediction rule to obtain the maximum and minimum expected throughput. The second way to obtain the range of the performance is to use the performance prediction errors of the past. For example, when the average and standard deviation of the throughput prediction error is 10 hours and 18 hours respectively and about normally distributed, then the range for the expected throughput for the next measurement period will be $[450 - 10 - 1.96 \times 18, 450 - 10 + 1.96 \times 18]$. For the remainder of this thesis we prefer the second way, because all kinds of errors that may have been made in the modelling phase are included in the prediction errors of the past, whereas the first method only can take into account the estimation errors of the variables that are included in the model. For example, behavior deviations from the policies, cannot be taken into account in the first method, while the second method will use past behavior deviations amongst other prediction errors.
CHAPTER 5

SIMULATION STUDY

In Chapter 4 we developed a model to predict the throughput of a production unit and production order completion times. In this chapter the quality of the prediction rules will be tested by simulation experiments for a number of different situations. In Section 5.1 we will discuss the requirements the prediction rules should meet. Then, in Section 5.2, the experimental design of the simulation study will be given. The specific prediction rules that are used in the simulation experiments are discussed in Section 5.3. In Section 5.4 the simulation results will be given, followed by the statistical testing of the hypotheses which are directly related to the requirements stated in Section 5.1. A discussion about the simulation results will follow in Section 5.5. In this discussion also an interpretation will be given regarding the meaning of these results for applying the prediction rules in practical situations.

5.1 Requirements

In the Chapters 3 we restricted the research to the short term performance of production units. In Chapter 4 we built a prediction model to calculate or predict performance targets that could serve the performance evaluation in the short term. Thereby, we assumed that actual state of the production unit (i.e. the actual work in process, the expected available capacity, and the expected work supply) to a large extent determines the performance in the short term. The prediction rules that were developed took into account this actual state, which made them state-dependent. Our implicit hypothesis so far is that we expect that these state-dependent predictions will perform better than predictions that do not take into account actual information about the state. Intuitively this seems to be true, and also in the literature some evidence can be found that supports our expectation. The literature that deals with the assignment of due dates to production orders suggests that using specific information regarding the state of the production unit and/or the production orders (e.g. total number of production orders in the production unit, total amount of work in process at the workcenters in a production order's routing) in the assignment of the due dates will give a better performance with respect to the average lateness, the spread in the lateness, and the
percentage of the production orders finishing tardy (Chang, 1994, Udo, 1994). Based on these results we also expect this positive influence on the performance of using actual information about the work in process, the expected available capacity and the expected work supply. If the prediction model we developed in the preceding chapter is built well, then this model should meet some requirements that are based on the expected increase in prediction reliability. Therefore, we will discuss in the remainder of this section to which requirements the developed prediction rules should meet.

With regard to the throughput of a production unit, we require that the state-dependent predictions of the throughput better "follow" the fluctuations in the actual throughput over time than state-independent predictions. The same holds for state-dependent predictions of production order completion times. A common measure to evaluate the quality of predictions is the standard deviation of the prediction error. Analogous to the estimation errors of the model variables (see Section 4.4), the prediction error will be defined as the difference between the predicted value and the actual value. The standard deviation of the prediction error is a measure that represents the quality or reliability of the predictions. This measure will be used for comparing the prediction results between the state-dependent predictions and the state-independent predictions. The developed state-dependent prediction rules should have a lower standard deviation of the prediction errors than the state-independent prediction rules, otherwise the adding of extra information about the actual state is of no use. Therefore, we can put forward the following two requirements.

**Requirement 1:** The standard deviation of the prediction error of the state-dependent throughput predictions should be lower than the standard deviation of the prediction error of the state-independent predictions.

**Requirement 2:** The standard deviation of the prediction error of the state-dependent production order completion time predictions should be lower than the standard deviation of the prediction error of the state-independent production order completion time predictions.

The main intention of the development of state-dependent prediction rules is to increase the reliability of the predictions, or in other words, to decrease the standard deviation of the prediction errors. This increase in the reliability however, may not lead to the introduction of systematic errors. On the long run, state-independent predictions should result in an mean prediction error of zero. Clearly, the state-dependent prediction rules should also meet this condition, otherwise systematic errors may have been introduced. From this point of view, the following two requirements can be put forward.
Requirement 3: The mean prediction error of the state-dependent throughput predictions should be equal to zero.

Requirement 4: The mean prediction error of the state-dependent production order completion time predictions should be equal to zero.

Generally, when a production order has only few operations remaining, the completion time of that production order will be in the nearer future than that of production orders that have relatively many operations remaining. This is caused by the amount of waiting time that is about proportional to the number of operations that remain. The longer the routings of the production orders, the longer the flow times will be of these production orders. This increase not only holds for the average production order flow times, but also for the standard deviation of the production order flow times. From a practical point of view, it can be said that the more operations are remaining in a production order's routing, the more disturbances and unknown factors there are that influence a production order's completion time. For these reasons, state-dependent predictions for the near future should be more reliable than predictions that are directed at a longer period of time. Consequently, the following requirement can be put forward.

Requirement 5: The standard deviation of the prediction error of the state-dependent predictions of production order completion times that are estimated to be completed in the near future should be lower than the standard deviation of the prediction error of the state-dependent production order completion times that are estimated to be completed over a longer time horizon.

In Chapter 4 we mentioned that in case there is no information about the expected available capacity or the expected work supply, one should use the average available capacity and the average amount of work supply respectively. In practice, often more information about the expected available capacity and expected work supply can be used than just taking the average amounts. For example, if one has planned preventive maintenance on several machines for the next measurement period, one should use this extra information to determine the expected available machine capacity for the next measurement period. Regarding the expected work supply, there usually is some knowledge about the characteristics of the production orders that are expected to be released to production. If extra relevant information is used by the prediction model, we require that the predictions become more reliable. To see whether this requirement is met, we restrict this extra information to knowledge about the volume of the work supply, which results in the following requirement:
Requirement 6: Adding volume information about the expected work supply should lead to lower standard deviations of throughput prediction errors and production order completion time prediction errors.

In the simulation experiments that are used to verify this requirement, the prediction results of situations in which the volume of the expected work supply is unknown will be compared with situations in which the volume of the work supply is known. In the situations where the volume of the expected work supply is unknown, the average amount of work supply is taken.

5.2 Experimental design

The simulation experiments that will be done to test the hypotheses underlying the requirements defined in the preceding section, will be based on the classical job-shop, as described by Jackson (1957). The characteristics of such a kind of a job-shop are:

- The arrival pattern of orders is characterized by a Poisson process.
- Processing times of orders on a machine are exponentially distributed.
- Orders are processed at the machines in a sequence that is independent of production order characteristics (e.g. processing times, number of workcenters in the routing) and independent of the queues at the workcenters.
- The routing matrix of the orders is symmetrical.
- There are no machine-breakdowns.

In the literature about production planning and control, the classical job-shop often is taken as starting point for analyses with respect to the performance measures we are interested in. In Chapter 3 we already concluded that the performance models that are commonly used for these analyses usually do not take into account the actual state of a production unit. The classical job-shop is considered to be the most complex type with regard to the predictability of the performance, i.e. the state as well as the performance strongly fluctuate over time. By using information about the actual state, it is expected that these fluctuations in the performance can be "followed" better compared with state-independent prediction rules. By taking this kind of job-shop as starting point for our simulation experiments, we thus link on the relevant literature.

Based on our model of a production unit as discussed in Chapter 2, three factors will be used to characterize our simulation situations:
The number of workcenters in the production unit. This measure determines the complexity of a production unit. Three situations will be distinguished, namely production units with 2, 5 and 9 workcenters. Each workcenter will consist of one machine that is operated by one operator.

The average machine capacity utilization in the production unit. This factor is set at 90% for highly utilized shops, 80% for medium utilized shops, and 60% for shops with a low average utilization. The utilization rate will be used in the simulations to determine the extent to which the actual work in process fluctuates from measurement period to measurement period in a production unit.

Information about the volume of the work supply. This factor will be used to show the impact of adding state information in the predictions. The volume of the work supply is set at "known" or "unknown". In the latter case the average work supply is taken.

In total 18 experiments for different situations should be executed. Table 5.1 gives an overview of these situations.

For all simulation experiments the First Come First Served (FCFS) sequencing rule is used. The length of the measurement period is set at 40 hours. This length is related to the average production order flow times for the different situations in such a way that not all production orders will finish within one measurement period, as we will see in the remainder of this section. This setting also makes a comparison with practical situations, where this is a very common reporting period of the performance, more easily. The average interarrival time of production orders entering the production unit, denoted by $1/\lambda$, for all experiments will be set at 1 hour. On the average about 40 production orders enter the production unit during a measurement period. Because the arrival process is Poisson and the processing times are exponentially distributed, the departure process of the production orders from the workcenters also can be characterized by a Poisson process. Therefore, the arrival process at individual workcenters can also be regarded as a Poisson process (e.g. Kleinrock, 1975). Consequently, each workcenter can be modelled as a M/M/1 queue. The average processing time, $\bar{p}$, for each experiment is set in such a way that the production unit's average utilization rate, $\bar{p}\lambda$, will be 90%, 80%, or 60%. For example, for the experiments 1, 2, 7, 8, 13, and 14, $\bar{p}$ is set at 0.9 hours.

A job-shop simulation model developed at Eindhoven University of Technology has been used as the basis model of the situations to be simulated. The simulation model is written in SIMULA. The experiments were run on a IBM PS/2 Model 72 computer (486-33). For the simulation runs, two aspects are important: the number of subruns per experiment, and the length of the starting up period.
First, the number of subruns per experiment should be determined. A rule of thumb is that the number of subruns should be between 10 and 20, as is stated by Schmeiser (1982). Therefore, we choose a number of 10 subruns to test our hypothesis. The number of measurements periods in a subrun has been set in such a way that independence would be the result. By some "trial and error" it appeared that 400 measurements periods per subrun were sufficient to get independent subrun results. This independence was tested with the Von Neumann statistic, as described by Kleijnen and Van Groenendaal (1992). The Von Neumann statistic requires a minimum of 100 observations. Therefore, we ran for the 90% utilized

<table>
<thead>
<tr>
<th>Situation number</th>
<th>Number of workcenters</th>
<th>Utilization rate</th>
<th>Volume of work supply</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>90%</td>
<td>unknown</td>
<td>9, 90, u</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>90%</td>
<td>known</td>
<td>9, 90, k</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>80%</td>
<td>unknown</td>
<td>9, 80, u</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>80%</td>
<td>known</td>
<td>9, 80, k</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>60%</td>
<td>unknown</td>
<td>9, 60, u</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>60%</td>
<td>known</td>
<td>9, 60, k</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>90%</td>
<td>unknown</td>
<td>5, 90, u</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>90%</td>
<td>known</td>
<td>5, 90, k</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>80%</td>
<td>unknown</td>
<td>5, 80, u</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>80%</td>
<td>known</td>
<td>5, 80, k</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>60%</td>
<td>unknown</td>
<td>5, 60, u</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>60%</td>
<td>known</td>
<td>5, 60, k</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>90%</td>
<td>unknown</td>
<td>2, 90, u</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>90%</td>
<td>known</td>
<td>2, 90, k</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>80%</td>
<td>unknown</td>
<td>2, 80, u</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>80%</td>
<td>known</td>
<td>2, 80, k</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>60%</td>
<td>unknown</td>
<td>2, 60, u</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>60%</td>
<td>known</td>
<td>2, 60, k</td>
</tr>
</tbody>
</table>
shops (i.e. experiments 1, 7, and 13) 100 subruns and tested independence on the average production order flow time. The results of these tests are given in Table 5.2.

Table 5.2 The Von Neumann ratio for the different situations.

<table>
<thead>
<tr>
<th>Situation</th>
<th>Von Neumann ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 workcenters</td>
<td>1.673</td>
</tr>
<tr>
<td>5 workcenters</td>
<td>1.694</td>
</tr>
<tr>
<td>9 workcenters</td>
<td>1.993</td>
</tr>
</tbody>
</table>

The 95% confidence interval of the Von Neumann ratio to accept independence is the interval [1.612 - 2.388]. From the results in Table 5.2 we can conclude that the average production order flow times per subrun of all three experiments can be considered as independent. It was assumed that the average production order flow times in the subruns in situations with a lower average capacity utilization are independent also, because production orders generally have a shorter flow time in the system and thus have less influence on each other. For that reason, no independence test is required for the other experiments.

The second aspect is the length of the starting up period. Each experiment starts with an empty system (i.e. no production orders in the production unit). Because this state usually is not very realistic for production units, we need to load the system with a number of production orders. Another reason to use a starting up period is that the state-independent predictions, which are compared with the state-dependent predictions, require that the system is in steady state. For that reason a starting up period of 15 subruns appeared to be long enough. In Figure 5.1 we graphically summarize the simulation settings.

![Simulation settings](image)

Figure 5.1 Simulation settings.

Simulation study
5.3 Prediction rules

The state-dependent prediction rules as described in Chapter 4 were included in the simulation model with some adjustments, which will be discussed in this section. Also the state-independent prediction rules will be given.

**Throughput prediction rules**

Because there is only one machine per workcenter that is continuously available (i.e. no breakdowns), the available capacity per workcenter is fixed and consequently equal to the measurement period length: \( T = 40 \). The state-dependent throughput prediction rule for a workcenter (see Chapter 4, equations 4.9 and 4.10) thus can be changed into:

\[
\hat{Y}_i(n) = \text{MIN}[\{I_i(n) + \alpha[I_i(n)]\hat{W}_i(n)\}, T]
\]  

(5.1)

where \( \alpha[I_i(n)]=0 \) if \( I_i(n) \geq T \).

The determination of \( \hat{W}_i(T) \) is done as follows:

\[
\hat{W}_i(n) = \begin{cases} 
W_i(n) & \text{, if volume work supply known} \\
\hat{W} = \lambda \beta T & \text{, if volume work supply unknown}
\end{cases}
\]  

(5.2)

The average fraction of the work supply that is expected to be processed at workcenter \( i \), \( \alpha[I_i(n)] \), can be calculated at the beginning of each simulation experiment, and thus needs to be done only once. Actually, this fraction should be calculated for each possible value of \( I_i(n) \), but because this value can take every real value, we restricted this calculation to integer values of \( I_i(n) \) only, starting from \( 0 \) up and till \( T-1=39 \) (recall that \( \alpha[I_i(n)]=0 \) for \( I_i(n) \geq T \)). As has been mentioned in Chapter 4, we need a method to determine \( \hat{I}_i(n+1) \) in order to calculate \( \alpha[I_i(n)] \). The only method we have found that can determine future amounts of work in process is an iteration method developed by de Kok (1989). Given a starting value of the work in process, this method determines the expected amount of work in process after an arbitrary number of production order arrivals. However, in de Kok's method the number of production order arrivals is not deterministically related to the period of time in which they arrive. For example, suppose that the method determines that the expected work in process will be 20 hours after 30 production order arrivals. Because there is no further information about the period of time in which these production orders arrive, it cannot be determined at which specific moment of time the expected amount of work in process will be 20 hours.
Anyhow, in our model we made as if that the amount of work in process after $\lambda T$ production order arrivals is equal to the amount of work in process at time $t+T$, thereby introducing a modeling error. In three experiments with different utilization rates it has been tested what the magnitude of this error was. It appeared that the error introduced by using de Kok's iteration method is only very small; the calculated $\alpha[I_i(n)]$'s differed only little from the real $\alpha[I_i(n)]$'s. Therefore we decided that the iteration method could be used in our simulation study. The results of these three experiments are described in Appendix A where also the description of the calculation method is given. The results of the calculations of the $\alpha[I_i(n)]$'s are shown in Figure 5.2. For ease of survey, the calculated fractions per integer value of $I_i(n)$ are connected by means of a line.

We define the state-independent prediction for the throughput at workcenter $i$ as the average throughput that will be realized in an arbitrary measurement period at that workcenter, that is calculated by:

$$\bar{Y}_i(n) = \lambda \bar{p} T$$  \hspace{1cm} (5.3)

For example, in the experiments with a utilization rate of 90%, each measurement period the state-independent throughput prediction is $0.9 \times 40 = 36$ hours. Clearly, the average throughput must equal the average work supply. Both the state-independent and the state-dependent throughput prediction will be made at the beginning of each measurement period for that specific measurement period only. Throughput predictions for subsequent measurement periods (i.e. measurement period $n+1$, $n+2$, etcetera) at the beginning of measurement period $n$ will not be made, because the
effect of using state information on the prediction reliability will decrease for two reasons. First, the actual work in process will become of less importance for the expected performance over a longer period of time. Second, the expected available capacity and the expected amount of work supply for a couple of measurement periods ahead shall be estimated by their averages, because usually no detailed information will be available for that period of time.

**Production Order Completion Time Prediction Rules**

For the state-dependent predictions of individual production order completion times, the priority rule is of great importance. Because in the simulation experiments First Come First Served is used, machine breakdowns do not occur, and the processing times are known, $\beta$ in equation 4.16 is equal to one, which for our model means that the expected amount of work in process at a workcenter at the moment a production order arrives, will be the expected waiting time for that order. The state-dependent production order completion time prediction rule as described in Chapter 4 will thus be used without any further adjustments. A state-dependent prediction of the production order completion times will be made for every production order at the time that production order enters the production unit. Further, for all production orders in the production unit a state-dependent completion time prediction will be made at the beginning of a measurement period to update the predictions for new information that is available about the work in process, the expected available capacity, and the expected work supply. So, for production orders that stay longer than one measurement period in the production unit, several completion time predictions will be made.

For the state-independent production order completion time predictions the following rule has been used:

$$\hat{D}_j = a_j + \sum_{i \in N_j} p_{j,i} + \sum_{i \in N_j} \hat{B}_{j,i} \tag{5.4}$$

where,

- $\hat{D}_j$ = the expected completion time of production order $j$
- $a_j$ = the arrival moment in the production unit of production order $j$
- $p_{j,i}$ = the processing time of production order $j$ at a workcenter $i$
- $\hat{B}_{j,i}$ = the expected waiting time of production order $j$ at workcenter $i$
- $N_j$ = the set of workcenters on which production order $j$ has to be processed

Due to our definition of the characteristics of the production units, the expected waiting time for a production order at a workcenter is the same for each workcenter and can be calculated with the following formula (e.g. Kleinrock, 1975):
\[ B = \frac{\rho}{1 - \rho} \] (5.5)

where \( \bar{B} \) denotes the average waiting time, and \( \rho \) the utilization rate, defined as \( \lambda \bar{p} \). Combining equations 5.4 and 5.5 results in a simpler expression for the expected completion time of a production order, which is as follows:

\[ \hat{D}_j = a_j + \sum_{i \in N_j} p_{j,i} + m_j \bar{B} \] (5.6)

where \( m_j \) is the number of workcenters in the routing of production order \( j \). This state-independent prediction is made only once for each production order, namely at the moment the production order enters the production unit. This state-independent prediction can thus be regarded as a completion time that is planned by a model directed at the longer term, such as an MRP system.

The expected production order completion times can lay in the near future as well as relatively far away in the future. In Section 5.1 we stated in requirement 5 that the predictions for the near future should be more reliable than the predictions farther away. In order to verify this requirement, each production order completion time prediction will be placed into a specific class that relates the predicted flow time to the length of the measurement period. In class "horizon 1" all predictions are placed of the production orders that are expected to have a flow time that is equal to or shorter than the length of one measurement period, i.e. 40 hours, while the class "horizon 2" contains all predictions of production orders having a predicted remaining flow time with a duration between one and two measurement period durations. More classes of horizons will not be considered in this simulation study. The above classification will be used for both the state-dependent and state-independent production order completion time predictions.

5.4 Simulation results

The simulation results of the state-dependent as well as state-independent predictions are presented by the means and standard deviations of the prediction errors. Recall that the prediction error is defined as the predicted value minus the actual value. Consequently, a positive prediction error for the throughput prediction means that the prediction is too high, while a positive prediction error for the production order completion time predictions implies that the prediction is too low.
In Table 5.3 the simulation results throughput predictions are given. For both the state-dependent and the state-independent situations, the mean prediction error of the ten subruns are shown as well as the average standard deviation of the prediction error of the ten subruns. The means and averages of the standard deviations of the prediction errors of the production order completion times can be found in Table 5.4. The detailed results per subrun for both types of prediction errors are given in Appendix B.

Table 5.3 Mean prediction errors and average standard deviations of prediction errors (s.d.) of throughput predictions.

<table>
<thead>
<tr>
<th>Number of workcenters</th>
<th>Utilization level</th>
<th>State-independent</th>
<th>State-dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>90% mean</td>
<td>0.06</td>
<td>3.62</td>
</tr>
<tr>
<td></td>
<td>s.d.</td>
<td>17.76</td>
<td>13.26</td>
</tr>
<tr>
<td>80% mean</td>
<td>s.d.</td>
<td>24.89</td>
<td>20.20</td>
</tr>
<tr>
<td>60% mean</td>
<td>s.d.</td>
<td>36.62</td>
<td>30.37</td>
</tr>
<tr>
<td>5</td>
<td>90% mean</td>
<td>0.26</td>
<td>2.03</td>
</tr>
<tr>
<td></td>
<td>s.d.</td>
<td>13.48</td>
<td>10.24</td>
</tr>
<tr>
<td>80% mean</td>
<td>s.d.</td>
<td>18.38</td>
<td>15.03</td>
</tr>
<tr>
<td>60% mean</td>
<td>s.d.</td>
<td>21.42</td>
<td>20.17</td>
</tr>
<tr>
<td>2</td>
<td>90% mean</td>
<td>-0.12</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>s.d.</td>
<td>7.94</td>
<td>6.13</td>
</tr>
<tr>
<td>80% mean</td>
<td>s.d.</td>
<td>10.02</td>
<td>8.51</td>
</tr>
<tr>
<td>60% mean</td>
<td>s.d.</td>
<td>9.96</td>
<td>9.51</td>
</tr>
</tbody>
</table>
Table 5.4 Mean and average standard deviations (s.d.) of production order completion time prediction errors.

<table>
<thead>
<tr>
<th>Number of work-centers</th>
<th>Utilization level</th>
<th>State-independent</th>
<th>State-dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>supply volume unknown</td>
<td>supply volume known</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hor. 1</td>
<td>Hor. 2</td>
</tr>
<tr>
<td>9</td>
<td>90%</td>
<td>mean</td>
<td>-0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>s.d.</td>
<td>19.55</td>
</tr>
<tr>
<td>80%</td>
<td></td>
<td>mean</td>
<td>-0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>s.d.</td>
<td>13.26</td>
</tr>
<tr>
<td>60%</td>
<td></td>
<td>mean</td>
<td>-0.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>s.d.</td>
<td>5.74</td>
</tr>
<tr>
<td>5</td>
<td>90%</td>
<td>mean</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>s.d.</td>
<td>19.11</td>
</tr>
<tr>
<td>80%</td>
<td></td>
<td>mean</td>
<td>-0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>s.d.</td>
<td>12.10</td>
</tr>
<tr>
<td>60%</td>
<td></td>
<td>mean</td>
<td>-0.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>s.d.</td>
<td>4.04</td>
</tr>
<tr>
<td>2</td>
<td>90%</td>
<td>mean</td>
<td>-0.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>s.d.</td>
<td>14.70</td>
</tr>
<tr>
<td>80%</td>
<td></td>
<td>mean</td>
<td>-0.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>s.d.</td>
<td>6.72</td>
</tr>
<tr>
<td>60%</td>
<td></td>
<td>mean</td>
<td>-0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>s.d.</td>
<td>2.29</td>
</tr>
</tbody>
</table>

Before verifying our requirements, we will first have a quick glance at Table 5.3 and Table 5.4 to see whether the simulation results support our expectation that state-dependent predictions are more reliable than state-independent predictions. Looking at the means of the standard deviations of both the throughput and production order completion time prediction errors.
error, it seems that the prediction reliability can be increased considerably by using state information. We can see that knowledge about the volume of the work supply has got a great influence on the prediction reliability of the throughput predictions. The effect of the use of information about the state of the production unit seems to be of less importance. The opposite holds for the prediction reliability of the production order completion time predictions. There, the greatest gain comes from the adding of work in process information of the production unit. A further, more detailed discussion about these results will be given in the next section.

The Student's t-test with a significance level of 95% has been used to test hypotheses that underlie the requirements (e.g. Kleijnen and Van Groenendaal, 1992). In Table 5.5 we present how the testing of the hypotheses is set up. The definition of the variables in the figure are as follows:

\[
\begin{align*}
x_{ij} &= \text{Either the mean prediction error or the standard deviation of the state-dependent prediction error of subrun } i \text{ from experiment } j. \\
z_{ij} &= \text{Either the mean prediction error or the standard deviation of the state-independent prediction error of subrun } i \text{ from experiment } j. \\
d_{ij} &= \text{The difference between the state-dependent result and the state-independent result: } x_{ij} - z_{ij}. \\
\mu_{ij} &= \text{The average of variable } f \text{ from experiment } j, \text{ where } f \text{ is either variable } x, \text{ variable } z, \text{ or variable } d. \\
s_{ij} &= \text{The standard deviation of variable } f \text{ from experiment } j, \text{ where } f \text{ is either variable } x, \text{ variable } z, \text{ or variable } d.
\end{align*}
\]

<table>
<thead>
<tr>
<th>Subrun</th>
<th>State-dependent</th>
<th>State-independent</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(x_{1j})</td>
<td>(z_{1j})</td>
<td>(d_{1j})</td>
</tr>
<tr>
<td>2</td>
<td>(x_{2j})</td>
<td>(z_{2j})</td>
<td>(d_{2j})</td>
</tr>
<tr>
<td>1 (\otimes)</td>
<td>(x_{10j})</td>
<td>(z_{10j})</td>
<td>(d_{10j})</td>
</tr>
</tbody>
</table>

Table 5.5 Definition of variables used for testing the hypotheses.
To use the Student's t-test, the variable $d_{ij}$ should be normally distributed. We checked this for the experiments with 90% utilization rate and 60% utilization rate and concluded that $d_{ij}$ indeed is normally distributed.

When it is hypothesized that there is a significant difference between two variables, which is the case for requirements 1, 2, 5, and 6, it is tested if $\mu_{xj} < \mu_{zj}$, which is equal to $\mu_{dj} < 0$. For this situation the hypothesis $H_0: \mu_{dj} = 0$ should be rejected (the alternative hypothesis is $H_1: \mu_{dj} < 0$). In this case the following test is done:

$$\mu_{dj} + t_{n-1}(\alpha) \cdot s_{dj} / \sqrt{n} < 0$$

where $n$ = the number of subruns
$\alpha$ = the 95% significance level
$s_{dj} = \sqrt{(s_{xj}^2 + s_{zj}^2 - 2\text{cov}(x, z))}$

For the remaining requirements (i.e. requirements 3 and 4), it will be tested if $\mu_{xj} = \mu_{zj}$, which is equal to $\mu_{dj} = 0$. In this situation $H_0: \mu_{dj} = 0$ should not be rejected. For this case the following test is done:

$$\mu_{dj} - t_{n-1}(\frac{1}{2}\alpha) \cdot s_{dj} / \sqrt{n} < 0 < \mu_{dj} + t_{n-1}(\frac{1}{2}\alpha) \cdot s_{dj} / \sqrt{n}$$

Regarding requirements 3 and 4, it should be noted that we are not going to test whether the mean prediction errors of the state-dependent predictions are equal to zero, because accidental prediction errors may occur in the simulation experiments. If accidental prediction errors occur, then this will have an impact on both the mean prediction error of the state-dependent predictions and the mean prediction error of the state-independent predictions. Therefore, it will be tested whether the means of the prediction errors of the state-dependent predictions and the state-independent predictions are equal.

**Requirements 1 and 2**

The results of the tests of the hypotheses underlying requirements 1 and 2 for the individual situations are summarized in Table 5.6. From Table 5.6 the following observations can be made. First, all state-dependent throughput predictions are significantly more reliable than the state-independent throughput predictions. Second, for some of the situations with a small number of workcenters in the production unit combined with a low utilization rates (i.e. situations [5,60,u], [5,60,k], [2,80,u], [2,80,k], [2,60,u], and [2,60,k]) there were no or not enough observations for doing statistical analyses on the standard deviations of production order completion times classified into horizon 2. In these experiments the average production order flow time is such low, that most of the orders are completed within 40 hours. For example, in the situation with 5 workcenters and a utilization rate of 60% the average production order flow time is 7.5 hours. Third, for the 90% and 80% utilized production units
the state-dependent production order completion time predictions are significantly better than the state-independent predictions, but for the lowest utilized production units the state-dependent predictions do not statistically outperform the state-independent predictions.

Table 5.6 T-test values for the hypotheses underlying requirements 1 and 2 (* = too few observations).

<table>
<thead>
<tr>
<th>Situation</th>
<th>Production order completion time, horizon 1</th>
<th>Production order completion time, horizon 2</th>
<th>Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>9, 90, u</td>
<td>-4.91 significant</td>
<td>-7.49 significant</td>
<td>-4.09</td>
</tr>
<tr>
<td>9, 90, k</td>
<td>-6.73 significant</td>
<td>-10.84 significant</td>
<td>-11.57</td>
</tr>
<tr>
<td>9, 80, u</td>
<td>-1.13 significant</td>
<td>-2.04 significant</td>
<td>-4.29</td>
</tr>
<tr>
<td>9, 80, k</td>
<td>-2.01 significant</td>
<td>-5.59 significant</td>
<td>-18.88</td>
</tr>
<tr>
<td>9, 60, u</td>
<td>0.58</td>
<td>1.39</td>
<td>-2.15</td>
</tr>
<tr>
<td>9, 60, k</td>
<td>0.61</td>
<td>1.29</td>
<td>-28.36</td>
</tr>
<tr>
<td>5, 90, u</td>
<td>-5.67 significant</td>
<td>-9.65 significant</td>
<td>-2.94</td>
</tr>
<tr>
<td>5, 90, k</td>
<td>-7.37 significant</td>
<td>-13.94 significant</td>
<td>-8.43</td>
</tr>
<tr>
<td>5, 80, u</td>
<td>-1.40 significant</td>
<td>-1.00 significant</td>
<td>-3.13</td>
</tr>
<tr>
<td>5, 80, k</td>
<td>-2.17 significant</td>
<td>-6.30 significant</td>
<td>-13.85</td>
</tr>
<tr>
<td>5, 60, u</td>
<td>0.69</td>
<td>*</td>
<td>-1.17</td>
</tr>
<tr>
<td>5, 60, k</td>
<td>0.75</td>
<td>*</td>
<td>-18.47</td>
</tr>
<tr>
<td>2, 90, u</td>
<td>-5.29 significant</td>
<td>-17.34 significant</td>
<td>-1.69</td>
</tr>
<tr>
<td>2, 90, k</td>
<td>-6.25 significant</td>
<td>-22.45 significant</td>
<td>-4.75</td>
</tr>
<tr>
<td>2, 80, u</td>
<td>-1.10 significant</td>
<td>*</td>
<td>-1.42</td>
</tr>
<tr>
<td>2, 80, k</td>
<td>-1.21 significant</td>
<td>*</td>
<td>-6.89</td>
</tr>
<tr>
<td>2, 60, u</td>
<td>0.02</td>
<td>*</td>
<td>-0.35</td>
</tr>
<tr>
<td>2, 60, k</td>
<td>0.02</td>
<td>*</td>
<td>-7.96</td>
</tr>
</tbody>
</table>

Requirements 3 and 4

In Table 5.7 the t-test values for the hypotheses underlying requirements 3 and 4, where the mean prediction errors are subject of analysis, are summarized. A first important observation of Table 5.7 is that no general indication can be given which situations lead to rejection of
the hypotheses. Only for all situations with two workcenters, the mean prediction error of the state-dependent production order completion time predictions is equal to the mean prediction error of the state-independent production order completion times. Further it can be observed that the intervals of the t-values depend on the utilization rate; the higher the utilization rate, the larger the intervals. This goes for the production order completion time predictions as well as for the throughput predictions. Because the hypotheses are rejected for many situations, it may be that modelling errors in the prediction rules have been made, resulting in systematic prediction errors. In the next section we will discuss this in more detail.

Table 5.7 T-test intervals of the hypotheses underlying requirements 3 and 4 (* = too few observations).

<table>
<thead>
<tr>
<th>Situation</th>
<th>Production order completion time, horizon 1</th>
<th>Production order completion time, horizon 2</th>
<th>Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>9, 90, u</td>
<td>-4.16, 0.72</td>
<td>-6.96, 0.95</td>
<td>2.09, 5.04 significant</td>
</tr>
<tr>
<td>9, 90, k</td>
<td>-1.67, 3.49</td>
<td>-1.14, 7.24</td>
<td>-11.15, -5.94 significant</td>
</tr>
<tr>
<td>9, 80, u</td>
<td>-1.81, -0.81 significant</td>
<td>-6.64, -4.43 significant</td>
<td>1.46, 2.79 significant</td>
</tr>
<tr>
<td>9, 80, k</td>
<td>0.07, 1.44 significant</td>
<td>-0.12, 3.36</td>
<td>-4.90, -0.66 significant</td>
</tr>
<tr>
<td>9, 60, u</td>
<td>-0.41, -0.19 significant</td>
<td>-1.44, 1.38</td>
<td>0.16, 0.47 significant</td>
</tr>
<tr>
<td>9, 60, k</td>
<td>-0.27, 0.00 significant</td>
<td>0.34, 3.55 significant</td>
<td>-0.18, 2.74</td>
</tr>
<tr>
<td>5, 90, u</td>
<td>-4.00, -0.27 significant</td>
<td>-11.08, -3.76 significant</td>
<td>0.98, 2.56 significant</td>
</tr>
<tr>
<td>5, 90, k</td>
<td>-1.98, 2.00</td>
<td>-5.20, 2.86</td>
<td>-6.46, -3.71 significant</td>
</tr>
<tr>
<td>5, 80, u</td>
<td>-1.32, -0.48 significant</td>
<td>-9.70, -5.93 significant</td>
<td>0.70, 1.53 significant</td>
</tr>
<tr>
<td>5, 80, k</td>
<td>-0.07, 1.06</td>
<td>-2.66, 0.36</td>
<td>-3.09, -0.68 significant</td>
</tr>
<tr>
<td>5, 60, u</td>
<td>-0.27, -0.10 significant</td>
<td>*</td>
<td>0.03, 0.24 significant</td>
</tr>
<tr>
<td>5, 60, k</td>
<td>-0.21, -0.03 significant</td>
<td>*</td>
<td>-0.43, 1.40</td>
</tr>
<tr>
<td>2, 90, u</td>
<td>-2.01, 1.03</td>
<td>-19.00, -0.99 significant</td>
<td>0.49, 1.16 significant</td>
</tr>
<tr>
<td>2, 90, k</td>
<td>-1.21, 1.94</td>
<td>-16.35, 3.38</td>
<td>-2.33, -1.27 significant</td>
</tr>
<tr>
<td>2, 80, u</td>
<td>-0.50, 0.07</td>
<td>*</td>
<td>0.23, 0.61 significant</td>
</tr>
<tr>
<td>2, 80, k</td>
<td>-0.25, 0.38</td>
<td>*</td>
<td>-1.08, -0.04 significant</td>
</tr>
<tr>
<td>2, 60, u</td>
<td>-0.07, 0.04</td>
<td>*</td>
<td>-0.02, 0.15</td>
</tr>
<tr>
<td>2, 60, k</td>
<td>-0.07, 0.05</td>
<td>*</td>
<td>0.00, 0.73 significant</td>
</tr>
</tbody>
</table>

Simulation study

65
Table 5.8 T-test values of the hypothesis underlying requirement 5 (* = too few observations).

<table>
<thead>
<tr>
<th>Situation</th>
<th>t-value</th>
<th>significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>9, 90, u</td>
<td>-13.27</td>
<td></td>
</tr>
<tr>
<td>9, 90, k</td>
<td>-11.78</td>
<td></td>
</tr>
<tr>
<td>9, 80, u</td>
<td>-15.58</td>
<td></td>
</tr>
<tr>
<td>9, 80, k</td>
<td>-12.98</td>
<td></td>
</tr>
<tr>
<td>9, 60, u</td>
<td>-11.90</td>
<td></td>
</tr>
<tr>
<td>9, 60, k</td>
<td>-12.08</td>
<td></td>
</tr>
<tr>
<td>5, 90, u</td>
<td>-16.34</td>
<td></td>
</tr>
<tr>
<td>5, 90, k</td>
<td>-13.91</td>
<td></td>
</tr>
<tr>
<td>5, 80, u</td>
<td>-21.81</td>
<td></td>
</tr>
<tr>
<td>5, 80, k</td>
<td>-17.74</td>
<td></td>
</tr>
<tr>
<td>5, 60, u</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>5, 60, k</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>2, 90, u</td>
<td>-19.74</td>
<td></td>
</tr>
<tr>
<td>2, 90, k</td>
<td>-15.93</td>
<td></td>
</tr>
<tr>
<td>2, 80, u</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>2, 80, k</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>2, 60, u</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>2, 60, k</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

**Requirement 5**

For requirement 5, it is verified whether predictions for a longer horizon result in larger standard deviation of the prediction error of the production order completion times than predictions for a shorter horizon. A distinction has been made in class horizon 1 for predicted production order completion times within one measurement period, and class horizon 2 for predictions between one measurement period and two measurement periods. The results of the t-tests are given in Table 5.8. It can be observed that for the situations with enough data all tests result in non-rejection of the hypothesis belonging to the fifth requirement.

**Requirement 6**

For the hypothesis underlying requirement 6 it is tested whether extra information, i.e. knowledge about the volume of the work supply for the next measurement period, would lead to lower standard deviations of the prediction errors. For this hypothesis the results of pairs of situations are compared. To test this hypothesis, the definition of the variables x and z are different from the definitions as described in the beginning of this section. Instead of representing the standard deviations of the state-dependent respectively state-independent prediction errors, x and z now represent both the standard deviations of the state-dependent predictions. Variable x thereby is used for the situations with known volume of the work supply, whereas variable z is used for the situations where the volume of the work supply is unknown. The results of the t-tests are given in Table 5.9.
Table 5.9 T-test values of the hypothesis underlying requirement 6 (* = too few observations).

<table>
<thead>
<tr>
<th>Situations</th>
<th>Production order completion time, horizon 1</th>
<th>Production order completion time, horizon 2</th>
<th>Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>9,90,u - 9,90,k</td>
<td>-1.78 significant</td>
<td>-3.14 significant</td>
<td>-7.26</td>
</tr>
<tr>
<td>9,80,u - 9,80,k</td>
<td>-0.85 significant</td>
<td>-3.42 significant</td>
<td>-14.44</td>
</tr>
<tr>
<td>9,60,u - 9,60,k</td>
<td>0.04 significant</td>
<td>0.50</td>
<td>-26.14</td>
</tr>
<tr>
<td>5,90,u - 5,90,k</td>
<td>-1.67 significant</td>
<td>-3.85 significant</td>
<td>-5.42</td>
</tr>
<tr>
<td>5,80,u - 5,80,k</td>
<td>-0.74 significant</td>
<td>-4.72 significant</td>
<td>-10.53</td>
</tr>
<tr>
<td>5,60,u - 5,60,k</td>
<td>0.06 *</td>
<td>-17.23 significant</td>
<td>4.5%</td>
</tr>
<tr>
<td>2,90,u - 2,90,k</td>
<td>-0.93 significant</td>
<td>-4.29 significant</td>
<td>-2.89</td>
</tr>
<tr>
<td>2,80,u - 2,80,k</td>
<td>-0.10 significant</td>
<td>*</td>
<td>-5.39</td>
</tr>
<tr>
<td>2,60,u - 2,60,k</td>
<td>0.01 *</td>
<td>-7.49 significant</td>
<td></td>
</tr>
</tbody>
</table>

From Table 5.9 it can be observed that for the production order completion time predictions the adding of volume information of the work supply only leads to significant more reliable predictions in situations with a moderate or high utilization rate. In all situations with a 60% utilization rate, the hypothesis is rejected. The adding of information about the volume of the work supply in the throughput predictions leads to a statistically significant improvement in the reliability of the predictions for all situations.

5.5 Discussion of results

The primary goal of the development of the state-dependent prediction rules was to get more reliable predictions (i.e. lower standard deviations of the prediction error) of the performance, which is especially important for production units evaluating their performance. From the simulation results and the statistical tests we can conclude that the constructed state-dependent predictions are more reliable than the state-independent predictions. From the situations that were statistically significant, the minimum and maximum decreases in standard deviation of the throughput prediction errors are 4.5% (situation [2,60,u]) and 90.6% (situation [9,90,k]) respectively. For the production order completion time predictions the minimum and maximum decreases for class horizon 1 are 12.0% (situation [9,80,u]) and 51.4% (situation [2,90,k]) respectively. For class horizon 2 these values are 9.1% (situation [9,80,u]) and
54.1\% (situation [2,90,k]) respectively. The extent to which the state-dependent predictions are more reliable depends on the utilization rate of the production unit and the number of workcenters in the production unit. Thereby, the following pattern can be observed: the higher the utilization rate, the less reliable the production order completion time predictions are, but the more reliable the throughput predictions are. This can be explained as follows. In situations with a high utilization rate, the work in process and the work supply during a measurement period at a workcenter are generally so high that often the maximum throughput can be realized. There is relatively little variation in the actual throughput compared to situations with a lower utilization rate. The throughput predictions also will be relatively high with little variation. Consequently, the reliability of the throughput predictions will be relatively high, which results in relatively low standard deviations of the throughput prediction errors. So, the state-dependent throughput predictions for high utilization rates can explain a larger part of the actual throughput variation than the state-dependent throughput predictions for lower utilization rates. On the other hand, completion times of individual production orders are more difficult to predict in highly utilized production units than in production units with a low utilization rate. Due to the high utilization rate, the mean and standard deviation of the actual production order flow time will be relatively high. The predictions of the subsequent waiting times in a production order's routing however will become more and more state-independent because the expected waiting times for workcenters at the end of a production order's routing go to the average waiting time. This will especially be the case for production orders with relatively many operations. The result is that the predicted production order completion times will show smaller variations compared to the actual variations in production order flow times. Because the number of operations in the routing are determined by the number of workcenters in the production system, we also have made the linkage of the observed pattern with the number of workcenters.

Another conclusion we can draw from the simulation experiments is that the adding of volume information about the work supply for the next measurement period in the predictions lead to more reliable predictions. In percentages, the minimum and maximum decrease in the standard deviation of the prediction error for the throughput predictions are respectively 50.7\% for the situation with 2 workcenters and 90\% utilization, and 88.7\% for the situation with 9 workcenters and 60\% utilization. The decrease in standard deviation of the prediction error of the production order completion time predictions range from 8.5\% for the situation with 5 workcenters and a utilization of 80\% to 14.7\% for the situation with 9 workcenters and 90\% utilization for the predictions in class horizon 1. For the predictions in class horizon 2 the decrease of the standard deviation range from 12.1\% for the situation with 9 workcenters and a utilization rate of 90\% to 16.7\% for the situation with 2 workcenters and
90% utilization. As can be noted, this positive effect on the production order completion time predictions is less strong than the effect on the throughput predictions. The reason for this is that for the production order completion time predictions the greatest effect on decrease of the standard deviation of the prediction error already has been gained by using information about the work in process, as explained in the preceding paragraph. For the reliability of the throughput predictions the greatest effect is obtained by using information about the volume of the work supply.

A final important conclusion resulting from the simulation experiments is that the mean prediction errors of both the state-dependent production order completion time predictions and the state-dependent throughput predictions are non-zero for many situations, and thus possibly systematic by nature. Because the throughput prediction for a measurement period is used in predictions of the production order completion times (see equation 4.17), this systematic error of the throughput predictions will affect the prediction errors of the production order completion times. Therefore, we will concentrate on the average throughput prediction errors as given in Table 5.4. In terms of percentage, the maximum mean prediction error is 2.7% (i.e. experiment 8, 5 workcenters, utilization 90%; \(-4.83/5\times36\)). From a practical point of view however, this error is negligible because this error amply falls in the prediction range where the actual throughput will probably be in. However, we want to dwell on this issue shortly, because this error not only affects the completion time predictions but it may also have an effect of the standard deviations of all predictions.

Generally, a distinction can be made between modeling errors and measurement errors. Modeling errors are introduced if the relationship between variables are not known exactly, or when relevant variables are not taken into account. Therefore, a model will always be a simplified view the modeler has about the system that is modeled. Measurement errors are errors that are introduced when not exactly is measured what is expected to be measured. If the sample sizes in a simulation study are large enough, there will be no measurement errors in simulation experiments, because the modeler can measure exactly what has to be measured. For example, in the measurement of the actual work supply, the actual work in process and the actual throughput in our simulation experiments no measurement errors are introduced. Therefore, the systematic prediction errors that were found in the simulation results can only be explained by the introduction of modeling errors. For the systematic prediction error for the throughput, four types of modeling errors can be given:

* Theoretically, the number of production order arrivals within a certain period of time can be infinitely large. However, in practice this number will be limited. In our calculations of the $\alpha[I_t(n)]$'s we limited this number of arrivals to 75 (see Appendix
A). Because we thereby neglect the probability of more than 75 arrivals (for the situations with a utilization rate of 90% this probability is approximately $4.2 \times 10^{-9}$), each calculated $\hat{I}_i(n+1)$ will be somewhat too small. Because of this, each $\alpha[I_i(n)]$ in fact is a little too large (see equation 4.16), which makes the throughput predictions also a little too large.

- The estimated work in process at time $T$, $\hat{I}_i(n+1)$, is the expected work in process that is measured at the moment the last production order of the work supply arrives. Almost never this arrival moment coincides with the end of a measurement period. So, the expected work in process at the end of a measurement period in fact will be a little smaller, because after the last arrival of a production order there will always be some time remaining for production, denoted by $x$ (see Figure 5.3).

![Figure 5.3](image)

Figure 5.3 Difference between a production order arrival time and the end of a measurement period.

The expected time remaining for production is directly related to the moment of arrival of the last production order. Because the production order arrivals are exponentially distributed, the time between $t+T$ and $t+T-x$ can be estimated by:

$$T \int_0^x \lambda e^{-\lambda x} dx + T e^{-\lambda x} = \frac{1}{\lambda} (1 - e^{-\lambda x})$$

(5.7)

As a result, $\hat{I}_i(n+1)$ in fact is estimated too large, which results in $\alpha[I_i(n)]$'s that are too small. The throughput predictions thus are somewhat too small.
The $\alpha[I_j(n)]$ that is used in equation 5.1, refers to the nearest integer value of the actual work in process, $I_j(n)$. Because the relationship between $I_j(n)$ and $\alpha[I_j(n)]$ is a concave function of which the first derivative descends (see Figure 5.2), this rounding leads on average to an underestimation of $\alpha[I_j(n)]$ and thus also to an underestimation of the throughput.

By using the minimum function in the throughput prediction rule (see equation 5.1) inherently an error is introduced. For all situations where $I_j(n)<T$, the actual throughput will range from $I_j(n)$ to $T$. In the prediction function however, $I_j(n)+\alpha[I_j(n)]\bar{W}_j(n)$ can exceed $T$, depending on the combination of the values for $I_j(n)$ and $\bar{W}_j(n)$. By limiting the predicted throughput to $T$, the average predicted throughput will be smaller than the average real throughput. In formula:

$$\lim_{x \to \infty} \frac{\sum_{n=1}^{x} \tilde{Y}_j(n)}{x} \leq \lim_{x \to \infty} \frac{\sum_{n=1}^{x} Y_j(n)}{x} = \bar{Y}_j$$

The predicted throughput thus will be on average smaller than the actual throughput, which results in a negative prediction error. The number of times that, due to the characteristic of the prediction function, the predicted throughput will be $T$, will be larger in situations with a high utilization rate than in situations with a low utilization rate. This explains why the mean throughput prediction error in Table 5.3 is smaller for lower utilization rates. In fact, this error is a good example of a modeling error that is made because the real relationship between $I_j(n)$, $W_j(n)$ and $Y_j(n)$ is not exactly known.

Summarizing, we can conclude that in a theoretical job-shop model the state-dependent prediction rules outperform the state-independent prediction rules with respect to the reliability of the predictions, measured in the standard deviation of the prediction error. However, due to modelling errors the state-dependent prediction rules introduce a systematic error. Further research is needed to decrease this error, but from a practical point of view the systematic prediction errors are of little importance. In the next chapter we will use state-dependent prediction rules in two real production units to see whether the quality of performance predictions can be improved by using state information in the same order of magnitude as obtained in the simulation experiments described in this chapter. Besides possible new modelling errors, we then also have to deal with measurement errors. Although we are going to pay considerable attention to the quality of the predictions in the two real life situations, the emphasis in the next chapter will be on the way in which the performance predictions are used to search for improvements in the performance of the production units.

**Simulation study**
In Chapter 3 we discussed the performance measurement and performance evaluation phases of the total process of continuous improvement. The prediction model for obtaining short term performance targets has been discussed in Chapter 4. By the simulation experiments described in Chapter 5, we demonstrated that generally it is very useful to use state information of a production unit in performance prediction rules to get more reliable performance targets for the short term compared with state-independent performance prediction rules. In this chapter, we will pay attention to the next phase in the performance management process: the performance diagnosis phase (see Figure 6.1).

From the above figure performance diagnosis seems to be an activity that should explain the difference between the actual performance and the performance targets, which is about similar to the variance analysis that can be found in the management accounting literature (Horngren et al., 1994). In this chapter we will make clear that it is more than just explaining differences. After a literature review about diagnosis presented in section 6.1, we will define performance diagnosis in a way that suits our research goals. In section 6.2, it will be
discussed how we can diagnose the performance of production units. For that purpose, we will present a framework in which performance diagnosis is linked to performance measurement, performance evaluation and the performance target setting procedure. Then, in Section 6.3, we will discuss to what extent a quantitative explanation about the realized performance can be given. The remaining deviation between the actual performance and the quantitatively explained part of the performance, is subject for a further, more qualitative diagnosis.

6.1 Definition

The term diagnosis is probably best known and used in the medical world. A dictionary definition of this term reads as "identification of disease by means of patient's symptoms ..." (Webster, 1989). If we consider the patient to be an organization or an organizational department, an analogy with organizational "illness" can be made easily. In the performance evaluation phase, the extent to which this illness manifests has been determined, namely the difference between the performance target and the actual performance.

A recent overview of the literature about diagnosis is given by Wagner (1993). He makes a distinction between causal diagnosis (cf. Smith, 1989) and situation understanding (cf. Bouwman, 1983). Causal diagnosis is triggered by an undesired situation, given the desired situation. In causal diagnosis it is relevant whether the problem source is, or can be treated as either structural or functional. A structural analysis such as "find the element (structure) that causes the problem" generally does not require an understanding of the causal model underlying the malfunction. Structural problems can often be observed directly. Functional analysis requires more detailed study of the situation, because the problem solver has to understand the system's behavior to decide which activity causes the problem. Consequently, less expertise is required for a structural analysis than a functional analysis. For situation understanding there is no desired situation. Therefore, it is less crisis oriented and more proactive. The task of the problem solver is to interpret the current situation without knowledge about the desired situation. This functional analysis can give the problem solver knowledge about the desired situation. As a result, situation understanding may lead to a causal diagnosis.

In the literature, the search process in the diagnosis phase is basically described in general terms, e.g. (Wagner, 1993):

- Perception of a stimulus which signals the undesirable situation.
• Description of the system. The purpose of this modelling step is to allow an easier search for performance deviations and to assure a complete search.
• Generation of a list of problem causes.
• Planning the search. Expert problem solvers rarely use trial-and-error, but they organize the search sequence efficiently.
• Analysis and model building. In this step the problem solver generates hypotheses, looks for clues to support or refute the hypotheses and develops a corresponding model of the problem. A successful completion of this step results in a determination of the problem cause.
• Labelling of the problem cause. The problem causes should be classified to start up appropriate actions for improvement.

These steps do not give any direction where and what exactly to search for. The experience of the problem solver plays an important role, especially at the modelling stage. By modelling the system, one automatically has to make a choice on the variables that play a role. The expertise of the problem solver will determine in which direction problem causes will be sought. It should be noted that it is not only the problem solver's expertise that plays a role, but also his individual biases.

Despite the knowledge about the steps to conduct a diagnosis such as described above, evidence suggests that managers do not carry out explicit, formal diagnosis (Wagner, 1993). At the production unit level, production unit management and personnel often claim to know what the causes are for performance deviations. Using their personal model of the production unit and their experiences with the processes that take place within the production unit, they can name numbers of explanations for the observed gap between the actual performance and the desired performance or the performance target. Examples of these explanations are: machine breakdowns, unavailability of raw materials, and unexpected operator absenteeism. These explanations usually are qualitative by nature rather than quantitative. For people that do not have the knowledge of the processes within the production units, the given explanations or causes for performance deviations sound very reasonable. However, these explanations for the deviations are never being tested quantitatively. The obvious reason for this is that there is no appropriate instrument available that can be used to test the hypothesized causes. Another reason, and probably the most important one, is that people believe in their own intuition or experience that they are right about the causes (the so-called "selective perception", Simon, 1976), which is magnified by the fact that others can easily be convinced; hardly any critical questions are asked about more detailed explanations.

Anyhow, the possessing of a diagnostic instrument is strongly recommended for the following reasons. First, it is possible that the assumed causes are not all the causes that explain the
observed performance gap. This also may include that the performance is quite insensitive for these causes. In that case, there are other causes that are not determined yet. Second, when the result of a diagnosis is that the assumed causes were the right ones, one can use that information to support the intuition. Third, due to all kinds of changes in the production unit itself or its environment, there is the danger that problem solvers only use their experience from the past to find possible causes, whereas there may have arisen new causes for performance deviations. The above reasons were the motive for developing an instrument that could be used at the operational level to explain quantitatively the actual performance of a production unit.

In our research performance diagnosis will be considered as a form of causal diagnosis. The desired situation is the performance prediction that serves as a performance target. In case there is a difference between the actual performance and the performance target, the explanation of this deviation can be regarded as an example of causal diagnosis. If there is no gap between the actual performance and the performance target, it still is useful to conduct a diagnosis to explain the achieved performance and to see whether the maximum achievable performance has been realized. In the subsequent sections of this chapter we will discuss this in more detail. Restricting diagnosis to quantitative explanations only, we will define performance diagnosis as: the process of finding causes of quantified performance deviations and explaining the achieved performance, with the requirement that the causal variables can be quantified.

An important implication of our focus on causes that are quantitative by nature, is that both variables that cannot be measured and variables that are not measured will not be part of our research. The reason is that if a certain variable cannot be measured or is not measured, one can never test quantitatively that this specific variable may have an effect in the specific production unit that is subject for research. An example of a variable that cannot be measured quantitatively is motivation of production personnel. When current attempts to quantify all kinds of motivational variables (e.g. O'Leary-Kelly and O'Leary-Kelly, 1993) become successful, this variable can and even should be taken into account in the diagnosis process, and consequently also in the performance prediction phase. The reason why a certain variable is not measured is that the assumed importance of that variable is considered to be low. For example, when there is only little scrap in production, people will not quickly consider this variable as important for the total performance of the production unit.

Based on the foregoing discussion, we split up performance diagnosis in a quantitative part and a qualitative part (see Figure 6.2). In this research only the quantitative part will be dealt
with. If the quantitative performance diagnosis does not result in a full explanation of the achieved performance, a complementary qualitative diagnosis must be conducted, especially a relatively large part of the actual performance remains unexplained. This qualitative diagnosis may lead to actions to measure new variables, which then are included in the performance prediction model as well as in the performance measurement system. In Section 6.3 a suggestion will be given how this qualitative diagnosis can be conducted.

Figure 6.2 Quantitative and qualitative performance diagnosis in the performance management process.

6.2 The performance evaluation and diagnosis method

The only model that we have found that has been developed specifically for diagnosis purposes in production units is the model developed by Wiendahl et al. (1994). Their model is based on expert learning and therefore called a knowledge based model. Based on the observed performance on different performance measures, such as total work in process, workcenter utilization, and order flow times, their model gives a list of possible causes for performance deviations. The strength of this model is that it provides a rather complete overview of the relevant performance measures that are used to evaluate a production unit's overall performance. Especially the reporting functions are worked out very well; data, diagrams and tables of the past and current performance at production unit as well as
workcenter level are given in a structured way. In that way, relationships between the performance on certain performance measures can be found more easily. However, with regard to the diagnostic part of the model there are several disadvantages that can be named. First, the model only gives a list of possible causes instead of the real causes. Thereby the quantified impact of a specific cause on the performance is not given. The second disadvantage is that some deeper, underlying causes will not be found with this model. The knowledge base consists of many occurrences of the performance coupled with a list of possible causes for deviations and a list with possible solutions. Usually this knowledge base is set up by consultant’s experiences (e.g. Armenakis et al., 1990). Although this knowledge base can be very large, it will always be limited. Third, the model does not provide a structured way to evaluate and diagnose the performance; the model shows only the actual performance together with possible causes for performance deviations. In addition, it is only focused at negative performance deviations. The strength of a diagnosis should be that also causes for better performances than expected should be included, because this knowledge should be used to maintain the current way of working.

Because of the disadvantages mentioned above, we will take a different approach towards performance diagnosis, that is quantitative by nature and more specific. Before dealing with this new approach in more detail, we will first present how performance measurement, evaluation and diagnosis can be linked to each other to get a structured way to come to performance improvements. A graphical representation of the method is given in Figure 6.3.

![Figure 6.3 The performance evaluation and diagnosis method.](image-url)
PERFORMANCE PREDICTION
At the beginning of a measurement period one starts to predict the performance for the coming measurement period. The input for this prediction consists of the actual and expected values of the variables and policies as described in Chapter 4. Based on the pre-predicted performance, one can decide to make changes in the input, for example one may try out another capacity allocation, to see if that will increase the production unit's throughput or decrease the number of orders expected to finish late. These analyses are the what-if analyses. When one is satisfied with the results of the pre-prediction, one accepts that this will be the performance target for this measurement period. One should realize that this performance target will be the maximum performance that can be achieved, given the current organization of the production unit.

PERFORMANCE MEASUREMENT AND PERFORMANCE EVALUATION
At the end of the measurement period the shop's performance of the foregoing period is measured, reported and fed back to all whom it may concern. For example, operators may get information about their efficiency, schedulers about last period's delivery reliability, and financial managers about machine utilization rates. For our model as described in Chapter 4, the production unit's actual throughput and the actual order completion dates are reported. The actual performance will now be compared with the predicted one, that served as the performance target for the evaluation. In many cases a performance deviation can be observed which has to be explained in the diagnosis phase. Again, we will emphasize that this not only holds for negative deviations, where the actual performance is worse than the performance target, but also for positive deviations. First, the causes that lead to positive deviations should be the motive for (re)defining formal policies to hold the actual performance at the higher level. Second, including these new policies in the prediction model will make the performance predictions more reliable. When no performance deviation is found, the performance diagnosis should be executed anyhow, because one still does not know the impact on the actual performance of deviations between expected values and actual values of the input variables.

PERFORMANCE DIAGNOSIS
All measured variables that may have an impact on the performance and that differ from the expected values should now be considered as possible causes for the observed performance deviation. A so-called post-prediction is made by replacing expected variable values by realized variable values. The starting point for the post-prediction is the initial state of the shop (at the beginning of the foregoing measurement period). The post-predicted performance can thus be regarded as the performance that could have been achieved, given the real order flow during the past measurement period. The post-predicted performance may differ from

Performance diagnosis
the actual performance. This deviation cannot be caused anymore by differences between the expected and actual values of the model variables, but they will be due to variations in the fulfillment of certain policies or due to model errors. These aspects will be dealt with in the next section where the diagnostic potential of the prediction model is discussed.

From our process of performance management it can be seen that the frequency of conducting a diagnosis is equal to the frequency of performance measurement and performance evaluation. Generally, the higher the reporting frequency, i.e. the shorter the feedback loop, the quicker one can conduct a diagnosis to find causes for performance deviations. At the production unit level this aspect may be even more important than on other organizational levels because of the many disturbances that may occur, each having an impact on the performance. If a more qualitative diagnosis is required, this short feedback loop demonstrates to be very important due to the short term memory of the human being (e.g. Anderson, 1990), which especially holds for people working at the production unit level because they are focused on the short term already; one can better remember special disturbances that occurred in the near past than disturbances that took place a relatively long time ago.

6.3 Explaining potential of the prediction model

In the method for performance evaluation and diagnosis three different performances play a role: the pre-predicted performance, the post-predicted performance, and the actual performance (see Figure 6.4).

![Diagram showing Actual performance, Pre-predicted performance, Post-predicted performance, Performance prediction quality, Quality of performance explanation, Estimation quality of input variables]

Figure 6.4 Differences between the actual, pre-predicted, and post-predicted performance.
The difference between the pre-predicted and actual performance reflects the quality of the prediction model. One poor prediction however does not necessarily mean that the prediction model is not good. That will depend to which extent that difference can be explained by the performance diagnosis. The difference between the pre-predicted and post-predicted performance gives knowledge about the quality of the estimates of the input variables (i.e. expected available capacity and expected work supply) of the performance prediction model. It is the difference between the maximum performance that was expected and the maximum performance that with hindsight could have been achieved. A relative large deviation between the pre-predicted performance and the post-predicted performance thus implies that the input variables were estimated poorly. Therefore, it is important to register the differences between the estimated and actual values of the input variables, as we already discussed in Section 4.4.

The deviation between the post-predicted and actual performance finally shows to which extent it is possible to explain the achieved performance with the given model variables. This will be called the explaining potential of the prediction model. The prediction model will have a high explaining potential when it can explain the realized performance for many subsequent measurement periods, or in other words, when the post-predicted performance will be close to the actual performance. A number of measurement periods will be required to exclude coincidental poor post-predictions.

From the above discussion, we can conclude that the current way of performance evaluation in fact is not the right way to conduct a performance evaluation. The performance targets that are used for the comparison with the actual performance are related to the future performance, where the future performance is an expectation based on expectations about variables that are considered to be relevant for the performance. On the contrary, a performance evaluation is related to the past, i.e. the actual performance resulting from past occurrences and decisions. So, for a fair-minded performance evaluation, the actual performance should be compared with the post-predicted performance. The post-predicted performance should thus be regarded as an ex-post performance target.

The explaining potential of a model in simulation experiments can be demonstrated relatively easy, because we can have all possible disturbances under control. In the simulation experiments that were described in Chapter 5, we only introduced variation in the work supply. We proved that the state-dependent predictions with known volume of the work supply were significantly better than the state-dependent predictions that only used the average work supply for the estimations. The first predictions refer to the post-predictions whereas the latter predictions can be regarded as the pre-predictions. In the experiments there
were no other variables that could cause performance differences between the predicted performance and the actual performance, so the explaining potential of the model was high.

In real life it will be more difficult to ascertain that the prediction model can fully explain the realized production unit performance. Three types of modelling errors can be named as cause for this:

- The absence of relevant model variables.
- Wrong interpretations.
- Simplifications.

First, modelling errors may have been introduced by excluding variables that have an important impact on the short term production unit performance, but that are not known yet or that are not regarded as important. Second, it may be that we still do not have enough knowledge about the processes that take place at the production unit level. This can introduce modelling errors by using wrong relationships between model variables, or by measuring other things that actually should be measured. For example, for the measuring of the net available operator capacity, several definitions can be used that resemble each other (Elmaghraby, 1991). Third, all kinds of simplifications are made by the modeler. An example of such a simplification is the aggregation of the work supply as discussed in Chapter 4. Another frequently observed simplification is the modelling of a certain behavior into some formal rules or policies. In practice, behavior deviations from the policies occur regularly at the production unit level. For example, a sequencing rule that should be used, will hardly be followed exactly (e.g. Stoop and Wiers, 1994). Because the observed behavior often cannot be represented by a fixed policy, the impact of behavior deviations on the production unit performance will also be difficult to quantify.

The result of all these modelling errors is that they automatically lead to variations in the quality of predictions. In practical situations, the individual impact of each of the above causes on the quality of the predictions cannot be determined, which implies that actually you never will be sure whether the model can explain the achieved performance to full extent. However, if the post-predicted performances show only small differences with the real performances over a number of measurement periods, we assume that the model has enough potential to explain the realized production unit performance.

Assuming that the model has a high explaining potential, the separate impact on the performance of some of the model variables can be determined. For analyzing the effect of one individual variable on the performance, one should use the expected value of that variable for the post-prediction instead of the actual value, while using the realized values for the
remaining variables. The impact can then be determined by comparing this adjusted post-predicted performance with the real post-predicted performance.

If the use of the prediction model in real life often leads to relatively large deviations between the post-predicted and actual performance, then all of the above mentioned modelling errors should be considered as candidate causes for these deviations. The people who are responsible for the performance of the production unit should make the decision whether the observed performance deviation justifies a further analysis. When the search process is focussed on the finding of new, relevant variables that affect the performance, this search process initially will be qualitative by nature. A method that can be used for this qualitative diagnosis is the stream analysis as described by Porras (1987). The result of this analysis should be a list of potential causes. The next step is to measure these causes and then to include the belonging variables in the model. From that moment, these causes will be part of the quantitative diagnosis.
CHAPTER 7

CASES

In Chapter 6 we have seen how performance measurement, performance evaluation, and performance diagnosis are linked together. All the steps in the process of continuous improvement have been discussed now. The simulation results of the prediction model were used to show the explaining potential of the model in a controlled environment. To illustrate the applicability of the method and prediction model in a non-controlled environment, namely in practical situations that are difficult to control because of all kinds of events that are not expected, we will now describe the results of two case studies. In Section 7.1 a description is given of the two production units. Then, in Section 7.2, we will describe the development process of the state-dependent prediction rules for the two production units. The proposed way to use this model within the performance management process will be explained in Section 7.3. Next, in Section 7.4, the results of the state-dependent throughput and production order completion time predictions will be shown and discussed. A more general discussion about the results obtained with the performance evaluation and diagnosis method will be discussed in Section 7.5. In that section we will first describe and explain the differences that were observed between the proposed way and the actual way of using the performance evaluation and diagnosis method. Next, the most important organizational adjustments and improvements by applying the performance evaluation and diagnosis method will be discussed. Finally, we will have a more detailed discussion about the prediction results.

7.1 The machining production units

The two production units considered are production departments of a Dutch aircraft manufacturer. Both are functionally organized production units that are specialized in metal removing operations, such as turning, milling, drilling, and bench working. The two production units are specialized in small products, such as special screws, pins, guards, brackets, and spacers. The quality of the products should meet the high quality standards that are used for aircraft components. After a major reorganization, the two production units were formed out of one large production unit. Since then, both production units operate more or less autonomously, although there still is interaction regarding capacity exchange and work
subcontracting. The difference between the production units mainly refers to the kind of products that have to be made. In the first production unit, which we will call production unit A, all newly developed products and all rush production orders are being processed, whereas production unit B produces the recurring part of the products. The focus in production unit A is directed at speed and flexibility, whereas production unit B concentrates on costs and efficiency.

In production unit A there are 16 workcenters. Production unit B consists of 9 workcenters. Each workcenter consists of one or more machines with the same functionality. The machines in a workcenter are technically about equal. Average processing times per operation are about 3.4 and 2.5 hours respectively. Average production order flow times are about 3.5 and 2.5 weeks for production units A and B respectively. This seems to be in contrast with a former description of the production units where it was said that production unit A is directed at speed. The reason why production unit A nevertheless is characterized by a larger average production order flow time is caused by the larger number of workcenters in the production unit compared to production unit B and the longer production order routings. Each production unit works in two 8-hour shifts, five days a week. Sometimes working overtime is used to reduce production order backlogs. Each shift has about twelve operators and one shift leader. In production unit A most operators can work at two or more workcenters, whereas this operator flexibility in production unit B is slightly lower. For the shift leaders the most important performance measure is the throughput of the production unit, defined as the total number of hours processed in a week. This throughput is measured by summing up the pre-calculated processing times of the production orders that were completed at the workcenters. This performance measure is a key measure for the financial department to make cost calculations. The feedback on the realized throughput is given weekly.

The goods flow of the total manufacturing process is planned and controlled by an MRP system. The MRP system generates operation due dates for each production order. Both production units are part of the chain that supplies products to the final assembly line. A simplified representation of the goods flow is depicted in Figure 7.1. The two machining production units considered here have two schedulers. Each of them manages a set of products. The most important performance measure for the schedulers is the delivery reliability which is defined as the percentage of the production orders that leave the shop too early or right on time compared to the operation due date of the last operation in the production unit. The main task for the schedulers therefore is to find solutions for production orders that probably will arrive tardy at the different customers (the main stream of the production orders go to the final assembly line via the painting and galvanizing department)
or production orders that are already tardy. Splitting up production orders, changing priorities, and negotiating with colleague schedulers form therefore a large part of the schedulers' work.

**Figure 7.1** Main goods flows for machining production units A and B.

In this chapter, we often shall discuss the results of both production units together, because of the many interactions between the two production units. This makes it difficult to explain the realized performance for each production unit separately. In Figure 7.2 an overview is given of the organizational structure with respect to the people directly involved with production unit and goods flow control.

**Figure 7.2** Organizational chart.
7.2 Development of the state-dependent prediction model

The main motive for the production units to participate in the research to the development of a performance evaluation and diagnosis method was that the production system manager as well as the goods flow manager of the two machining production units did not have a good notion about what the maximum performance could be, given the current organization and specific characteristics of the production units. For the throughput of the production units, the production manager and the capacity planner monthly agreed upon a throughput target, which we will call the MRP target, for the medium term (i.e. a 4 week period). This total amount is divided by four to get the throughput target for a week. However, the strong fluctuations in the real throughput in each week, demonstrated that these performance targets were not very realistic for each individual week. The fluctuations were considered to be inevitable due to the strong fluctuations in available capacity and work supply. So, for making more reliable judgements, there was a need for a more realistic performance target for the throughput.

For the delivery reliability, no explicit performance targets had been set. This was mainly due to the low production order release reliability; only 60% of the production orders that had one or more operations to be performed in the production units were released on time. So, 40% of the production orders were released to the production unit on a time later than the planned operation due date of the first operation to be performed. The delivery reliability of the production units was better than the production order release reliability, but still relatively low (about 70%). This implied that on the average the production orders recovered some of their back-log. For some production orders, a timely delivery is not always necessary, but the problem is that nobody knows which production orders will probably arrive tardy. Due to the relatively large amount of production orders that are behind schedule, the schedulers made promises about new delivery dates for the customers only when customers complained about late deliveries. The reliability of these promises often was very low, mainly because of many unexpected disturbances in the progress of the production orders and because no progress monitoring took place which could lead to an update of the promises. The promises had no formal status in the sense that they were not used to reschedule production orders, which led to an attitude of both schedulers and customers to take promises not too seriously. As long as there were no complaints from customers, no promises were made for production orders, because the schedulers said that they did not have enough time for that.

Another problem faced by the production manager and the goods flow manager was that the explanation of the realized performance in throughput and delivery reliability was only qualitative by nature and merely based on intuition. Many times, common disturbances, such
as machine breakdowns, and operator unavailability, were mentioned as causes for a
disappointing throughput and unreliable production order completions. Usually, these
explanations sounded very reasonable, but a quantitative test of the extent to which these
explanations were correct had never been made.

For getting more reliable throughput performance targets and production order completion
date promises as well as for getting a quantification of causes for performance deviations, the
use of a state-dependent prediction model, as described in Chapter 4, might be the solution.
Therefore, the production manager and goods flow control manager decided to develop and
implement such a model, specifically directed at the two production units concerned. To the
managers' opinion, the outcome of the project could be either a verification of what was
already known, which is the worst case scenario, or could give new insights in the processes
at the shop floor.

After consultations with the production system manager, the goods flow control manager, the
capacity planner, the shift leaders, and the schedulers, it was agreed that the model should
give estimates for the weekly throughput as well as estimates for the production order
completion dates (i.e. the dates on which production orders are estimated to leave the
production unit) of all production orders that already are in process. Therefore, the
measurement period for the prediction model had been set at one week, which corresponded
to the performance reporting period. The model should also be automated so that the
predictions could be made relatively quickly and could be stored for analyses. The
development of the state-dependent prediction model was done in close connection with the
people involved, because initially some of the final end-users of the model (i.e. schedulers
and shift leaders) were not convinced immediately of the need for such an automated tool.
To increase the level of acceptance, we set ourselves the task to make the model as
transparent as possible, by programming the prediction rules in a structured way so that the
physical processes at the shop floor as well as the decision processes could easily be traced
back in the prediction rules. We took some results from queuing theory as a starting point to
model the production units and included the decision behavior with respect to production
order release and production order sequence policies as was stated by the people involved.
The model thus resembled part of the simulation model as described in Chapter 5.

The workcenters in the machining production units were the fixed elements of the prediction
model. The actual state of the shop consisted of all production orders in the shop (work in
process). The following production order characteristics were used for the performance
predictions:
• Arrival date in the shop.
• Actual position in the shop.
• Remaining routing.
• Pre-calculated processing times for each operation.
• Operation due dates for each operation.
• Extra information: a mark for rush production orders, a code for the cause of a standstill.

Another variable in the model was the work supply that was expected to be released by the schedulers to the production unit in the next measurement period. Further, the available operator and machine capacity were considered to be important variables that determine the performance of the next period. Finally, two policies were considered to be part of the prediction model. The production order release policy determines the external work supply for the next week. The release policy was that there would be as much new work released (in hours) as actually had been processed in the preceding week (i.e. input-output control). This rule stabilized the total work in process level at the shop floor. The second policy is the production order sequence rule which determines the sequence in which production orders are processed at the workcenters. In the two production units, the policy was that operators first should process rush production orders, and then production orders in sequence of Operation Due Date. The Operation Due Dates were calculated by the MRP system using standard production order flow times per workcenter. According to the people involved, especially the strong fluctuations in work supply per workcenter and the available operator capacity caused the differences in the performance from week to week. We verified this assumption by some sample measurements and concluded that these statements were true. Therefore, we expected that a state-dependent model could contribute greatly to the reliability of performance predictions and explanations in the short term.

The general prediction rules developed in Chapter 4 had been modified to capture the specific characteristics and policies in the two production units. Many discussions were held with the final end-users of the model (shift leaders and schedulers) in order to modify the general prediction rules to make them realistic with respect to the characteristics of the production units. For verification of the assumptions underlying the prediction rules, the production manager and the goods flow manager were questioned regularly. They were also asked for suggestions and help to facilitate the process of implementation. The definitive prediction rules were programmed in a software tool. A description of the specific prediction rules is given in Appendix C.
7.3 The designed performance evaluation and diagnosis method

In conformance with the method described in Chapter 6, the following method to evaluate and diagnose the performance of the two machining production units was introduced. Each morning on the first working day of the week the capacity planner makes a download of the actual state (work in process) of both production units and a list of the completed operations in the production units of the foregoing week. These data are given to the shift leaders who read the downloaded data in the software programma and then enter the expected capacity for the current week. For the machine capacity they estimate the available machine hours per workcenter, taking into account for example preventive maintenance. For the operator capacity they first have to estimate the available operator capacity and then they have to make an allocation decision. Regarding this allocation decision, the two shift leaders can discuss whether some operators should be exchanged between the production units. Also the schedulers may ask to allocate extra operator capacity to certain workcenters to decrease waiting times or to give priority to some specific production orders. Next, a linkage should be made by the shift leaders between operator and machine capacity (for example in case an operator can handle two machines in a workcenter at the same time) by determining which capacity type will be the constraining capacity. Further, the expected efficiency rate, defined as throughput divided by available capacity, for the whole production unit has to be entered. The average efficiency of the last 10 measurement periods is taken as default value. The actual efficiency rate is measured already by one of the available performance measurement systems, and is used at the weekly workmeetings for performance evaluation purposes. When the capacity input has been entered, they can run the model to make a pre-prediction. Based on the throughput of the foregoing week, the model first calculates which production orders are expected to be released for the next week assuming that the release policy is to release production orders in sequence of operation due date till the sum of the operation times of the first workcenter in the routing equals the production unit's throughput of last week. Table 7.1 shows an example of such a screen. In the table we can see that for each workcenter the actual work in process is given in hours in the second column named "queue length". In the third column the result of the pre-prediction is given. As has been mentioned in Appendix C, the pre-predicted throughput is amongst other variables based on the expected efficiency level, which in this example was 80%. For example, for workcenter 4 where the actual work in process exceeds the available capacity, the pre-predicted throughput is calculated by multiplying the efficiency level and the available capacity: 0.8 * 40 = 32 hours. Analyzing the pre-predicted throughput together with a list of production orders that probably will finish tardy, the shift leaders can decide, in close consultation with the schedulers, to make a different capacity allocation decision. For example, given the results of Table 7.1 it seems

Cases
Table 7.1 Example of throughput screen (machining production unit B, all values in hours).

<table>
<thead>
<tr>
<th>Workcenter</th>
<th>Queue length</th>
<th>Pre-predicted throughput</th>
<th>Expected capacity</th>
<th>Actual throughput</th>
<th>Post-predicted throughput</th>
<th>Actual capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pre-prediction efficiency 80%</td>
<td></td>
<td></td>
<td>Post-prediction efficiency 80%</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0</td>
<td>19.0</td>
<td>24</td>
<td>3.6</td>
<td>19.0</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>83.0</td>
<td>118.2</td>
<td>160</td>
<td>93.9</td>
<td>122.0</td>
<td>152</td>
</tr>
<tr>
<td>3</td>
<td>227.2</td>
<td>0.0</td>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>59.4</td>
<td>32.0</td>
<td>40</td>
<td>32.0</td>
<td>32.0</td>
<td>40</td>
</tr>
<tr>
<td>5</td>
<td>6.0</td>
<td>19.0</td>
<td>24</td>
<td>12.3</td>
<td>19.0</td>
<td>24</td>
</tr>
<tr>
<td>6</td>
<td>145.7</td>
<td>38.0</td>
<td>48</td>
<td>49.6</td>
<td>45.0</td>
<td>56</td>
</tr>
<tr>
<td>7</td>
<td>69.6</td>
<td>32.0</td>
<td>40</td>
<td>24.0</td>
<td>32.0</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>275.5</td>
<td>38.0</td>
<td>48</td>
<td>31.2</td>
<td>38.0</td>
<td>48</td>
</tr>
<tr>
<td>9</td>
<td>46.0</td>
<td>72.3</td>
<td>128</td>
<td>47.9</td>
<td>90.0</td>
<td>112</td>
</tr>
<tr>
<td>total</td>
<td>912.4</td>
<td>368.5</td>
<td>512</td>
<td>301.2</td>
<td>397.0</td>
<td>496</td>
</tr>
</tbody>
</table>

It is wise to move one operator (40 hours) from workcenter 2, where the amount of available capacity is much higher than the expected throughput, to workcenter 8, where the expected capacity utilization is at its maximum (based on the assumption that there is an efficiency of the capacity of 80%) and the work in process level is relatively high. For workcenter 3, that week there were no operators available with the right skills. This new allocation decision should be entered in the model, and then a new pre-prediction can be made. This process of "what-if analyses" continuous until shift leaders as well as schedulers are satisfied with the allocation with respect to the expected throughput and the production orders expected to finish tardy. The latest pre-predicted performance now serves as a short term performance target.

The next step is the evaluation and explanation of last week's performance. For that purpose shift leaders enter last week's actual available operator and machine capacity. The available operator capacity at each workcenter is not always easy to determine when operators are reallocated for relatively short time periods; there is no formal registration of the allocated operator capacity at workcenter level. Having entered the actual capacity, one can make the
post-prediction. This is the performance that could have been achieved, knowing the actual available capacity and the actual work supply. The actual work supply has been calculated by the model based on a comparison of the actual state of the production unit and the state at the beginning of the foregoing period. The difference between post-predicted and actual performance is now subject to discussion. In the example of Table 7.1 one first should look at the difference between the actual throughput (fifth column) and the post-predicted throughput (sixth column) of the whole production unit (301.2 versus 397.0 hours) and then focus on individual workcenters in order to determine at which workcenter there is a serious deviation (in this case workcenters 2 and 9). Sometimes shift leaders can explain these differences, but often they can not. In the latter case they can go to the workcenter and ask the operator who is working there for an explanation. The given explanations may lead to actions for performance improvement. In cases where the actual throughput is higher than the post-predicted one, this often is caused by the timing of the reporting of the completion date; when a production order has started the operation in the foregoing week, and the actual completion date falls into the current week, the total processing time of this production order is reported for the current week independently of the amount of hours actually spent on this production order in the current week.

After the post-prediction, a copy of the results is given to the schedulers. During the week they can use the data for two purposes. First, they can use the data for taking measures regarding production orders that are expected to finish tardy. The software model automatically generates a list of production orders that are expected to arrive tardy by comparing the pre-predicted completion time with the planned due date of the latest operation in the shop. A scheduler can start up three possible actions to try to let the production orders arrive on time:

- They can ask the shift leaders to allocate more capacity at the workcenters the production orders have to go through in their remaining routing.
- They can decide to give a production order a priority mark so that the production order can be considered as a rush production order.
- They can change the production order due date. This usually is only done after deliberations with the customer.

The second purpose to use the data is to answer questions of customers when production orders are expected to arrive. The pre-predicted completion time plus some fixed transportation time is then promised to be the delivery date for the customer. The reliability of this promise can be obtained by looking at the standard deviation of the prediction error of past production order completion time predictions, or by looking at the other performance
measures showing prediction errors, such as the minimum and maximum error. An example of such a screen is shown in Table 7.2.

Table 7.2 Example of screen showing the production order completion time prediction quality.

<table>
<thead>
<tr>
<th>Completion time prediction errors week 8 - week 14</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>horizon</td>
<td>#</td>
<td>Average error</td>
<td>Standard deviation</td>
<td>Relative error</td>
</tr>
<tr>
<td>Errors pre-predictions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-1</td>
<td>260</td>
<td>-5.10</td>
<td>7.31</td>
<td>0.34</td>
</tr>
<tr>
<td>1-2</td>
<td>140</td>
<td>-4.14</td>
<td>9.21</td>
<td>0.30</td>
</tr>
<tr>
<td>2-3</td>
<td>86</td>
<td>1.24</td>
<td>7.37</td>
<td>0.14</td>
</tr>
<tr>
<td>3-4</td>
<td>72</td>
<td>3.63</td>
<td>7.55</td>
<td>0.26</td>
</tr>
<tr>
<td>&gt;4</td>
<td>27</td>
<td>5.63</td>
<td>11.26</td>
<td>1.15</td>
</tr>
<tr>
<td>Errors post-predictions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-1</td>
<td>311</td>
<td>-2.49</td>
<td>4.88</td>
<td>0.11</td>
</tr>
<tr>
<td>1-2</td>
<td>203</td>
<td>-0.46</td>
<td>7.92</td>
<td>0.05</td>
</tr>
<tr>
<td>2-3</td>
<td>87</td>
<td>0.86</td>
<td>9.01</td>
<td>0.15</td>
</tr>
<tr>
<td>3-4</td>
<td>51</td>
<td>3.45</td>
<td>6.59</td>
<td>0.24</td>
</tr>
<tr>
<td>&gt;4</td>
<td>27</td>
<td>6.59</td>
<td>13.49</td>
<td>0.74</td>
</tr>
<tr>
<td>Errors scheduler predictions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-1</td>
<td>106</td>
<td>-3.33</td>
<td>8.36</td>
<td>0.08</td>
</tr>
<tr>
<td>1-2</td>
<td>144</td>
<td>-0.28</td>
<td>7.44</td>
<td>0.32</td>
</tr>
<tr>
<td>2-3</td>
<td>101</td>
<td>2.72</td>
<td>8.18</td>
<td>0.26</td>
</tr>
<tr>
<td>3-4</td>
<td>19</td>
<td>-3.74</td>
<td>11.62</td>
<td>0.03</td>
</tr>
<tr>
<td>&gt;4</td>
<td>11</td>
<td>-3.91</td>
<td>17.56</td>
<td>0.35</td>
</tr>
</tbody>
</table>

For example, the prediction errors of the pre-predictions can be read as follows. For production orders that are predicted to leave the production unit within one week, the average prediction error for week 8 up and till 14 was 5.1 days, based on 260 predictions. The minus sign indicates that the production orders in reality on average finish 5.1 day later than predicted. The most extreme prediction error was that a production order finished 35 days
later than predicted (depicted in the column "minimum error"). Other measures that give an indication of the reliability of the prediction errors are the standard deviation, the relative error (defined as the ratio of the actual prediction error and the actual flow time), and the absolute prediction error (defined as the absolute difference between the predicted and actual production order flow time).

7.4 Prediction results

**THROUGHPUT**

As can be observed from Figure 7.3, the predictions with the state-dependent model follow the fluctuations in the real performance quite well for both production units. In the weeks 51, 52 and 1, no predictions were made because of the Christmas holiday.

![Figure 7.3 Throughput predictions and actual throughput per week.](image)

In week 9 an adjustment of the model had been made. The predictions were first based on an operator efficiency target level of 80% stated by the production manager, based on time losses due to time for personal care, extra time needed for setups, learning time for new
products, and time required for teaching new colleagues. Also idle time is included in the efficiency, although this factor cannot be determined easily. The real operator efficiency (measured by the actual throughput divided by the actual available operator capacity) however deviated per production unit from this target. To make more reliable predictions it was decided to make also the efficiency an input variable; for every pre-prediction shift leaders had to enter the expected efficiency-level, as mentioned in Section 7.3. The average efficiency over the last 10 weeks thereby was given as a suggestion. For the post-predictions the actual efficiency was used. The efficiency differences are an important cause for differences between the predicted and real throughput, especially at workcenter level. The averages and standard deviations of the throughput prediction errors are shown in Figure 7.4. These results are based on 33 predictions.

![Figure 7.4 Throughput prediction errors.](image)

It can be observed that the predicted throughput in production unit A on the average is higher than the real performance (about 50 hours, which is about 10% of the actual throughput), whereas the opposite holds for production unit B (about 30 hours, or about 6% of the actual throughput). According to the production system manager and the shift leaders the main cause for this difference between the production units is that the operators in production unit A are more skilled than the operators in department B. The reliability of the pre-predictions in both production units is higher than the planned throughput by MRP; for production unit A the standard deviation of the throughput prediction error is about 40% lower than the standard deviation of the planned throughputs by MRP, whereas this difference is about 20% for
production unit B. Further, we can observe that the model indeed can explain to some extent the realized throughput since the standard deviation of the prediction errors of the post-predictions is lower than the standard deviation of the prediction errors of the pre-predictions.

**Production Order Completion Times**

To show the quality of the predictions of production order completion times, the prediction errors for all production orders that were completed in the production units have been determined by taking the difference between predicted completion date and real completion date. For each production unit, the averages as well as standard deviations of the prediction errors over a number of 33 weeks are shown in Figure 7.5. The prediction errors are shown of i) the pre-predictions, ii) the rule of thumb that the schedulers currently use to predict a production order’s completion time (i.e. a standard flow time per workcenter), iii) the post-predictions, and iv) the completion dates as planned by MRP. Just like the predictions in the simulation studies, a distinction is made in the number of weeks (the horizon) a production order completion time prediction is made for.

![Figure 7.5 Prediction errors of production order completion date predictions.](image)

From Figure 7.5 it can be observed that almost all prediction methods on the average are too optimistic; the actual completion times are over due in comparison with the estimated or planned completion times. The reliability of the predictions, represented by the standard deviation of prediction error, is high for the post-predictions and low for the pre-predictions.
deviation of the prediction errors, of both the state-dependent model and the scheduler method is much better than the MRP predictions; for both production units the standard deviation of the prediction errors is about 3 to 4 time lower than the standard deviation of the prediction error by using the planned completion dates of MRP. Further, it seems that the quality of the predictions of the model and the scheduler are not influenced by the prediction horizon of a production order. Surprisingly, the state-dependent predictions perform about the same as the predictions of the schedulers. In the next section some possible explanations will be given of the prediction errors of both the production order completion date predictions and the throughput predictions.

7.5 Evaluation of results

In Section 7.3 we designed a way to use the performance evaluation and diagnosis method by the shift leaders and the schedulers. The actual way they used the method differed somewhat from the designed way. The actual way they used the method and the reasons for the difference will be explained in this section. Further, a merely qualitative evaluation will be given of the results that were obtained with the new, structured way of performance evaluation and diagnosis. Finally, an evaluation will be given about the quality of the predictions.

The actual use of the performance evaluation and diagnosis method

The aim of the designed way to use the performance evaluation and diagnosis method was to let schedulers and shift leaders discuss about the predicted and achieved performance from week to week. The idea was that more communication would lead to a better understanding of each other's problems and that one could use each other's knowledge to improve the production units' throughput and delivery reliability. However, the involvement of the schedulers decreased in such a way that they only get the backups of the predictions. In other words, the schedulers were not involved with the what-if analyses to find an acceptable capacity allocation, did rarely use the list of production orders expected to finish tardy, and did rarely use the pre-predicted production order completion times to make promises to customers. There are five important causes for the diminishing involvement of the schedulers.

First, due to the many informal capacity and production order exchanges between the two production units, the progress of individual production orders was difficult to monitor, because the actual position of a production order might be different from the administrative position. This behavior would make some of the starting points for the predictions unreliable
(i.e. actual state and remaining workcenters in the routing), and consequently also the prediction themselves. This reason was valid until these informal production order exchanges were not allowed anymore by the production manager.

Second, many production order routings contained an unbalanced mix of workcenters distributed over both production units. Together with the splitting up into two distinct production units, the product routings had to be allocated over the two production units. But due to a lack of time in the reorganization phase, the adjustments of the routings were not done satisfactorily which resulted in many production orders that should go from production unit A to B, back to A and so on. Because the production order completion time predictions are only aimed at the particular production unit a production order is currently in, the use of the predictions was very limited because the schedulers wanted to have a completion time prediction based on the flow in the two production units.

Third, schedulers should often make promises of production order deliveries to their colleagues of the final assembly line which could not be supported by the prediction model that is applicable at production unit level only. Most of the production orders have some operations in the painting and galvanizing department after completion in the machining production units, before they arrive at the final assembly line. The schedulers were responsible for making promises for this total stage, although they had very little insight in the processes and the expected flow times at the painting and galvanizing department. Therefore, schedulers still had to consult the main computer system to see which operations followed the operations in the machining production units and then to estimate the flow time for these remaining operations. The fraction of the flow time in the machining production units compared to the total flow time to the final assembly line, including some transportation time, is relatively low, which makes it not attractive for the schedulers to use the state-dependent predictions apart from the main computer system.

Fourth, schedulers did not feel the need to communicate with shift leaders about the predicted and actual performance of the throughput. Their main task is to monitor and reschedule individual production orders and not to schedule the complete production order flow. Consequently, they were mainly focused on customer’s complaints about late deliveries as said they had no time for discussing operational problems with shift leaders regarding the throughput. Only if the individual progress of a production order could be influenced by a specific capacity allocation, the schedulers contacted the shift leaders.
Fifth, the schedulers had the feeling that a good delivery reliability could to a large extent be achieved by a good capacity allocation. Because the task of capacity allocation is a part of the shift leaders' function and not part of the schedulers' function, the schedulers didn't feel responsible for giving advice to the shift leaders about the allocation of capacity. As a consequence, they didn't feel the need to be present at the time the input of the performance prediction model was entered by the shift leaders.

RESULTS OBTAINED BY APPLYING THE PERFORMANCE EVALUATION AND DIAGNOSIS METHOD

The use of the performance evaluation and diagnosis method contributed mainly to organizational adjustments and improvements that are related to production unit control issues. However, many of these improvements will also have an effect on goods flow control. The following organizational improvements can be mentioned.

First, the use of the prediction model as a what-if instrument for capacity allocation decisions gave shift leaders more insight in relationships between available capacity, the amount of work in process, work supply, and throughput. This insight led to better allocation decisions resulting in a higher production unit's throughput with the same operator capacity available. If the higher throughput level can be maintained, then the average production order flow time will decrease which is beneficial for goods flow control purposes.

Second, the informal capacity and production order exchanges between the production units have been abandoned by the production manager to be able to improve the evaluation of the individual production units. Although this policy in practice still is more an intention than an action, these exchanges are expected to be abandoned in the future. The production units will then become more self-contained.

Third, differences between pre-calculated and actual processing times could not be used anymore by shift leaders to increase a production unit's throughput. In the past one had the possibility to ask for more throughput if the actual processing time of a production order's operation considerably exceeded the pre-calculated processing time. This extra throughput actually has never been used in the reporting of the actual throughput, but always has been reported separately. For some weeks this extra throughput was about 20% of the total production unit's throughput. Because it was recognized that the processing times should be more realistic to make reliable predictions, the production system manager and the goods flow manager decided to stop making this distinction. A project at the work preparation department had then been started up to make more reliable processing times.
Fourth, one could now identify the workcenters where throughput problems occur regularly, i.e. where it was difficult to predict the throughput (it should be noted that a bad throughput prediction for a workcenter also influences the quality of the production order completion date predictions). More importantly, one could now quantify these problems, whereas in the past only intuitive and qualitative reasons were used to explain the actual performance. This kind of detailed information could be obtained by a screen were the quality of the past post-predictions is presented. An example of such a screen is presented in Table 7.3.

Table 7.3 Example of a screen showing the prediction quality of the post-predicted throughput.

<table>
<thead>
<tr>
<th>workcenter</th>
<th>#</th>
<th>Average error</th>
<th>Standard dev. error</th>
<th>Relative error</th>
<th>Absolute error</th>
<th>Minimum error</th>
<th>Maximum error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>5.91</td>
<td>7.24</td>
<td>3.34</td>
<td>7.60</td>
<td>-3.40</td>
<td>14.00</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>6.03</td>
<td>38.59</td>
<td>0.10</td>
<td>21.20</td>
<td>-32.10</td>
<td>89.10</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>-0.90</td>
<td>7.74</td>
<td>0.18</td>
<td>6.56</td>
<td>-10.00</td>
<td>8.50</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>-5.86</td>
<td>8.65</td>
<td>0.04</td>
<td>8.00</td>
<td>-18.10</td>
<td>2.90</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>5.57</td>
<td>10.35</td>
<td>0.80</td>
<td>9.26</td>
<td>-6.80</td>
<td>18.40</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>11.81</td>
<td>13.50</td>
<td>0.21</td>
<td>16.39</td>
<td>-16.00</td>
<td>24.10</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>-13.30</td>
<td>17.44</td>
<td>0.19</td>
<td>14.21</td>
<td>-47.50</td>
<td>3.20</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>3.31</td>
<td>38.59</td>
<td>0.13</td>
<td>29.69</td>
<td>-70.20</td>
<td>45.70</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>-15.03</td>
<td>32.29</td>
<td>0.17</td>
<td>19.57</td>
<td>-84.40</td>
<td>12.00</td>
</tr>
<tr>
<td>total</td>
<td>7</td>
<td>-2.44</td>
<td>29.52</td>
<td>0.00</td>
<td>16.84</td>
<td>-51.20</td>
<td>50.40</td>
</tr>
</tbody>
</table>

It can be observed that, for example, for workcenter 9 the predicted throughput on the average is 15.03 hours lower than the actual one. Also the standard deviation of this prediction error is relatively high (i.e. 32.29 hours), which quantifies the difficulties with explaining the realized throughput for this workcenter. Based on further analyses the following causes for this kind of problems at workcenter level could be identified: setup times at some workcenters generally take more time than was pre-calculated, technical problems with some machines due to the old age, incomplete product drawings and specifications, difficulties with the use of certain tools, and differences between individual operators about the operator-efficiency or skills.

Cases
Fifth, the throughput performance targets for each next week were accepted to be more realistic than the available throughput targets that had been calculated for a period of four weeks. Further, this target corresponds better to the performance reporting period which enables fast feedback and thus quick corrective actions. An important observation was that the throughput performance target obtained from the pre-predictions, was not used by the production manager as a target to judge individual shift leaders, but more to open the discussion about how this performance target could be reached. In other words, it was asked which preparative activities (for example checking the availability of drawings or production tools) should be taken to make it possible to reach the pre-predicted throughput. For some workcenters this has led to a new structuring of the work, which resulted in a higher throughput. For example, at the benchworking workcenter, one operator of the four usually working there, was demanded to do the preparative actions, which resulted in a significant higher throughput at this workcenter.

**Evaluation of the prediction results**

Given the prediction results as shown in Section 7.5 and the discussion above, it seems that there are still possibilities to improve the predictions. We note however that possible prediction improvements were not demanded from the users of the prediction model, because the model generates predictions that are good enough so far to find evidences for performance improvements. The first matter of importance is the use of the performance evaluation and diagnosis method and not the specific prediction results. Although the prediction results themselves are considered to be good enough, improvements in these predictions should always be striven for. Therefore we try to explain the sometimes disappointing prediction results in the remainder of this section. In Chapter 6 we already clarified that for conducting quantitative diagnoses we only take into account variables that are currently being measured, either automatically or manually. The quantitative diagnosis seemed to be incomplete because there were systematic errors in the predictions. In the subsequent qualitative diagnosis we found together with the people involved the following five possible explanations for which further quantitative analyses are required to see whether they are true or not.

First, the rework was considered to be a variable that may have an impact on the predictions. Rework consists of products rejected by the quality inspection that have to be processed again. For this rework no extra processing times are calculated and as a result this is no part of the reported throughput. For the throughput predictions the exclusion of rework means that they should have been too high compared to the actual throughput. In Figure 7.4 we can see that this can only hold for production unit B, because of the positive average throughput prediction errors.
Second, the lack of following the production order sequencing policy may be a cause for the sometimes disappointing production order completion date predictions. From some sample observations of the actual sequence we concluded that the Operation Due Date policy was not followed strictly. Some reasons why operators show a different behavior with respect to the sequencing are: individual specialisms which make operators chose production orders they prefer, avoiding production orders for which the setups are difficult, and the filling of the remaining shift time with production orders that have processing times not larger than this remaining shift time. The problem with this issue is that one can not show how this different behavior affects the actual performance. The negative effect of this deviating behavior on the quality of the predictions holds for both pre-predictions and post-predictions.

A third explanation is that the actual work supply can deviate from the expected work supply, not in total volume but in the specific production orders that are released. In some weeks about 50% of the production orders actually released, were different from the list of production orders that were expected to be released. Part of this is caused by rush production orders that are released to production. These differences in work supply have a negative influence on the pre-predictions of the production order completion dates, and to a lesser degree also on the pre-predictions of the throughput.

Fourth, it happened that production orders during their presence in a production unit encountered some standstills due to the unavailability of raw material, drawings, or tools. In that case production orders got a mark for this standstill. For production orders with a mark for a standstill, no predictions were made because the duration of the standstill often could not be determined or estimated. However, the pre-predictions that already had been made in earlier measurement periods still were used in the reporting of the quality of the predictions. The standstills generally will lead to negative prediction errors of the production order completion dates (i.e. the actual completion date will be later than the pre-predicted production order completion date).

A last explanation is that the pre-predicted performance may lead to a different behavior than expected. For example, the list of production orders that are expected to finish tardy can be used by the schedulers to change priorities or to negotiate with the customers about new delivery dates. These actions may result in worse pre-predictions of the production order completion dates. However, because the schedulers rarely used this list of production orders, as has been mentioned in the beginning of this section, this effect may be expected to be small.
A suggestion for further research is to test the developed prediction rules by simulation experiments. In this way, it can be researched if any modeling errors have been made, resulting in systematic prediction errors. The results in Figure 7.4 and Figure 7.5 show that the average prediction errors are relatively large, but they all fall in the range the actual throughput and production order completion dates will probably be in. The simulation experiments can also be used to evaluate the observed values of the standard deviations of the prediction errors. In that way, it can be judged whether the observed standard deviations of the prediction errors are inherent to the dynamic behavior of the processes that take place in a production unit, or if they are (partly) caused by modeling errors or certain modeling simplifications. For example, by means of a simple simulation experiment (1 workcenter, Poisson arrivals, exponential processing times, First Come First Served discipline) we tested the effect of using the average work in process instead of the iteration method of de Kok (1989) for future amounts of work in process. In this experiment, with 6000 throughput predictions, the simpler method resulted in an increase of almost 25% of the standard deviation of the throughput prediction. This example makes clear that further research is highly recommended.
CHAPTER 8

TOWARDS IMPLEMENTATION

In the preceding chapters we developed a state-dependent prediction model and tested it by simulation experiments. It was shown that the state-dependent predictions generally performed better than the state-independent predictions with regard to the prediction reliability. Further, we reported on the successful application of the evaluation and diagnosis method in two production units. Based on these theoretical and empirical results we now will make some generalizations with respect to the implementation and use of the method in practical situations. In Section 8.1 we will first describe which efforts have to be made to implement the method and which results can be expected when the method is being used. Then, in Section 8.2, it will be discussed which conditions should be fulfilled to smoothly implement the performance evaluation and diagnosis method.

8.1 Efforts and expected benefits

Before the decision has been made to implement and use the performance evaluation and diagnosis method generally a trade-off should to be made between the required efforts and the expected benefits. With respect to the required efforts the modelling phase is the most critical phase. In particular the development of the prediction rules may take considerable time, because the processes that take place at the shop floor should be understood very well by the person (or team) that builds the model. The model builder should be someone with knowledge about queuing theory and relationships between the main production unit performance variables. Because this person usually won’t be someone working at the operational level, it is very important that the modelling is done in close connection with the practitioners the model is built for, because they are the ones that have the best knowledge about the processes that take place at the shop floor. In this way the different expertises are combined which increases the chance of a successful implementation. Another important reason to let participate the people the model is built for in the modelling phase is to increase the acceptance of the model.
Regarding the model, we expect that the actual state of the production unit, the available capacity, and the work supply will be the main variables that influence production units' performance. For the two production units described in Chapter 7 it was shown that this expectation is confirmed. It is mainly the complexity of the release and sequencing policy that determines the complexity of the prediction rules. The higher the complexity, the more time will be required to develop the prediction rules. Generally, each production unit requires its own prediction rules so that all specific relevant characteristics can be part of the model. Therefore, the use of standard prediction rules should be avoided, because they will have a more general character and thus also independent of the production unit for which they should be used. The modelling of the prediction rules may be supported by simulation experiments to test the quality of the prediction rules. This is recommended especially for complex prediction rules.

Besides the modelling effort, the development of a software tool that contains the prediction rules and that gives overview and graphs of the results may also take some considerable time and capacity. Important issues for the development are the determination of the performance measures, the construction of the screens, and the response times of the model. The computer programmer of the model should consult the final end-users of the model many times during this technical development phase. This participation will also increase the speed with which the model will be accepted.

The above efforts are non-recurring efforts. The only recurring effort is the amount of time that has to be spent each measurement period with the model. This amount of time consists of entering the input data, analyzing the results, and discussing the results. When these activities are included into the regular workmeetings, there is almost no extra time involved. The only thing is that these discussions about the work should take place at the beginning of each measurement period, due to the concentration of the prediction model on distinct measurement periods that correspond with the performance reporting periods.

The specific results that can be obtained with the use of the performance evaluation and diagnosis method will differ from production unit to production unit. The strength of the method is that it due to its formalized way of working people are forced to think about both the expected performance and achieved performance every measurement period again. Because the performance targets that are obtained from the prediction model are more objective than other targets that are set by individuals, a more useful and realistic discussion can be held between people responsible for the production unit control and people controlling the goods flows. The communication can lead to a better understanding of each others
problems and can help to avoid future problems. Further, quick responses are possible due
to the short term focus of the model. Fast feedback enables people to react quickly on
problems. The model will also stimulate a proactive decision behavior by giving realistic
performance estimations. For example, in the two production units of the preceding chapter
it was said that the pre-predicted production order completion dates could be used to
reschedule the production orders or to change the priorities of them. The disadvantage of this
proactive decision behavior is that systematic deviations in the average pre-prediction errors
may be the result, and possibly also higher standard deviations of the pre-prediction errors.
In the two case studies this decrease in prediction quality had no influence in the degree to
which people accepted the model. Actually, the goal is to reach the best actual performance
as possible, and not the best pre-predicted performance. By using the method each
measurement period one soon gets more insight in relationships between key variables. The
expected as well as the achieved performance becomes more open to discussion. The rousing
of these kinds of discussions is the main goal of the use of the performance evaluation and
diagnosis method.

Besides the above more general benefits, the benefits for production unit control purposes in
particular are:

- The ability to do what-if analyses to improve the allocation of capacity with regard
to workcenter utilization and throughput.
- The prediction model calculates realistic throughput targets for one measurement
  period ahead, because it uses important information about the state of the production
  unit. This has been demonstrated in the simulation experiments as well as in the two
case studies.
- Via the post-prediction insight can be gained about which throughput could have been
  achieved in the latest measurement period.
- A comparison of the post-predicted throughput with the actual throughput gives a
  quantitative rather than qualitative or intuitive explanation of the throughput that has
  been achieved by the total production unit.
- The prediction model makes it possible to evaluate the expected and realized
  throughput at each single workcenter. In this way relative differences between
  workcenters can be traced that can be used to set priorities with respect to the
diagnosis at individual workcenters.

For goods flow control in particular, the following benefits can be summed:

- Schedulers may be interested to do what-if analyses with respect to the production
  orders that are planned to release for each measurement period. Instead of using a
production order release policy, one should manually enter the selected production orders as expected work supply. This can increase the production unit's performance with regard to the utilization rate; the production orders can be spread more over the individual workcenters so that a more equal utilization will be the result which results in a smoother flow of the production orders within the production unit.

- The model gives schedulers the opportunity to contact customers before they start complaining about late deliveries. The model thus can be used as a customer oriented tool that gives reliable promises of order completion dates. If schedulers and customers have agreed upon new delivery dates of production orders, these new delivery dates can be used to reschedule all production orders on hand.

- Compared to manual methods to estimate individual production order completion times, the prediction model gives the expected completion times for all production orders at once. It won't be necessary anymore to calculate new estimates of production order completion times separately, which can save a lot of time for the schedulers.

- The updating of expected production order completion times takes place each measurement period, and is based on the current work in process of a production unit and the near-future expectations. Rescheduling suggestions provided by MRP do not take into account this available information, and will therefore be worse than the estimations given by the state-dependent prediction model.

8.2 Conditions and guidelines

Based on our experiences with the use of the performance evaluation and diagnosis method in both a theoretical context, i.e. the simulation studies, and a practical context, i.e. the two case studies, we will draw up some conditions and guidelines regarding the use of the method in practical situations. When these conditions are fulfilled and the guidelines are followed, we expect that the potential benefits that can be obtained by using the method will be maximized. The conditions and guidelines can be subdivided into:

- organizational conditions and guidelines;
- modelling and measurement conditions and guidelines;
- technical conditions and guidelines.

ORGANIZATIONAL CONDITIONS AND GUIDELINES

- The production units for which the prediction model will be used should be relatively independent of other production units with respect to the subsequent operations in a production order's routing. In other words, production orders should not return to the
same production unit many times because this makes it difficult to make production order completion time estimations over the whole production order routing. Actually, the predictions of the production order completion times are only directed at subsequent workcenters in a production order’s routing within the same production unit; if there are more operations to be done in the same production unit in a later stage of the production order’s routing, then these operations will not be taken into account because for the intermediate operations no flow time estimation is made.

The goals for production unit control and goods flow control should be unambiguous and should have a clear ranking in priority, otherwise conflicts may arise about the kind of decisions that will be made to overcome or avoid certain performance problems. Usually, this goal-setting is involved with trade-offs that have to be made between goods flow control aspects (e.g. delivery reliability), production unit control aspects (e.g. throughput) and costs (e.g., extra capacity allocation, outsourcing). Unambiguous goals facilitate the process of discussing about the performance, not only within production units but also between different organizational functions.

The distinction of responsibilities and authorities in an organization should be in conformance with the stated goals and the developed performance measures. One must have the means to influence the performance, otherwise a performance measure is of limited use. For example, in the two production units as described in Chapter 6, the schedulers did not have enough possibilities to improve their delivery reliability. The shift leaders of the production units appeared to have the greatest influence on this performance measure by making the capacity allocation decisions.

Groups should be created in which all people should participate who have an influence on a certain performance. This stimulates an open discussion about the expected as well as the achieved performance, and it provides a better attunement of operational decisions. To structure the discussions a kind of feedback report, such as used by the ProMES system of Pritchard et al. (1988), can be used which is a formal description of the performance for the past period. The structure of such a feedback report specifically directed at the performance of a production unit is given in Table 8.1.

**MODELLING AND MEASUREMENT CONDITIONS AND GUIDELINES**

- Knowledge about queuing theory and relationships between the key production unit performance measures are the most important prerequisites to build a performance prediction model. Knowledge about the production unit itself is less required by the modeler, because this knowledge can be obtained from the people working within the production unit.
Table 8.1 Example of the structure of a feedback report.

<table>
<thead>
<tr>
<th>Input</th>
<th>Expected</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Work supply</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance</td>
<td>Pre-predicted</td>
<td>Post-predicted</td>
</tr>
<tr>
<td>Throughput</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delivery reliability</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• One should be careful to use prediction rules that are already developed for a specific production unit. Each production unit has to be modelled anew, each using its own, specifically developed prediction rules, taking into account the specific characteristics of the production unit. If there already have been developed several state-dependent prediction rules, these prediction rules thus should only be used as a starting point for developing other prediction rules, and they should not be copied and used immediately in different situations.

• The length of the measurement period determines the frequency of the performance evaluation. The performance targets that are used for the performance evaluation thereby always should be in conformance with the length of the measurement period. So, if performance targets are set for a week then the evaluation should take place at the end of the week, but if a performance target is set for a month then the performance evaluation should not take place earlier than over a month. In this way the performance evaluation is done in a fair-minded way with regard to the people that are judged by the performance evaluation.

• The length of the measurement period should be set such that:
  i) important changes in the state as well as the performance of a production unit are observable;
  ii) decisions can be made to overcome performance problems that are expected in the near future;
  iii) the horizon within which the main part of the external work supply is known (with respect to volume and mix) is longer than the length of the measurement period.

With respect to the second issue, the planning horizon of the customer is of importance. For example, if a production order is predicted to be completed tardy and
the predicted completion date falls within the planning horizon of the customer, then
the production unit should undertake extra efforts to complete the production order on
time. If the predicted completion date however lays such far in the future that it falls
largely out of the customer's planning horizon, then it can be asked to change the
customer's future planning. The third condition should be fulfilled in order to make
the pre-predictions of throughput and production order completion dates as reliable as
possible. Actually, in the simulation experiments we have shown that knowledge about
the work supply can significantly improve the reliability of the performance
predictions.

- All performance measures that are used in the state-dependent prediction model should
  be related to the organizational goals, otherwise the performance measure doesn't have
  a function. In the example of the two cases, the production order completion times
  were in fact directly related to the production unit's delivery reliability, which was one
  of the most important performance measures. A method that focuses on the laying of
  these kinds of relationships between performance measures and the organizational
goals on the one hand and between performance measures among themselves is given
  by Flapper et al. (1996).

- Systematic prediction errors that cannot be explained should be kept in the model and
  made visible until the causes are found. So, the prediction rules should not be adjusted
  with the systematic errors. We think this forces users of the model every measurement
  period to search for causes of the observed systematic deviations in order to improve
  the prediction rules. If a prediction is modified with the systematic prediction error,
  there will be no drive to search for the causes for this error. In practice, the user of
  the model can adjust the prediction for himself with the systematic error. For example,
  a scheduler should take into account the systematic error to make promises of
  production order deliveries to the customer.

- In order to evaluate the quality of the state-independent prediction rules it can be
  worthwhile to include state-independent prediction rules in the prediction model.
  Especially within the introduction phase of the model, the users of the model can be
  convinced more easily on the benefit of the state-dependent prediction model. In the
  two production units in the preceding chapter we gave already gave an example by
  including the "prediction" results of the planned operation due dates by MRP.

**TECHNICAL CONDITIONS AND GUIDELINES**

- The variables of the prediction model and the performance measures should be defined
  unambiguously otherwise the predictions will be difficult to interpret. For the capacity
  for example, one should agree whether the net available capacity or the gross available

_Towards implementation_
capacity is meant. Therefore, it should be checked whether all people using the prediction model use the same definitions for the variables. Somewhat related to this issue is the actual execution of the measurement. It is important that all variables and performance measures are measured objectively. Automated measurements therefore are being preferred to manual measurements or users' estimations. In the two case studies for example, it was difficult for the shift leaders to exactly determine the amount of allocated capacity per workcenter, which may lead to poor predictions.

- The time required for calculating the pre-predicted and post-predicted performance should be as short as possible. The longer the response times, the greater the probability is that after a while the model will not be used anymore. For this reason, the prediction rules for the two production units described in Chapter 7 were simplified by replacing the complex method of de Kok to calculate future work in process levels (which took about 8 minutes) with simple calculations of the average work in process (which only took a few seconds). The kind of computer the program is run on is the most essential factor with regard to this issue.

- If a software tool is built that contains the prediction model, then this tool should be as user-friendly as possible. Therefore, requirements as stated by the users of the model should be included as much as possible.
CHAPTER 9

CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

The goal of this research was to develop a method that could be used for the evaluation and diagnosis of the performance of complex production units in the short term. The performance considered in the research was restricted to the throughput realized by a complex production unit and the completion time of the production orders of which the delivery reliability performance measure is a derivative. The throughput is an important performance measure for internal control and evaluation purposes, whereas the completion times of production orders in a production unit are relevant with respect to decisions made at the goods flow control level and negotiations with customers about (changes in) delivery dates. In this chapter we will report on the main findings related to the development and use of the performance evaluation and diagnosis method. Also some suggestions for further research will be given.

9.1 Conclusions

In Chapter 2 we defined the performance management process and used the process as an important starting point for the development of an appropriate performance evaluation and diagnosis method. This performance management process, which is generally applicable, has two important characteristics. First, there is an explicit distinction between performance measurement, performance targets, performance evaluation, and performance diagnosis. Each aspect is dealt with separately in this research. This distinction is necessary, because each of these aspects requires its own approach. The belief that performance measurement itself can lead to performance improvements should therefore be approached with some reluctancy. The second important characteristic is that the process is a closed loop system, which means that a feedback mechanism is built in that provides that the method will be used again and again. In this way the performance management process can be regarded as a process that is directed at continuous improvement of the performance.

Throughout the research the process of performance management has developed into a specific method for performance evaluation and diagnosis for production units in the short
An essential characteristic of this newly developed method is that not only the near future performance is regarded, but also the near past performance; the function of the method in this way is predictive as well as explaining. This is in contrast to existing performance models which are directed at the prediction function only. Another characteristic of the newly developed performance evaluation and diagnosis method is that relevant, actual information about the state of the production unit is used. The state of a production unit is defined as the combination of the actual work in process, the expected available capacity, and the expected work supply. This information is needed because the method focuses at the performance in the short term. Most existing performance models are more directed at the performance on the longer term. Because no state information is required for these models, these models are state-independent. Finally, short term feedback is very important in complex production units, because the explanation of the achieved performance in the performance diagnosis phase should be given as quick as possible to take appropriate improvement actions on the one hand and to get more insight on the other hand.

The performance prediction model and the method to evaluate and diagnose the performance of complex production units in the short term are closely connected with each other. Actually, each measurement period a performance target (i.e. the pre-predicted performance) has to be calculated by the prediction model for the performance evaluation, and for the diagnosis of the achieved performance an ex-post performance target is required which also has to be calculated by the prediction model. The method itself is generally applicable, but the prediction model should be made specific for the production unit the model will be used for. The latter means that the specific characteristics of the production unit and the policies that are defined (e.g. production order release and sequencing policies) and should be part of the performance prediction model.

The purpose of the performance prediction model is to provide realistic ex-ante and ex-post performance targets for the short term by using relevant information of the state of a production unit. The current performance level of a production unit thereby is not important; the model works equally well for production units with a relatively poor performance as for production units that perform relatively well. This makes the performance prediction model applicable for every production unit. The pre-predicted performance should be regarded as a realistic performance target that can be achieved if the work is organized well (that is, in a way that is in conformance with the assumptions underlying the prediction model). The pre-predicted performance is the basis for discussions between the people involved to determine how the work should be organized to realize the performance targets. The performance target should not be used as an absolute value that is used for the performance evaluation at the end.
of a measurement period, but as a value that gives insight in the direction the performance will probably go. The pre-prediction results should encourage the people involved to discuss about the near future performance. In fact, the activation of these kinds of discussions is the primarily goal of the use of the performance evaluation and diagnosis method. Just like the pre-predicted performance target, in practice the ex-post performance target for the diagnosis (determined by the post prediction) should also not be treated as an exact, absolute value for judging individual employees, but it should be used as a basis for the search for the causes for the performance deviations. The post-predicted performance shows which performance could have been achieved given the actual realizations of the input variables. A performance diagnosis should therefore always be conducted, because it is the only way for a fair-minded evaluation of the actual performance.

From a practical point of view the predictions should be "good enough", in the sense that they show the direction in which the performance will go (i.e. the pre-predictions) or should have been gone (i.e. the post-predictions). Because of the many modeling errors and measurement errors that can occur in a practical context, we tested the state-dependent prediction rules in a controlled environment by simulation experiments. The main goal of this simulation study was to evaluate to what extent the use of actual state information can improve the performance predictions in comparison with state-independent predictions. As a measure of the quality of the predictions, the standard deviation of the prediction error was chosen. The situations that were simulated were all types of the classical job-shop as defined by Jackson (1957), which is regarded as the most difficult type of a job-shop to predict the performance of. This type of job-shop had been chosen to link up with the relevant literature, because performance analyses of state-independent models often take this kind of a job-shop as starting point. The situations we distinguished for the simulation experiments differed in the number of workcenters and in the utilization rate. For the throughput predictions the maximum decrease in the standard deviation of the prediction error was 90.6% by using a state-dependent prediction rule instead of a state-independent prediction rule. For the production order completion time predictions the maximum decrease was 54.1%. The extent to which the state-dependent predictions are more reliable than the state-independent predictions depends on the utilization rate and the number of workcenters in the production unit. With respect to these relationships, the following observations were made:

- The higher the utilization rate, the relatively more reliable the state-independent throughput predictions are. The opposite holds for the state-dependent throughput predictions with known volume of the work supply.

**Conclusions and suggestions for further research**
The increase in the prediction reliability by using a state-dependent prediction rule for the throughput instead of a state-independent prediction rule is relatively smaller for higher utilization rates.

The increase in the prediction reliability by using a state-dependent prediction rule for the throughput instead of a state-independent prediction rule is relatively larger for a larger number of workcenters in a production unit.

The reliability of both state-independent and state-dependent production order completion time predictions for production orders that are predicted to be completed within one measurement period, decrease by higher utilization rates. This effect is relatively larger for the state-independent production order completion time predictions.

The larger the number of workcenters in a production unit, the smaller the relative increase in prediction reliability for both the state-dependent and state-independent production order completion time predictions.

A final conclusion about the simulation experiments is that the use of information about the volume of the work supply for the next measurement period leads to more reliable predictions compared with situations where the volume of the work supply is unknown. For the throughput predictions the maximum decrease in the standard deviation of the prediction error was 88.7%, whereas this maximum decrease for the production order completion time predictions is 16.7%.

From the simulation experiments as well as the case studies it appears that the actual work in process, the available capacity, and the work supply are indeed important variables that should be part of a state-dependent prediction model. The use of this model within the performance evaluation and diagnosis method in the two case studies we carried out thereby resulted in a better insight in the processes that take place in the production units, a better insight in the factors that cause a lower performance than expected, and a better decision making. However, due to several kinds of practical causes, the performance evaluation and diagnosis method was and could not be used in a way that would lead to all intended results. Based on these causes, we developed some organizational conditions and guidelines regarding the use of the performance evaluation and diagnosis method. These organizational guidelines are:

- A production unit should be relatively independent of other production units with respect to the subsequent operations in the routings of the production orders.
- The goals for production unit control and goods flow control should be unambiguous and should have a clear ranking in priority.
• The distinction of responsibilities and authorities of organizational functions regarding production unit control and goods flow control should be in conformance with the stated goals and developed performance measures.

• Groups should be created in which all people should participate who have an influence on the performance to discuss the expected performance as well as the achieved past performance.

If these organizational conditions and guidelines are fulfilled (together with some modeling and measurement conditions and guidelines and technical conditions and guidelines) then the following benefits for production unit control and goods flow control can be expected:

• An improvement of the operator capacity allocation decision resulting in a higher production unit's throughput and a higher delivery reliability.

• More realistic and objective performance targets to be used for discussions about the expected near future performance and diagnosis of the achieved performance.

• The possibility to quantify throughput problems at workcenter level.

• An improvement of the production order release decision, which can result in a higher production unit's throughput and a higher delivery reliability.

• Quick and more reliable updates of expected production order completion times that can be used to inform customers on time about changes in the planned deliveries.

9.2 Suggestions for further research

In our opinion, the implementation of the newly developed performance evaluation and diagnosis method in other production units is the most important direction for further research. The reason for more empirical research is that we think that we still don't understand to full extent what kind of processes really take place at the production unit level. In this research we mainly concentrated on three explaining variables only: the actual work in process, the work supply, and the available capacity. For the simulation experiments and the two cases these variables had quite a large explaining potential, which means that these variables are indeed the most important variables to be considered. This is in line with the existing (long term directed) performance models that also use these variables. However, we think that there are also production units with more and/or other relevant variables. These variables may however be very specific. Knowledge about these relevant variables will lead to a better understanding of how the performance of a production unit is realized. Next, this knowledge can be used to build more realistic prediction models for performance predicting and performance explaining functions. As a result of that, the errors that are currently made by the modeling process can be reduced.
The implementation of the performance evaluation and diagnosis method in other production units is highly recommended to get more insight in how a certain performance is achieved. Thereby, the prediction rules for the throughput and production order completion times, as generally described in Chapter 4, should be adjusted to the specific characteristics of the production units considered. In many situations it will be uncertain to what extent the prediction reliability can be increased by using state-dependent prediction rules instead of state-independent prediction rules. Therefore, it is useful to start with conducting simulation experiments with different settings as we used. These settings should be based on empirical data. For example, the impact of other production order release and sequencing rules on the prediction reliability of state-dependent as well as state-independent prediction rules is unknown so far. Also, the effect of using state-dependent prediction rules instead of state-independent prediction rules in less complex production units can be tested by simulation studies. Summarizing, we think there is a broad area left for which the reliability of state-dependent prediction rules can be tested.

The search to the causes of the systematic prediction errors in the model and the standard deviations of the prediction errors as described in chapter 5 and 7 is another direction for further research. Although the systematic prediction errors fall in the range of the prediction reliability, they can be regarded as annoying by the people using the prediction results. Therefore, we encourage mathematicians to develop prediction rules without systematic errors and with at least the same level of prediction reliability. Simulation experiments are recommended to find out which values of the standard deviation of the prediction errors can be expected for specific situations, given the proposed way of working. We expect that this search towards better prediction rules at the same time can lead to more insight in the processes that take place at the production unit level.

Another interesting topic for further research concerns the behavioral aspect in production units. In the case studies we have seen that the sequencing policy was not fulfilled completely. This negatively influences the quality of the predictions, because the actual behavior differs from the expected behavior. However, it is impossible so far to quantify to what extent the actual performance is influenced by a deviating decision behavior as proposed. Generally, in each model assumptions are made about relevant agreements that are more or less formalized into some policies. For the production unit control the way in which production orders should be released and the way in which sequence production orders should be selected for an operation are typical examples of agreements that are formalized. The basic question now is: how can we make visible the impact is on the performance of a decision that
deviates from the policy? The answer to this question can decision makers make more aware of their decision behavior.

In this research we only dealt with non-financial performance measures. In Chapter 3 we stated that the financial performance ultimately determines the survival of an organization. At some level, a linkage should be made between the financial and non-financial performance measures. At the production unit level, we think there are some opportunities for further research. In this respect, a basic research questions can be: when should operational decision makers make a decision that positively influences a non-financial performance measure (e.g. delivery reliability), but that also results in extra costs (e.g. the decision to work overtime)? The goal of this research direction is to relate operational decisions with non-financial performance measures on the one hand, and financial performance measures on the other hand.
References


References


References
Appendix A

Calculation of $\alpha[I_t(n)]$

In this appendix we will explain how the method works to calculate the fraction of the work supply, denoted by $\alpha[I_t(n)]$, that on average will be processed during a measurement period. We will also demonstrate by three simulation experiments that the modeling error we make in the calculations has only a small effect on the calculated values of the fractions compared with the "actual" values of the fractions, which are obtained from the simulation experiments. The calculated fractions will be used in the simulation experiments that are described in Chapter 5. Because of the symmetric routing matrix of the production orders, each workcenter has the same characteristics with respect to the arrival and departure pattern of production orders. Consequently, the index $i$ that represents the workcenter number can be neglected. So, instead of calculating $\alpha[I_t(n)]$'s for each workcenter, it suffices to calculate $\alpha[I(n)]$ for an arbitrary workcenter. To prevent confusion between the calculated and simulated fractions, we will index the fractions with "c" and "s" respectively.

**Calculation Method**

To make the calculation method better understandable, the variables needed to calculate $\alpha_c[I(n)]$ for each value of $I(n)$ are given in Figure A.1. Recall that we only need to calculate the fraction of the work supply that on average will be processed for situations where $I(n)<40$, because $\alpha_c[I(n)]$ is zero for situations where $I(n)\geq 40$.

![Figure A.1 Variables needed to calculate $\alpha[I(n)]$.](image)

Calculation of $\alpha[I_t(n)]$
As has been mentioned in Chapter 5, the fraction of the work supply that on average will be processed will be calculated only once (i.e. at the beginning of a simulation experiment) to save computational time. Because the value of I(n) can take any possible real value, it is impossible to calculate \( a[I(n)] \) for each of these possible values. Therefore, the following steps in the calculation method are done for integer values of I(n) only, starting at 0 hours and ending at 39 hours. Values of I(n) greater than 40 will not be considered in the calculation method, because all available capacity will then be used for processing the actual work in process, which means that \( a[I(n)] \) will be zero for these situations.

**Step 1**

Given the actual work in process at the beginning of a measurement period, the expected amount of work supply is determined that will have arrived at the moment all the actual work in process will have been processed. Because the work in process is given in hours, the actual work in process will be processed completely at time \( t+I(n) \). The average amount of work supply that will have arrived during the time period \( [t, t+I(n)] \) is given by \( \lambda \cdot \hat{p} \cdot I(n) \) where \( I/\lambda \) is the average interarrival time of production orders to the workcenter and \( \hat{p} \) the average processing time. This expected amount of work supply is equal to the expected amount of work in process at that time, and is denoted by \( \hat{I}(I(n)) \).

**Step 2**

Given \( \hat{I}(I(n)) \), we now determine \( \hat{I}(I(n+1)) \). In the simulation experiments, the capacity is continuously available. Hence, \( \hat{I}(I(n+1)) \) depends only on the number of production order arrivals during \( T-I(n) \). Let X(h) be the number of production order arrivals during the time interval of length h. Let h be equal to \( T-I(n) \), then \( \hat{I}(n+1) \) can be written as

\[
\sum_{k=0}^{75} P(X(h)=k) \hat{G}(k),
\]

where \( P(X(h)=k) = \frac{e^{-\lambda h}(\lambda h)^k}{k!} \) according to the Poisson probability density function, and \( \hat{G}(k) \) is the expected amount of work in process after \( k \) production order arrivals, calculated with the iteration method developed by de Kok (1989). At this point we make a modeling error, because the method of de Kok does not state during which time interval the arrivals of the production orders will be. In other words, this procedure is time-independent. By a number of simulation experiments, that are described later in this appendix, we will show how large this modeling error is. Finally, the calculation is restricted to 75 arrivals of production orders, because the probability that more than 75 production orders arrive is negligible (i.e. approximately \( 4.2 \times 10^{-5} \)).

128

**Appendix A**
Step 3
In the preceding step we calculated $\hat{I}(n+1)$ for each starting value of $I(n)$. Now, the fraction of the work supply that on average will be processed can be calculated by:

$$\alpha_c[I(n)] = \frac{\hat{W}(n) - \hat{I}(n+1)}{\hat{W}(n)}.$$

Experiments
Three experiments (one experiment per utilization level) have been carried out to evaluate the quality of the above procedure to calculate the $\alpha[I(n)]$'s. The simulation settings are equal to the settings described in Section 5.2. We took the data of the first 6000 measurement periods of each experiment for our evaluation. For each measurement period we registered the starting value of the work in process, $I(n)$, the actual work supply during a measurement period, $W(n)$, and the actual work in process at the end of a measurement period, $I(n+1)$. The measurements were sorted in decreasing order of $I(n)$. Next, we rounded each $I(n)$ value to an integer value by using the formula INT(0.5 + $I(n)$), where INT is the function that rounds a value to the nearest integer value. For each of the 40 classes of $I(n)$ we calculated the average actual work supply, $W(n)$, and the average work in process at the end of the measurement period, $\bar{I}(n+1)$. Finally, $\alpha_s[I(n)]$ was calculated by:

$$\alpha_s[I(n)] = \frac{\bar{W}(n) - \bar{I}(n+1)}{\bar{W}(n)}.$$

In Figure A.2 we show the calculated and "actual" (i.e simulated) fractions of the work supply that on average is processed. The calculated values are marked with a square, while the actual values are marked with a dot. For the ease of survey, the values of the integer values of $I(n)$ are connected by means of a line. The shaded areas in the figure represent the ranges in which the simulated values will probably be, based on the standard deviations on each simulated value of $\alpha_s[I(n)]$.

Generally, it can be observed that for the three experiments only little deviations occur between the simulated and calculated $\alpha[I(n)]$'s. The large deviations that occur stem from having too few observations for certain values of $I(n)$. This can especially be observed for the larger values of $I(n)$ for the experiments with a utilization level of 80% and 60%. In the latter experiment there were no measurement periods with values of $I(n)$ greater than 11 hours. For that situation the probability that $I(n)$ is larger than 11 is equal to $1 - e^{-0.6*40}$ which is about equal to 1.

Calculation of $\alpha[I_i(n)]$
As a measure for the reliability of the calculated values, the average error is taken for the situations where exist both a calculated and simulated value. This measure of reliability is defined as

\[ \frac{1}{N} \sum_{i=0}^{40} (\alpha_{c}(l(n)) - \alpha_{s}(l(n))) \]

where N is the number of situations with a calculated and simulated value. For the experiments with utilization level 90%, 80%, and 60%, the values of the above measure were -0.006, -0.041, and -0.011 respectively. A somewhat easier way to evaluate the reliability of the calculated fractions is by looking in the figure at the variance on the simulated fractions. For all experiments, the variance on the simulated values is relatively large, so that the calculated values almost in all situations fall in this range. From these results we may conclude the difference between the calculated and simulated values is very small. From this we may conclude that the modeling error we make by using the iteration method of de Kok can be ignored.

Figure A.2 Calculated and simulated \( \alpha(l(n)) \)'s.
Appendix B

Simulation results

This appendix contains the data from the simulation experiments described in Chapter 5. For each experiment the mean prediction error and the standard deviation (st.dev.) of the prediction error are presented per subrun. An asterisk (*) means that there are no or too few observations.
### Prediction errors production order completion times

<table>
<thead>
<tr>
<th>Subrun</th>
<th>Experiment 1</th>
<th>Experiment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Subrun</td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-1.71</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-2.24</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-2.11</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>-1.38</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>-1.89</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>-1.80</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>-1.90</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>-1.45</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>-2.20</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>-1.15</td>
</tr>
</tbody>
</table>

### Prediction errors throughput

<table>
<thead>
<tr>
<th>Subrun</th>
<th>Experiment 1</th>
<th>Experiment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Subrun</td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.96</td>
</tr>
</tbody>
</table>
## Prediction errors production order completion times

<table>
<thead>
<tr>
<th>Subrun</th>
<th>Horizon 1 Mean</th>
<th>Horizon 1 St.dev.</th>
<th>Horizon 2 Mean</th>
<th>Horizon 2 St.dev.</th>
<th>Horizon 2 Mean</th>
<th>Horizon 2 St.dev.</th>
<th>Horizon 2 Mean</th>
<th>Horizon 2 St.dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-1.03</td>
<td>10.94</td>
<td>1.19</td>
<td>11.22</td>
<td>-0.33</td>
<td>24.51</td>
<td>4.91</td>
<td>25.26</td>
</tr>
<tr>
<td>2</td>
<td>-1.95</td>
<td>12.34</td>
<td>-1.61</td>
<td>14.85</td>
<td>-11.90</td>
<td>31.20</td>
<td>-7.41</td>
<td>36.86</td>
</tr>
<tr>
<td>3</td>
<td>-2.09</td>
<td>11.99</td>
<td>-1.44</td>
<td>13.36</td>
<td>-8.70</td>
<td>28.19</td>
<td>-5.65</td>
<td>31.91</td>
</tr>
<tr>
<td>4</td>
<td>-1.14</td>
<td>11.50</td>
<td>0.92</td>
<td>12.90</td>
<td>-3.93</td>
<td>29.12</td>
<td>2.32</td>
<td>30.38</td>
</tr>
<tr>
<td>5</td>
<td>-1.60</td>
<td>11.71</td>
<td>-1.29</td>
<td>14.59</td>
<td>-9.42</td>
<td>31.07</td>
<td>-4.04</td>
<td>35.41</td>
</tr>
<tr>
<td>6</td>
<td>-1.65</td>
<td>12.12</td>
<td>-0.19</td>
<td>13.98</td>
<td>-7.13</td>
<td>29.77</td>
<td>-0.94</td>
<td>34.59</td>
</tr>
<tr>
<td>7</td>
<td>-1.28</td>
<td>11.48</td>
<td>0.19</td>
<td>13.18</td>
<td>-4.53</td>
<td>27.42</td>
<td>1.12</td>
<td>30.94</td>
</tr>
<tr>
<td>8</td>
<td>-0.99</td>
<td>11.09</td>
<td>0.88</td>
<td>12.33</td>
<td>-3.45</td>
<td>26.88</td>
<td>3.65</td>
<td>29.43</td>
</tr>
<tr>
<td>9</td>
<td>-1.51</td>
<td>11.40</td>
<td>-0.49</td>
<td>12.01</td>
<td>-3.47</td>
<td>25.14</td>
<td>0.28</td>
<td>27.09</td>
</tr>
<tr>
<td>10</td>
<td>-1.17</td>
<td>12.14</td>
<td>0.53</td>
<td>14.22</td>
<td>-5.56</td>
<td>31.49</td>
<td>2.74</td>
<td>32.93</td>
</tr>
</tbody>
</table>

<p>| Experiment 4 |
| 1     | 10.42          | 0.78              | 11.22          | 1.19              | 22.33          | 5.58              | 25.26          | 4.91              |
| 2     | 10.94          | 0.50              | 14.85          | -1.61             | 25.86          | -3.09             | 36.86          | -7.41             |
| 3     | 10.86          | 0.28              | 13.36          | -1.44             | 24.61          | -0.10             | 31.91          | -5.65             |
| 4     | 10.60          | 0.72              | 12.90          | 0.92              | 25.08          | 2.22              | 30.38          | 2.32              |
| 5     | 10.61          | 0.78              | 14.59          | -1.29             | 25.38          | -0.10             | 35.41          | -4.04             |
| 6     | 10.68          | 0.57              | 13.98          | -0.19             | 25.05          | -0.85             | 34.59          | -0.94             |
| 7     | 10.53          | 0.64              | 13.18          | 0.19              | 23.48          | 1.92              | 30.94          | 1.12              |
| 8     | 10.32          | 0.74              | 12.33          | 0.88              | 24.43          | 3.38              | 29.43          | 3.65              |
| 9     | 10.65          | 0.62              | 12.01          | -0.49             | 22.00          | 3.37              | 27.09          | 0.28              |
| 10    | 10.87          | 0.60              | 14.22          | 0.53              | 26.77          | 0.81              | 32.93          | 2.74              |</p>
<table>
<thead>
<tr>
<th>Subrun</th>
<th>Experiment 5</th>
<th>Experiment 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>State-dependent Horizon 1</td>
<td>State-independent Horizon 1</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>St.dev.</td>
</tr>
<tr>
<td>1</td>
<td>-0.22</td>
<td>5.97</td>
</tr>
<tr>
<td>2</td>
<td>-0.62</td>
<td>6.65</td>
</tr>
<tr>
<td>3</td>
<td>-0.61</td>
<td>6.54</td>
</tr>
<tr>
<td>4</td>
<td>-0.42</td>
<td>6.20</td>
</tr>
<tr>
<td>5</td>
<td>-0.59</td>
<td>6.30</td>
</tr>
<tr>
<td>6</td>
<td>-0.52</td>
<td>6.55</td>
</tr>
<tr>
<td>7</td>
<td>-0.33</td>
<td>6.16</td>
</tr>
<tr>
<td>8</td>
<td>-0.28</td>
<td>5.93</td>
</tr>
<tr>
<td>9</td>
<td>-0.54</td>
<td>6.18</td>
</tr>
<tr>
<td>10</td>
<td>-0.43</td>
<td>6.19</td>
</tr>
</tbody>
</table>

**Prediction errors production order completion times**

**Prediction errors throughput**
### Prediction errors production order completion times

<table>
<thead>
<tr>
<th>Subrun</th>
<th>State-dependent Horizon 1 Mean</th>
<th>State-dependent Horizon 1 St.dev.</th>
<th>State-independent Horizon 1 Mean</th>
<th>State-independent Horizon 1 St.dev.</th>
<th>State-dependent Horizon 2 Mean</th>
<th>State-independent Horizon 2 Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-1.13</td>
<td>11.89</td>
<td>4.65</td>
<td>15.40</td>
<td>0.98</td>
<td>28.46</td>
</tr>
<tr>
<td>2</td>
<td>-1.76</td>
<td>12.35</td>
<td>-2.36</td>
<td>20.75</td>
<td>-7.60</td>
<td>31.24</td>
</tr>
<tr>
<td>3</td>
<td>-1.16</td>
<td>11.83</td>
<td>2.84</td>
<td>18.73</td>
<td>-3.76</td>
<td>30.07</td>
</tr>
<tr>
<td>4</td>
<td>-1.17</td>
<td>12.39</td>
<td>4.41</td>
<td>17.38</td>
<td>-0.28</td>
<td>28.61</td>
</tr>
<tr>
<td>5</td>
<td>-1.54</td>
<td>11.87</td>
<td>-1.65</td>
<td>18.59</td>
<td>-7.48</td>
<td>28.01</td>
</tr>
<tr>
<td>6</td>
<td>-1.45</td>
<td>12.06</td>
<td>-0.57</td>
<td>20.22</td>
<td>-6.65</td>
<td>30.17</td>
</tr>
<tr>
<td>7</td>
<td>-1.76</td>
<td>12.41</td>
<td>-3.16</td>
<td>24.17</td>
<td>-14.38</td>
<td>34.88</td>
</tr>
<tr>
<td>8</td>
<td>-1.22</td>
<td>11.66</td>
<td>2.67</td>
<td>16.05</td>
<td>-0.02</td>
<td>26.06</td>
</tr>
<tr>
<td>9</td>
<td>-1.58</td>
<td>12.04</td>
<td>-1.05</td>
<td>19.79</td>
<td>-7.33</td>
<td>28.43</td>
</tr>
<tr>
<td>10</td>
<td>-1.33</td>
<td>12.26</td>
<td>1.47</td>
<td>20.06</td>
<td>-6.99</td>
<td>31.02</td>
</tr>
</tbody>
</table>

### Prediction errors throughput

<table>
<thead>
<tr>
<th>Subrun</th>
<th>State-dependent Mean</th>
<th>State-dependent St.dev.</th>
<th>State-independent Mean</th>
<th>State-independent St.dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2.14</td>
<td>9.62</td>
<td>1.49</td>
<td>12.50</td>
</tr>
<tr>
<td>2</td>
<td>1.69</td>
<td>10.47</td>
<td>-1.12</td>
<td>13.22</td>
</tr>
<tr>
<td>3</td>
<td>3.11</td>
<td>11.05</td>
<td>2.69</td>
<td>14.22</td>
</tr>
<tr>
<td>4</td>
<td>2.82</td>
<td>11.45</td>
<td>2.63</td>
<td>15.13</td>
</tr>
<tr>
<td>5</td>
<td>1.57</td>
<td>9.54</td>
<td>-1.75</td>
<td>12.79</td>
</tr>
<tr>
<td>6</td>
<td>1.34</td>
<td>9.11</td>
<td>-0.66</td>
<td>11.44</td>
</tr>
<tr>
<td>7</td>
<td>1.42</td>
<td>9.72</td>
<td>-1.43</td>
<td>13.28</td>
</tr>
<tr>
<td>8</td>
<td>1.88</td>
<td>10.63</td>
<td>0.12</td>
<td>14.40</td>
</tr>
<tr>
<td>9</td>
<td>1.39</td>
<td>9.41</td>
<td>-1.07</td>
<td>12.39</td>
</tr>
<tr>
<td>10</td>
<td>2.91</td>
<td>11.36</td>
<td>1.68</td>
<td>15.45</td>
</tr>
</tbody>
</table>

### Experiment 7

<table>
<thead>
<tr>
<th>Subrun</th>
<th>Mean</th>
<th>St.dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>11.89</td>
<td>4.65</td>
</tr>
<tr>
<td>2</td>
<td>12.35</td>
<td>2.36</td>
</tr>
<tr>
<td>3</td>
<td>11.83</td>
<td>2.84</td>
</tr>
<tr>
<td>4</td>
<td>12.39</td>
<td>4.41</td>
</tr>
<tr>
<td>5</td>
<td>11.87</td>
<td>1.65</td>
</tr>
<tr>
<td>6</td>
<td>12.06</td>
<td>0.57</td>
</tr>
<tr>
<td>7</td>
<td>12.41</td>
<td>3.16</td>
</tr>
<tr>
<td>8</td>
<td>11.66</td>
<td>2.67</td>
</tr>
<tr>
<td>9</td>
<td>12.04</td>
<td>1.05</td>
</tr>
<tr>
<td>10</td>
<td>12.26</td>
<td>1.47</td>
</tr>
</tbody>
</table>

### Experiment 8

<table>
<thead>
<tr>
<th>Subrun</th>
<th>Mean</th>
<th>St.dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10.40</td>
<td>4.65</td>
</tr>
<tr>
<td>2</td>
<td>10.45</td>
<td>2.36</td>
</tr>
<tr>
<td>3</td>
<td>10.22</td>
<td>2.84</td>
</tr>
<tr>
<td>4</td>
<td>10.75</td>
<td>4.41</td>
</tr>
<tr>
<td>5</td>
<td>9.84</td>
<td>1.65</td>
</tr>
<tr>
<td>6</td>
<td>10.32</td>
<td>0.57</td>
</tr>
<tr>
<td>7</td>
<td>10.48</td>
<td>3.16</td>
</tr>
<tr>
<td>8</td>
<td>10.02</td>
<td>2.67</td>
</tr>
<tr>
<td>9</td>
<td>10.16</td>
<td>1.05</td>
</tr>
<tr>
<td>10</td>
<td>10.36</td>
<td>1.47</td>
</tr>
<tr>
<td>Subrun</td>
<td>Mean</td>
<td>St.dev.</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>1</td>
<td>-0.81</td>
<td>9.85</td>
</tr>
<tr>
<td>2</td>
<td>-1.40</td>
<td>10.15</td>
</tr>
<tr>
<td>3</td>
<td>-0.66</td>
<td>10.00</td>
</tr>
<tr>
<td>4</td>
<td>-0.88</td>
<td>10.24</td>
</tr>
<tr>
<td>5</td>
<td>-0.90</td>
<td>10.09</td>
</tr>
<tr>
<td>6</td>
<td>-1.19</td>
<td>10.04</td>
</tr>
<tr>
<td>7</td>
<td>-1.62</td>
<td>11.16</td>
</tr>
<tr>
<td>8</td>
<td>-1.06</td>
<td>10.11</td>
</tr>
<tr>
<td>9</td>
<td>-1.55</td>
<td>10.75</td>
</tr>
<tr>
<td>10</td>
<td>-0.88</td>
<td>10.22</td>
</tr>
</tbody>
</table>

**Experiment 10**

<table>
<thead>
<tr>
<th>Subrun</th>
<th>Mean</th>
<th>St.dev.</th>
<th>Mean</th>
<th>St.dev.</th>
<th>Mean</th>
<th>St.dev.</th>
<th>Mean</th>
<th>St.dev.</th>
<th>Mean</th>
<th>St.dev.</th>
<th>Mean</th>
<th>St.dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.39</td>
<td>9.14</td>
<td>0.68</td>
<td>11.39</td>
<td>0.47</td>
<td>29.99</td>
<td>1.87</td>
<td>40.18</td>
<td>-2.05</td>
<td>3.85</td>
<td>1.00</td>
<td>18.07</td>
</tr>
<tr>
<td>2</td>
<td>0.19</td>
<td>9.47</td>
<td>-1.08</td>
<td>12.85</td>
<td>-6.21</td>
<td>32.03</td>
<td>-1.49</td>
<td>40.43</td>
<td>-1.90</td>
<td>4.11</td>
<td>-1.15</td>
<td>18.47</td>
</tr>
<tr>
<td>3</td>
<td>0.46</td>
<td>9.07</td>
<td>0.90</td>
<td>11.48</td>
<td>-0.03</td>
<td>29.63</td>
<td>4.15</td>
<td>32.73</td>
<td>-2.02</td>
<td>3.97</td>
<td>2.21</td>
<td>19.06</td>
</tr>
<tr>
<td>4</td>
<td>0.29</td>
<td>9.69</td>
<td>0.77</td>
<td>10.86</td>
<td>5.44</td>
<td>25.86</td>
<td>6.57</td>
<td>29.89</td>
<td>-2.18</td>
<td>4.06</td>
<td>2.22</td>
<td>18.96</td>
</tr>
<tr>
<td>5</td>
<td>0.61</td>
<td>9.29</td>
<td>-0.65</td>
<td>11.80</td>
<td>-1.10</td>
<td>25.49</td>
<td>-2.42</td>
<td>32.96</td>
<td>-1.53</td>
<td>4.12</td>
<td>-1.85</td>
<td>17.18</td>
</tr>
<tr>
<td>6</td>
<td>0.20</td>
<td>9.14</td>
<td>-0.19</td>
<td>11.32</td>
<td>-1.64</td>
<td>25.38</td>
<td>-1.90</td>
<td>34.80</td>
<td>-1.69</td>
<td>4.02</td>
<td>-0.85</td>
<td>16.98</td>
</tr>
<tr>
<td>7</td>
<td>0.16</td>
<td>9.87</td>
<td>-1.53</td>
<td>14.11</td>
<td>-8.42</td>
<td>29.94</td>
<td>-8.77</td>
<td>43.79</td>
<td>-1.84</td>
<td>4.20</td>
<td>-1.48</td>
<td>18.75</td>
</tr>
<tr>
<td>8</td>
<td>0.22</td>
<td>9.39</td>
<td>-0.04</td>
<td>11.38</td>
<td>-0.54</td>
<td>27.02</td>
<td>0.38</td>
<td>31.61</td>
<td>-1.66</td>
<td>4.20</td>
<td>-0.26</td>
<td>18.13</td>
</tr>
<tr>
<td>9</td>
<td>0.03</td>
<td>9.70</td>
<td>-1.10</td>
<td>13.43</td>
<td>-5.69</td>
<td>28.48</td>
<td>-3.29</td>
<td>43.00</td>
<td>-1.87</td>
<td>4.13</td>
<td>-0.95</td>
<td>18.91</td>
</tr>
<tr>
<td>10</td>
<td>0.48</td>
<td>9.18</td>
<td>0.29</td>
<td>12.33</td>
<td>-6.71</td>
<td>32.50</td>
<td>-8.06</td>
<td>42.60</td>
<td>-1.88</td>
<td>4.08</td>
<td>1.32</td>
<td>19.26</td>
</tr>
<tr>
<td>Subrun</td>
<td>Experiment 11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>State-dependent</td>
<td>State-independent</td>
<td>State-dependent</td>
<td>State-independent</td>
<td>State-dependent</td>
<td>State-independent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Horizon 1</td>
<td>Horizon 1</td>
<td>Horizon 2</td>
<td>Horizon 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>St.dev.</td>
<td>Mean</td>
<td>St.dev.</td>
<td>Mean</td>
<td>St.dev.</td>
<td>Mean</td>
<td>St.dev.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-0.38</td>
<td>4.94</td>
<td>0.03</td>
<td>4.17</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.32</td>
<td>4.79</td>
<td>-0.12</td>
<td>3.91</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-0.16</td>
<td>4.42</td>
<td>0.08</td>
<td>3.83</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-0.20</td>
<td>4.54</td>
<td>0.04</td>
<td>3.95</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.16</td>
<td>4.40</td>
<td>0.04</td>
<td>3.82</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-0.22</td>
<td>5.03</td>
<td>-0.22</td>
<td>4.32</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-0.27</td>
<td>4.79</td>
<td>-0.19</td>
<td>4.29</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-0.08</td>
<td>4.48</td>
<td>0.00</td>
<td>3.98</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-0.20</td>
<td>4.50</td>
<td>0.00</td>
<td>3.90</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-0.25</td>
<td>4.70</td>
<td>-0.05</td>
<td>4.27</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subrun</th>
<th>Experiment 12</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>State-dependent</td>
<td>State-independent</td>
<td>State-dependent</td>
<td>State-independent</td>
<td>State-dependent</td>
<td>State-independent</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Horizon 1</td>
<td>Horizon 1</td>
<td>Horizon 2</td>
<td>Horizon 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>St.dev.</td>
<td>Mean</td>
<td>St.dev.</td>
<td>Mean</td>
<td>St.dev.</td>
<td>Mean</td>
</tr>
<tr>
<td>1</td>
<td>-0.32</td>
<td>5.03</td>
<td>0.03</td>
<td>4.17</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2</td>
<td>-0.26</td>
<td>4.84</td>
<td>-0.12</td>
<td>3.91</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>3</td>
<td>-0.12</td>
<td>4.44</td>
<td>0.08</td>
<td>3.83</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>4</td>
<td>-0.14</td>
<td>4.59</td>
<td>0.04</td>
<td>3.95</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>5</td>
<td>-0.10</td>
<td>4.46</td>
<td>0.04</td>
<td>3.82</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>6</td>
<td>-0.12</td>
<td>5.11</td>
<td>-0.22</td>
<td>4.32</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>7</td>
<td>-0.18</td>
<td>4.83</td>
<td>-0.19</td>
<td>4.29</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>8</td>
<td>-0.02</td>
<td>4.51</td>
<td>0.00</td>
<td>3.98</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>9</td>
<td>-0.14</td>
<td>4.55</td>
<td>0.00</td>
<td>3.90</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>10</td>
<td>-0.18</td>
<td>4.74</td>
<td>-0.05</td>
<td>4.27</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Subrun</td>
<td>Mean</td>
<td>St.dev.</td>
<td>Mean</td>
<td>St.dev.</td>
<td>Mean</td>
<td>St.dev.</td>
<td>Mean</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>---------</td>
<td>------</td>
<td>---------</td>
<td>------</td>
<td>---------</td>
<td>------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Experiment 13</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-0.42</td>
<td>7.88</td>
<td>2.24</td>
<td>11.52</td>
<td>3.00</td>
<td>22.46</td>
<td>12.45</td>
</tr>
<tr>
<td>2</td>
<td>-0.55</td>
<td>8.07</td>
<td>-0.22</td>
<td>16.54</td>
<td>-23.41</td>
<td>40.73</td>
<td>-22.79</td>
</tr>
<tr>
<td>3</td>
<td>-0.54</td>
<td>7.79</td>
<td>1.59</td>
<td>11.62</td>
<td>5.40</td>
<td>26.61</td>
<td>19.82</td>
</tr>
<tr>
<td>4</td>
<td>-0.75</td>
<td>8.32</td>
<td>0.31</td>
<td>14.37</td>
<td>-10.48</td>
<td>29.27</td>
<td>-2.10</td>
</tr>
<tr>
<td>5</td>
<td>-0.66</td>
<td>8.64</td>
<td>-1.27</td>
<td>16.42</td>
<td>-17.61</td>
<td>30.54</td>
<td>-9.73</td>
</tr>
<tr>
<td>6</td>
<td>-0.54</td>
<td>7.77</td>
<td>1.70</td>
<td>12.92</td>
<td>-10.79</td>
<td>31.96</td>
<td>5.08</td>
</tr>
<tr>
<td>7</td>
<td>-0.58</td>
<td>8.49</td>
<td>-0.85</td>
<td>16.31</td>
<td>-13.48</td>
<td>31.68</td>
<td>-4.37</td>
</tr>
<tr>
<td>8</td>
<td>-0.77</td>
<td>8.42</td>
<td>-5.42</td>
<td>18.92</td>
<td>-15.95</td>
<td>29.66</td>
<td>-31.86</td>
</tr>
<tr>
<td>9</td>
<td>-0.54</td>
<td>8.58</td>
<td>-0.28</td>
<td>14.63</td>
<td>-10.60</td>
<td>35.25</td>
<td>5.65</td>
</tr>
<tr>
<td>10</td>
<td>-0.42</td>
<td>7.77</td>
<td>1.34</td>
<td>13.71</td>
<td>-7.79</td>
<td>28.71</td>
<td>26.05</td>
</tr>
<tr>
<td><strong>Experiment 14</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.35</td>
<td>7.02</td>
<td>2.24</td>
<td>11.52</td>
<td>4.11</td>
<td>18.21</td>
<td>12.45</td>
</tr>
<tr>
<td>2</td>
<td>0.27</td>
<td>7.12</td>
<td>-0.22</td>
<td>16.54</td>
<td>-15.02</td>
<td>32.91</td>
<td>-22.79</td>
</tr>
<tr>
<td>3</td>
<td>0.26</td>
<td>6.92</td>
<td>1.59</td>
<td>11.62</td>
<td>3.34</td>
<td>19.98</td>
<td>19.82</td>
</tr>
<tr>
<td>4</td>
<td>0.20</td>
<td>7.12</td>
<td>0.31</td>
<td>14.37</td>
<td>-8.19</td>
<td>25.18</td>
<td>-2.10</td>
</tr>
<tr>
<td>5</td>
<td>0.18</td>
<td>7.75</td>
<td>-1.27</td>
<td>16.42</td>
<td>-11.25</td>
<td>25.93</td>
<td>-9.73</td>
</tr>
<tr>
<td>6</td>
<td>0.24</td>
<td>6.83</td>
<td>1.70</td>
<td>12.92</td>
<td>-7.35</td>
<td>26.70</td>
<td>5.08</td>
</tr>
<tr>
<td>7</td>
<td>0.28</td>
<td>7.26</td>
<td>-0.85</td>
<td>16.31</td>
<td>-8.92</td>
<td>26.23</td>
<td>-4.37</td>
</tr>
<tr>
<td>8</td>
<td>0.30</td>
<td>7.06</td>
<td>-5.42</td>
<td>18.92</td>
<td>-11.31</td>
<td>27.19</td>
<td>-31.86</td>
</tr>
<tr>
<td>9</td>
<td>0.39</td>
<td>7.38</td>
<td>-0.28</td>
<td>14.63</td>
<td>-7.84</td>
<td>30.19</td>
<td>5.65</td>
</tr>
<tr>
<td>10</td>
<td>0.32</td>
<td>6.90</td>
<td>1.34</td>
<td>13.71</td>
<td>-4.19</td>
<td>22.98</td>
<td>26.05</td>
</tr>
</tbody>
</table>
## Simulation results

<table>
<thead>
<tr>
<th>Subrun</th>
<th>Mean</th>
<th>St.dev.</th>
<th>Mean</th>
<th>St.dev.</th>
<th>Mean</th>
<th>St.dev.</th>
<th>Mean</th>
<th>St.dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Horizon 1</strong></td>
<td>State-dependent</td>
<td>State-independent</td>
<td>State-dependent</td>
<td>State-independent</td>
<td>State-dependent</td>
<td>State-independent</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Experiment 15</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-0.23</td>
<td>5.40</td>
<td>0.15</td>
<td>6.65</td>
<td>*</td>
<td>*</td>
<td>0.48</td>
<td>8.44</td>
</tr>
<tr>
<td>2</td>
<td>-0.28</td>
<td>5.67</td>
<td>-0.24</td>
<td>7.39</td>
<td>*</td>
<td>*</td>
<td>0.29</td>
<td>8.73</td>
</tr>
<tr>
<td>3</td>
<td>-0.29</td>
<td>5.43</td>
<td>-0.19</td>
<td>6.62</td>
<td>*</td>
<td>*</td>
<td>0.41</td>
<td>8.71</td>
</tr>
<tr>
<td>4</td>
<td>-0.34</td>
<td>5.83</td>
<td>-0.34</td>
<td>7.16</td>
<td>*</td>
<td>*</td>
<td>0.05</td>
<td>8.78</td>
</tr>
<tr>
<td>5</td>
<td>-0.21</td>
<td>5.13</td>
<td>0.56</td>
<td>6.04</td>
<td>*</td>
<td>*</td>
<td>0.37</td>
<td>8.78</td>
</tr>
<tr>
<td>6</td>
<td>-0.32</td>
<td>5.37</td>
<td>0.08</td>
<td>6.35</td>
<td>*</td>
<td>*</td>
<td>0.05</td>
<td>8.33</td>
</tr>
<tr>
<td>7</td>
<td>-0.28</td>
<td>5.46</td>
<td>-0.30</td>
<td>7.14</td>
<td>*</td>
<td>*</td>
<td>0.74</td>
<td>8.59</td>
</tr>
<tr>
<td>8</td>
<td>-0.24</td>
<td>5.69</td>
<td>-0.85</td>
<td>7.36</td>
<td>*</td>
<td>*</td>
<td>-0.73</td>
<td>7.84</td>
</tr>
<tr>
<td>9</td>
<td>-0.09</td>
<td>4.98</td>
<td>0.37</td>
<td>5.96</td>
<td>*</td>
<td>*</td>
<td>0.26</td>
<td>8.11</td>
</tr>
<tr>
<td>10</td>
<td>-0.20</td>
<td>5.41</td>
<td>0.41</td>
<td>6.49</td>
<td>*</td>
<td>*</td>
<td>0.50</td>
<td>8.83</td>
</tr>
<tr>
<td><strong>Horizon 2</strong></td>
<td>State-dependent</td>
<td>State-independent</td>
<td>State-dependent</td>
<td>State-independent</td>
<td>State-dependent</td>
<td>State-independent</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Experiment 16</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.03</td>
<td>5.34</td>
<td>0.15</td>
<td>6.65</td>
<td>*</td>
<td>*</td>
<td>-0.59</td>
<td>3.03</td>
</tr>
<tr>
<td>2</td>
<td>0.04</td>
<td>5.33</td>
<td>-0.24</td>
<td>7.39</td>
<td>*</td>
<td>*</td>
<td>-0.63</td>
<td>2.75</td>
</tr>
<tr>
<td>3</td>
<td>0.03</td>
<td>5.34</td>
<td>-0.19</td>
<td>6.62</td>
<td>*</td>
<td>*</td>
<td>-0.89</td>
<td>2.72</td>
</tr>
<tr>
<td>4</td>
<td>-0.02</td>
<td>5.62</td>
<td>-0.34</td>
<td>7.16</td>
<td>*</td>
<td>*</td>
<td>-0.85</td>
<td>2.49</td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>5.08</td>
<td>0.56</td>
<td>6.04</td>
<td>*</td>
<td>*</td>
<td>-0.81</td>
<td>2.87</td>
</tr>
<tr>
<td>6</td>
<td>-0.02</td>
<td>5.29</td>
<td>0.08</td>
<td>6.35</td>
<td>*</td>
<td>*</td>
<td>-0.91</td>
<td>2.95</td>
</tr>
<tr>
<td>7</td>
<td>0.00</td>
<td>5.21</td>
<td>-0.30</td>
<td>7.14</td>
<td>*</td>
<td>*</td>
<td>-0.76</td>
<td>2.60</td>
</tr>
<tr>
<td>8</td>
<td>0.11</td>
<td>5.50</td>
<td>-0.85</td>
<td>7.36</td>
<td>*</td>
<td>*</td>
<td>-0.42</td>
<td>3.08</td>
</tr>
<tr>
<td>9</td>
<td>0.10</td>
<td>4.87</td>
<td>0.37</td>
<td>5.96</td>
<td>*</td>
<td>*</td>
<td>-0.76</td>
<td>3.07</td>
</tr>
<tr>
<td>10</td>
<td>0.06</td>
<td>5.25</td>
<td>0.41</td>
<td>6.49</td>
<td>*</td>
<td>*</td>
<td>-0.82</td>
<td>2.80</td>
</tr>
</tbody>
</table>

## Prediction errors production order completion times

## Prediction errors throughput
<table>
<thead>
<tr>
<th>Subrun</th>
<th>Mean</th>
<th>St.dev.</th>
<th>Mean</th>
<th>St.dev.</th>
<th>Mean</th>
<th>St.dev.</th>
<th>Mean</th>
<th>St.dev.</th>
<th>Mean</th>
<th>St.dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Experiment 17</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-0.02</td>
<td>2.28</td>
<td>0.06</td>
<td>2.24</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>-0.11</td>
<td>8.89</td>
</tr>
<tr>
<td>2</td>
<td>-0.02</td>
<td>2.35</td>
<td>-0.07</td>
<td>2.32</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>-0.34</td>
<td>9.63</td>
</tr>
<tr>
<td>3</td>
<td>-0.09</td>
<td>2.09</td>
<td>-0.04</td>
<td>2.25</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>-0.34</td>
<td>10.20</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>2.19</td>
<td>0.05</td>
<td>2.17</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>-0.34</td>
<td>9.80</td>
</tr>
<tr>
<td>5</td>
<td>-0.07</td>
<td>2.24</td>
<td>0.00</td>
<td>2.23</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>-0.04</td>
<td>9.44</td>
</tr>
<tr>
<td>6</td>
<td>-0.08</td>
<td>2.16</td>
<td>0.00</td>
<td>2.19</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>-0.39</td>
<td>9.59</td>
</tr>
<tr>
<td>7</td>
<td>-0.03</td>
<td>2.51</td>
<td>-0.14</td>
<td>2.63</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>0.13</td>
<td>9.53</td>
</tr>
<tr>
<td>8</td>
<td>-0.03</td>
<td>2.42</td>
<td>-0.17</td>
<td>2.42</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>-1.25</td>
<td>9.23</td>
</tr>
<tr>
<td>9</td>
<td>-0.02</td>
<td>2.27</td>
<td>0.03</td>
<td>2.21</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>0.00</td>
<td>9.17</td>
</tr>
<tr>
<td>10</td>
<td>-0.06</td>
<td>2.13</td>
<td>0.00</td>
<td>2.26</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>-0.09</td>
<td>9.59</td>
</tr>
<tr>
<td><strong>Experiment 18</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-0.02</td>
<td>2.28</td>
<td>0.06</td>
<td>2.24</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>-0.02</td>
<td>1.84</td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>2.35</td>
<td>-0.07</td>
<td>2.32</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>0.12</td>
<td>1.83</td>
</tr>
<tr>
<td>3</td>
<td>-0.09</td>
<td>2.09</td>
<td>-0.04</td>
<td>2.25</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>-0.09</td>
<td>1.61</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>2.19</td>
<td>0.05</td>
<td>2.17</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>0.03</td>
<td>1.77</td>
</tr>
<tr>
<td>5</td>
<td>-0.07</td>
<td>2.24</td>
<td>0.00</td>
<td>2.23</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>0.00</td>
<td>1.65</td>
</tr>
<tr>
<td>6</td>
<td>-0.08</td>
<td>2.17</td>
<td>0.00</td>
<td>2.19</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>-0.12</td>
<td>1.64</td>
</tr>
<tr>
<td>7</td>
<td>-0.02</td>
<td>2.54</td>
<td>-0.14</td>
<td>2.63</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>0.12</td>
<td>1.88</td>
</tr>
<tr>
<td>8</td>
<td>-0.02</td>
<td>2.43</td>
<td>-0.17</td>
<td>2.42</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>0.19</td>
<td>1.97</td>
</tr>
<tr>
<td>9</td>
<td>-0.01</td>
<td>2.27</td>
<td>0.03</td>
<td>2.21</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>-0.02</td>
<td>1.79</td>
</tr>
<tr>
<td>10</td>
<td>-0.06</td>
<td>2.13</td>
<td>0.00</td>
<td>2.26</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>0.00</td>
<td>1.70</td>
</tr>
</tbody>
</table>
Appendix C

Prediction rules for the machining production units

In this appendix we will describe how the general prediction rules for the throughput and production order completion times as described in Chapter 4 have been modified to capture the specific characteristics of the two machining departments that are described in Chapter 7. In real life situations three aspects are important with regard to the development and implementation: robustness, user-friendliness, and simplicity. Robustness refers to the extent to which the predicted performance is sensitive to variations in the input variables. Especially a very detailed prediction model can lead to large variations in the predicted performance if, for example, demand or capacity characteristics change regularly. To avoid that such a prediction model should be continuously adjusted to the new characteristics, the robustness of the model is of great importance. User-friendliness refers to the extent to which the users have problems with the process of using the prediction model. In a real life situation considerable attention should be paid to this aspect to increase the degree of acceptance of the model. A last criterium we used is the criterium of simplicity. The prediction rules should be kept as simple as possible to keep the prediction rules comprehensible for the users (which also increases the degree of acceptance), to save time in the model building phase, and to save computational time with the performance predictions themselves. The requirement of simplicity should of course not be at the expense of the reliability of the predictions. Throughout the description of the development of the prediction rules in this appendix, we will indicate at which points the three above mentioned criteria played a role.

**THROUGHPUT PREDICTION RULE**

The description of the throughput prediction rule shall be limited to a single workcenter only, because the expected throughput for the production unit as a whole can be obtained easily by the summation of the expected throughput at the individual workcenters. Recall that the basic equation for the throughput prediction at workcenter i is:

\[ \hat{Y}_i(n) = \text{MIN}(I_i(n) + \alpha[I_i(n)]\hat{W}_i(n),\hat{C}_i(n)) \]  \hspace{1cm} (c.1)

where \( \hat{Y}_i(n) \) = the expected throughput at workcenter i in measurement period n (in hours)
I\(_i\)(n) = \text{the actual work in process at workcenter } i \text{ at the beginning of}
\text{measurement period } n \text{ (in hours)}

\alpha[I_i(n)] = \text{the fraction of the work supply that on average is processed in}
\text{measurement period } n, \text{ given the actual work in process}

\hat{W}_i(n) = \text{the expected work supply at workcenter } i \text{ in measurement period } n \text{ (in}
\text{hours)}

\hat{C}_i(n) = \text{the expected available capacity at workcenter } i \text{ in measurement period}
\text{n (in hours)}

Based on the three criteria we discussed in the introduction of this appendix, we modified
equation c.1 into:

\[ \hat{Y}_i(n) = \min[I_i(n) + \hat{W}_i(n), \hat{f}(n)\hat{C}_i(n)] \tag{c.2} \]

where \( \hat{f}(n) \) is the expected efficiency factor for the production unit in measurement period
n. An explanation for the differences will be given next.

Two differences can be observed between equations c.1 and c.2. The first difference is that
in the latter one the expected available capacity is multiplied by the expected efficiency in
measurement period n, \( \hat{f}(n) \). This efficiency factor has to be entered in the model and is
meant for the production unit as a whole. To avoid that the users of the model should enter
such a factor for each workcenter, it was decided from a user-friendliness point of view to
use one overall efficiency factor for the production unit as a whole instead of an efficiency
factor per workcenter. The idea behind this efficiency factor is that the gross available
capacity can not be used completely for production. Some of the available time is spent on
non-producing activities such as time for personal care, and time for teaching colleague
operators. The efficiency factor already was used by the production unit for performance
evaluation purposes.

The second difference between the basic and specifically developed prediction rule is that in
the specific throughput prediction rule it is assumed that on average the entire expected work
supply is processed (under the condition that the sum of the actual work in process and the
expected work supply do not exceed the expected available capacity). There are two reasons
why the expected work supply is not multiplied by a sort of fraction such as \( \alpha[I_i(n)] \) in the
basic prediction rule. First, from historical data about the actual work in process at the
different workcenters and the available capacity, it appeared that the actual work in process
at these workcenters usually was so high, that it exceeded the average available capacity on
these workcenters many times. The consequence is that the expected capacity will often be the restraining factor with respect to the expected throughput, regardless the expected amount of work supply. This first reason thus relates to the criterium of simplicity. Second, the dynamics in the production units with respect to demand characteristics (e.g. changes in production order mix and volume) and capacity characteristics (e.g. changes in the number of operators or the types of machines) make it impossible to determine more or less stable values of such kind of fraction for each workcenter. Demand characteristics as well as capacity characteristics change regularly which leads to fluctuations in the fraction of the work supply that on average is processed at a workcenter in the next measurement period. Ideally, each time there is an important change with respect to demand or capacity characteristics, the fraction actually should be calculated anew. From the simulation study we knew that this calculation takes relatively much computational time. Because the data processing already takes some considerable time, we decided to leave out these calculations for the moment. So, not only the robustness criterium led to the decision to leave out a kind of fraction such as $\alpha[I_i(n)]$, but also the user-friendliness criterium.

The expected work supply at workcenter $i$ in measurement period $n$, $\hat{W}_i(n)$, consists of the expected external work supply for workcenter $i$ in measurement period $n$, denoted by $\hat{W}_{i,\text{exr}}(n)$, and the expected internal work supply at workcenter $i$ in measurement period $n$, denoted by $\hat{W}_{i,\text{int}}(n)$. The expected external work supply for workcenter $i$ in measurement period $n$ is calculated by summing up the processing times of the production orders that are expected to be released to the production unit and that have to undergo their first operation at workcenter $i$. The expected internal work supply at workcenter $i$ in measurement period $n$ consists of the sum of the processing times of the production orders that are currently being processed or waiting at all workcenters other than workcenter $i$ and that are expected to be processed at those other workcenters within the current measurement period, and that have their next operation at workcenter $i$. For this last condition the actual work in process is compared with the expected available capacity corrected for the expected efficiency, $\bar{\eta}(n)\bar{C}_j(n)$. The concept of this calculation is illustrated in Figure C.1 for a simple situation with only two workcenters.

Suppose we want to know the expected amount of internal work supply in measurement period $n$ going from workcenter $j$ to $i$. The production orders that are waiting at workcenter $j$ are sorted on priority and then numbered from 1 to 7. Suppose that production orders 1, 3, and 7 have their next operation at workcenter $i$. Because the expected available capacity at workcenter $j$ in measurement period $n$, $\bar{\eta}(n)\bar{C}_j(n)$, equals the sum of the processing times of
the production orders 1 up and till 4, only production orders 1 and 3 are considered to be the expected internal work supply in measurement period n from workcenter j to workcenter i.

Sorted queue of production orders

\[ W_{i \text{int}}(n) \]

Work center

\[ \sum_{k=1}^{4} \hat{p}_{k,j} = \hat{\eta}(n) \hat{C}_j(n) \]

Figure C.1 Example showing the calculation of the internal work supply.

**PRODUCTION ORDER COMPLETION TIME PREDICTION RULE**

To predict the completion time of a production order, the expected production order flow time should be determined and added to the actual time. The expected production order flow time for production order k, denoted by \( \hat{F}_k \), is split up into two parts: the expected flow time at workcenter i where the production order is currently waiting at (denoted by \( \hat{F}_{k,i} \)) and the expected flow time at the remaining workcenters in the production order’s routing (denoted by \( \hat{F}_{k,\text{rem}} \)).

The expected flow time at workcenter i where a production order j is currently waiting at, can be determined by:

\[
\hat{F}_{k,i} = \hat{M}_{k,i} + \hat{E}_{k,i} + \hat{P}_{k,i}
\]

where
- \( \hat{M}_{k,i} \) = minimum waiting time of production order k at workcenter i (in hours)
- \( \hat{E}_{k,i} \) = expected extra waiting time of production order k at workcenter i (in hours)
- \( \hat{P}_{k,i} \) = pre-calculated or expected processing time (including setup time) of production order k at workcenter i (in hours).
\( \hat{M}_{k,i} \) is calculated as follows:

\[
\hat{M}_{k,i} = \sum_{x \in V_{k,i}} \hat{p}_{x,i}
\]

where \( V_{j,i} \) is the set of production orders at workcenter \( i \) with a higher priority (i.e. earlier Operation Due Date) than production order \( k \).

\( \hat{E}_{k,i} \) is determined as follows. In a period of six subsequent weeks, every week we randomly choose 5 production orders at different workcenters. For each of these workcenters the actual work supply during that week was registered. Next, for each of these workcenters we determine that part of the actual work supply that had an earlier operation due date than the chosen production order. From these calculations it appeared that about 50% of the production orders that arrived during the week had a higher priority than the production orders considered. For that reason we let \( \hat{E}_{k,i} = 0.5 \hat{W}_{i}(n) \).

The second step in the prediction of a production order's completion time consists of the estimation of the flow time for the remaining workcenters in a production order's routing. The expected waiting times at the other workcenters depend on the amount of work in process at these workcenters at the moment the production order arrives and the relative priority of the production order in comparison with the other production orders in the queue. Due to the dynamics in the two production units with respect to all kinds of unexpected disturbances (e.g. machine breakdowns, and unplanned activities like rework), it is impossible to predict precisely which production orders are waiting at a workcenter at the expected moment the production order arrives for which the expected completion time is predicted. Hence, the relative priority in the queue can not be determined. Therefore, we decided to leave out the relative priority and to use only the expected amount of work in process as the basis for the expected flow times of the production orders.

In comparison with the basic production order completion time prediction rule as described in Section 4.3, the prediction of future work in process levels at workcenters is strongly simplified by leaving out the iteration procedure developed by de Kok (1989). This has been done because of the user-friendliness criterium; the procedure took too much computational time which was not acceptable for the users of the model. This complex iteration procedure has been replaced by the calculation of the average work in process based on the latest 10 measurements periods. This average work in process is used as estimation for the measurements periods \( n+2, n+3, \) etcetera (For measurement period \( n+1 \), the estimation is used that is based on the throughput prediction as given in equation 4.17.) Interpolation is used,
if the expected production order arrival time falls between two measurement periods. So, the course of the expected amount of work in process at workcenter i with starting value $I_i(n)$ is

$$\hat{I}_i(n+1) = I_i(n) + \hat{W}_i(n) - \hat{Y}_i(n), \quad \hat{I}_i(n+2) = \bar{I}_i, \quad \hat{I}_i(n+3) = \bar{I}_i,$$

etcetera, where $\bar{I}_i = \sum_{l=n-d}^{n} I_i(l)$.

The expected flow time at the next workcenter, say m, in the remaining routing of production order i can now be calculated as follows:

$$\hat{F}_{k,m} = \frac{\hat{I}_m(t+\hat{F}_{k,i})}{\bar{C}_m} + \hat{f}_{k,m} \quad (c.5)$$

where $\bar{C}_m$ is the average available capacity at workcenter m over the last 10 weeks, and $\hat{I}_m(t+\hat{F}_{k,i})$ can be obtained by interpolation of the future work in process levels as described earlier. If there are more operations remaining, the expected flow time at these other workcenters can be determined in the same way. Then, $\hat{F}_{k,rem}$ can be determined by summing these expected flow times.
This research focuses on performance management of job-shop like production departments with respect to the short term performance of performance measures regarding production department control and goods flow control. For such kinds of production departments a method has been developed: i) to explain the achieved performance, ii) to evaluate the achieved performance, and iii) to predict the near-future performance. The developed method uses relevant information about the actual state of the department, which makes the method different from existing performance methods and models. The ultimate goal of the method is to establish the expected as well as the realized performance in an objective way, which results in the creation of a base for an open discussion. By using this method in a process of performance management, which takes the continuous improvement philosophy as starting point, an ongoing process directed at performance improvements is created.

The set of production orders that enter the type of production department we consider in this research are characterized by a large variety in volume and mix; the production orders differ in the number of operations, the sequence of the routings, and the processing times per operation. Often, the operator-capacity in this type of production department is multi-skilled. In addition, many disturbances, such as machine breakdowns, unforeseen absenteeism, a lack of raw materials, and a lack of product specific tools, may occur that influence the production department's performance. The inherent complexity of this type of production department makes it difficult to give reliable expectations about the near-future performance. But also the diagnosis or explanation of the realized performance is difficult. Often, qualitative explanations are given that sound plausible, but a quantitative explanation normally is not given. In addition, there is no instrument yet that gives clues for directions for performance improvements.

This research focuses on the development of an instrument for the performance evaluation and diagnosis of the production department's throughput (as a performance measure for departmental control) and production order completion times (for goods flow control). In practice these performance measures are the most important ones for performance evaluation purposes. Other, more aggregated performance measures such as the average production order flow time and delivery reliability can be deduced immediately by combining the above performance measures with available data about production order characteristics and capacity.
Performance evaluation is the process of comparing the actual performance with performance targets. In practice, these targets usually are directed at and set over a longer period of time (from several weeks till several months). However, the actual performance often is measured, reported, and evaluated on a shorter term (e.g. daily or weekly). This actual performance often fluctuates heavily in the short term caused by the many daily disturbances that may occur and the dynamics of the production processes themselves. Clearly, over a longer period of time the long term performance target should be met, but this will probably not the case for each measurement period. As a result, the long term performance target will not be realistic for the performance in the short term and thus not appropriate for the evaluation in the short term. Therefore, the frequency of the target setting should be equal to the frequency of the measurements and reporting. This means that short term performance targets are needed for the production departments. To generate short term performance targets we developed a prediction model which should make more realistic performance targets in the short term by using information about the actual work in process, the expected work supply in the short term, and the expected available capacity in the short term. These three variables indicate the actual state of the production department. The prediction model thus is state-dependent, which makes it different from existing performance prediction models for production departmental control. Besides the actual state of the production department, two policies will be considered that have an impact on the production department's performance in the short term: the production order release policy and the sequencing rule.

In a simulation study the influence of using information about the actual state on the prediction reliability of the performance is tested, where prediction rules that are state-independent were used as a baseline for the prediction reliability. For all defined situations a positive influence could be observed. The magnitude of the effect depends on the production department's utilization rate and the number of workcenters in the production department. For situations with a relatively high utilization rate the addition of actual information about the state is relatively less effective for throughput predictions compared with production order completion time predictions. For situations with a relatively low utilization rate the opposite holds. In addition, the more workcenters in a production department, the relatively better the throughput predictions will be, but the relatively worse the predictions for the production order completion times will be. From these simulation results it can be concluded that using information about the actual state is useful for every type of job-shop like production department. Using state information is the least effective in production departments with a high utilization rate and only a little number of workcenters.
The short term performance evaluation and diagnosis method contains the following steps. Each measurement period the prediction model generates an expected performance based on the actual work in process, the expected available capacity, and the expected work supply. By varying the input variables (i.e., capacity and work supply), what-if analyses can be executed, which may lead to better decisions with regard to the performance. This so-called pre-prediction serves as a realistic performance target for the short term, given the actual work in process and estimates about capacity and work supply. When the actual performance is measured and reported, this actual performance is compared with the performance targets, i.e., the performance evaluation. Next, a performance diagnosis should be executed, independently whether a performance deviation is observed or not, because it may be a coincidence that the actual performance corresponds with the performance targets. By a performance diagnosis, the actual performance is explained objectively and quantified. For that purpose, the prediction model is used again to make a so-called post-prediction. This post-prediction uses the initial actual work in process and the actual values instead of the expected values of the available capacity and the actual work supply. The post-predicted performance thus is the performance that could have been realized. This means that the comparison of the post-predicted performance with the actual performance in fact is the only fair-minded way to objectively evaluate the actual performance. A possible deviation between the pre-predicted and actual performance can be considered as a measure for the prediction quality of the prediction model, whereas a possible difference between the pre-predicted and the post-predicted performance is a measure for the quality of the estimations of the input variables. If the post-predicted performance and the actual performance deviate from each other, then this deviation is the unexplained part of the realized performance. To explain this remaining difference, a further diagnosis is needed, which will be qualitative by nature in contrast to the first diagnosis which is quantitative by nature. The qualitative diagnosis has not been worked out in this research.

The developed performance evaluation and diagnosis method is tested in two production departments on its usefulness for practice. The performance prediction rules of the generally described prediction model were adjusted to the specific characteristics of the two production departments. Thereby, the performance prediction model was constructed as a software tool. The method of performance evaluation and diagnosis appeared to be very useful in the two cases. First, what-if analyses were made to determine the allocation of the operator-capacity, which increased the insight in the relationships between actual work in process, capacity, work supply, throughput and production order completion dates. Second, the near-future performance became more open to discussion by the objective pre-predicted performance. Discussions were focused on the organization of the work to strive for realizing the pre-
predicted performance targets. Third, the real bottle-necks in the production departments could be shown in a quantitative rather than an intuitive or qualitative way. In this way the achieved performance could largely be explained quantitatively by deviations between the expected and actual values of capacity and work supply. When these deviations were analyzed, further discussions could then be focused and restricted on finding remaining causes (i.e. the qualitative diagnosis). The expected improvements of the prediction reliability by using information about the actual state could be observed only for the throughput predictions compared with the planned throughput by MRP. The state-dependent production order completion time prediction rule resulted in more reliable completion time predictions than the planned due dates by MRP, but the predictions were only equally well as the rules of thumb used by the schedulers.

With respect to the implementation of the performance evaluation and diagnosis method, several conditions and guidelines should be met to get the most benefits. The most important conditions and guidelines are:

- A production department should be relatively independent of other production departments with respect to the subsequent operations in production orders' routings.
- Goals for production departmental control and goods flow control should be unambiguous and should have a clear ranking in priority. Authorities and responsibilities of people who can influence the performance should be related to those goals.
- The persons who can influence the performance of a production department, should discuss the expected as well as the realized performance in groups. In this way the influence of each other's decisions on the performance can be understood better, and it will make the search to further performance improvements more effective.
- The performance targets should reflect the period of time for which a performance evaluation is done.

Generally, the performance evaluation and diagnosis method offers the following potentialities for production departmental control. First, what-if analyses can be conducted to improve the production order release decisions and the capacity allocation decisions. The pre-predicted throughput thereby serves as a subject for discussion how the work should be organized to achieve the generated performance targets. Second, the method shows quantitatively which workcenters in a production department structurally realize a larger or smaller throughput than expected. This can be used by departmental management to set priorities with regard to the problem solving. For goods flow control, the method offers the potentiality to anticipate on production orders that are expected to finish tardy by generating production order completion
time predictions. This information can lead to a different priority setting (after consultation with departmental management and/or customers), and can be used as communication tool towards (internal or external) customers.
Dit onderzoek richt zich op de korte termijn logistieke prestatie van job-shop-achtige produktie-afdelingen. Voor dit type afdeling is een methode ontwikkeld waarmee men i) de geleverde prestatie ten dele kwantitatief kan verklaren (de zgn. diagnose), ii) de geleverde logistieke prestatie kan beoordelen (de zgn. evaluatie), en iii) de logistieke prestatie kan voorspellen. In de ontwikkelde methode wordt zoveel mogelijk relevante informatie over de actuele toestand van de afdeling betrokken. De toevoeging van deze toestandsinformatie maakt de methode verschillend van bestaande methoden en modellen. Het uiteindelijke doel van de methode is om zowel de te leveren als de geleverde logistieke prestatie beter bespreekbaar te maken door gebruik te maken van objectieve data, waarbij een proces wordt gecreëerd waarin men voortdurend gericht is op het verbeteren van deze prestatie. Dit proces, het zogenaamde performance management proces, heeft de "continuous improvement" gedachte als uitgangspunt.

Het type produktie-afdeling dat we in dit onderzoek beschouwen krijgt orders te verwerken met een hoge variëteit in het aantal bewerkingen per order, bewerkingsvolgordes en bewerkingstijden. De beschikbare man-capaciteit is veelal in te zetten op verschillende soorten werkplekken en de vrijgave van nieuwe orders aan de afdeling vertoont een grote diversiteit in zowel aantal als soort. Bovendien kunnen er in dit soort omgevingen vele verstoringen optreden die invloed hebben op de logistieke prestatie, zoals machine-storingen, afwezigheid van personeel door ziekte, een tekort aan uitgangsmaterialen en een tekort aan produktspecifieke gereedschappen. Deze complexiteit maakt het voor het afdelingsmanagement moeilijk om een uitspraak te doen over welke prestatie men zal realiseren. Ook het beoordelen van de geleverde prestatie is moeilijk. Kwalitatief kunnen vaak oorzaken gevonden worden voor afwijkingen ten opzichte van de verwachte prestatie, maar een kwantitatieve toets wordt zelden of nooit uitgevoerd. Men heeft op dit moment bovendien geen instrument ter beschikking dat aangeeft waar in een afdeling mogelijkheden voor prestatie-verbetering te vinden zijn.

Dit onderzoek is gericht op ontwikkelen van instrumenten voor de beoordeling en diagnose van de totale output (hoeveelheid verzet werk in uren) van een produktie-afdeling ten behoeve van de interne afdelingsbeheersing, en op gereedheidsdata van orders ten behoeve van de goederenstroombeheersing. In praktijksituaties blijken deze twee prestatie-maten het meest relevant te zijn voor de evaluatie of beoordeling. Andere, meer geaggregeerde prestatie-maten

Samenvatting

Dit onderzoek richt zich op de korte termijn logistieke prestatie van job-shop-achtige produktie-afdelingen. Voor dit type afdeling is een methode ontwikkeld waarmee men i) de geleverde prestatie ten dele kwantitatief kan verklaren (de zgn. diagnose), ii) de geleverde logistieke prestatie kan beoordelen (de zgn. evaluatie), en iii) de logistieke prestatie kan voorspellen. In de ontwikkelde methode wordt zoveel mogelijk relevante informatie over de actuele toestand van de afdeling betrokken. De toevoeging van deze toestandsinformatie maakt de methode verschillend van bestaande methoden en modellen. Het uiteindelijke doel van de methode is om zowel de te leveren als de geleverde logistieke prestatie beter bespreekbaar te maken door gebruik te maken van objectieve data, waarbij een proces wordt gecreëerd waarin men voortdurend gericht is op het verbeteren van deze prestatie. Dit proces, het zogenaamde performance management proces, heeft de "continuous improvement" gedachte als uitgangspunt.

Het type produktie-afdeling dat we in dit onderzoek beschouwen krijgt orders te verwerken met een hoge variëteit in het aantal bewerkingen per order, bewerkingsvolgordes en bewerkingstijden. De beschikbare man-capaciteit is veelal in te zetten op verschillende soorten werkplekken en de vrijgave van nieuwe orders aan de afdeling vertoont een grote diversiteit in zowel aantal als soort. Bovendien kunnen er in dit soort omgevingen vele verstoringen optreden die invloed hebben op de logistieke prestatie, zoals machine-storingen, afwezigheid van personeel door ziekte, een tekort aan uitgangsmaterialen en een tekort aan produktspecifieke gereedschappen. Deze complexiteit maakt het voor het afdelingsmanagement moeilijk om een uitspraak te doen over welke prestatie men zal realiseren. Ook het beoordelen van de geleverde prestatie is moeilijk. Kwalitatief kunnen vaak oorzaken gevonden worden voor afwijkingen ten opzichte van de verwachte prestatie, maar een kwantitatieve toets wordt zelden of nooit uitgevoerd. Men heeft op dit moment bovendien geen instrument ter beschikking dat aangeeft waar in een afdeling mogelijkheden voor prestatie-verbetering te vinden zijn.

Dit onderzoek is gericht op ontwikkelen van instrumenten voor de beoordeling en diagnose van de totale output (hoeveelheid verzet werk in uren) van een produktie-afdeling ten behoeve van de interne afdelingsbeheersing, en op gereedheidsdata van orders ten behoeve van de goederenstroombeheersing. In praktijksituaties blijken deze twee prestatie-maten het meest relevant te zijn voor de evaluatie of beoordeling. Andere, meer geaggregeerde prestatie-maten
zoals gemiddelde orderdoorlooptijd en leverbetrouwbaarheid zijn door combinatie van bovenstaande prestatie-maten met bekende order- en capaciteitsgegevens direct afleidbaar.

Het evalueren van de geleverde logistieke prestatie houdt in dat deze prestatie vergeleken wordt met een prestatie-norm. Kenmerk is dat deze prestatie-norm bij de in de praktijk gebruikelijke werkwijzen betrekking heeft op een wat langere periode (van enkele weken tot enkele maanden). De werkelijk geleverde logistieke prestatie door een produktie-afdeling wordt echter vaak op kortere termijn (bijv. op weekbasis) gemeten, gerapporteerd en geanalyseerd. Deze werkelijke prestatie vertoont vaak grote fluctuaties op de korte termijn, enerzijds veroorzaakt door de vele verstoringen die zich kunnen voordoen en anderzijds door de dynamiek in de bedrijfsprocessen zelf. Als de prestatie-norm die voor een langere periode is opgesteld reëel is, dan zal deze norm op de lange duur gemiddeld genomen gerealiseerd worden. Dit zal echter niet het geval zijn voor elke meet- of rapportageperiode. Dit betekent dat de langere termijn norm niet altijd even realistisch is voor de korte termijn prestatie en daarmee dus ook niet geschikt voor de beoordeling van deze op korte termijn gerealiseerde prestatie. Om het bovenstaande probleem te voorkomen dient de frequentie van de normstelling gelijk te zijn aan de meet- en rapportagefrequentie van de geleverde prestatie. Voor de beschouwde produktie-afdelingen betekent dit dat er prestatie-normen voor de korte termijn nodig zijn. Voor het genereren van deze korte termijn normen hebben we een voorspellingsmodel ontwikkeld dat voor de korte termijn een meer realistische prestatie voorspelt. Dit voorspellingsmodel maakt gebruik van informatie over het actuele onderhanden werk, het op korte termijn verwachte werkaanbod en de op korte termijn verwachte beschikbare capaciteit. Deze drie variabelen samen vormen de zogenaamde actuele toestand van een produktie-afdeling. Het voorspellingsmodel is dus toestands-afhankelijk gemaakt, hetgeen het belangrijkste verschil is met bestaande voorspellingsmodellen in de produktie-logistiek. Naast de bovengenoemde toestandsinformatie vormen twee belangrijke gedragslijnen de basis van de ontwikkelde toestandsafhankelijke voorspellingsregels voor de output en de ordergereedheidsdata: de order vrijgave regel en de verwerkingsvolgorde regel.

Met simulatie-experimenten is de invloed van het gebruik van toestandsinformatie op de voorspelkwaliteit nagegaan ten opzichte van het gebruik van toestandsonafhankelijke voorspellingen. Voor al de onderzochte situaties blijkt het toevoegen van toestandsinformatie een positief effect te hebben op de voorspelkwaliteit van de prestatie. De sterkte van dit effect is daarbij afhankelijk van de gemiddelde bezettingsgraad en van het aantal werkplekken in een afdeling. Bij een relatief hoge bezettingsgraad levert het meenemen van toestandsinformatie relatief weinig verbeteringen op in de output-voorspellingen, maar relatief veel in de ordergereedheidsdata voorspellingen. Bij een lage bezettingsgraad is het
omgekeerde het geval. Verder heeft het toestandsafhankelijke voorspellen van de output meer 
zin naarmate er meer werkplekken in een afdeling zijn. Ook hier geldt het omgekeerde voor 
de ondergereedheidsdata voorspellingen. De algemene conclusie volgend uit het simulatie-
onderzoek is dat het toevoegen van toestandsinformatie de voorspelijkwaltie verbetert. De 
winst in voorspelijkwaltie is het laagst voor produktie-afdelingen met een hoge 
bezettingsgraad en weining werkplekken.

De methode van prestatie-evaluatie en diagnose werkt nu als volgt. Het voorspellingsmodel 
genereert per meetperiode een verwachte prestatie op grond van het actuele onderhanden 
werk, de verwachte beschikbare capaciteit en het verwachte werkbaar. Door de 
verwachtingen ten aanzien van de capaciteit en het werkbaar te variëren, kan het 
voorspellingsmodel gebruikt worden als "what-if" instrument, zodat de verwachte prestatie 
verhoogd kan worden door bijvoorbeeld een andere beslissing over de allocatie van de 
verwachte beschikbare capaciteit. Deze zogenaamde pre-voorspelde prestatie dient beschouwd 
met worden als een realistische prestatie-norm gegeven het actuele onderhanden werk en de 
verwachte capaciteit en het verwachte werkbaar. Als de actuele prestatie is gemeten en 
gerapporteerd, kan deze bij de prestatie-evaluatie vergeleken worden met de pre-voorspelde 
prestatie. Na de prestatie-evaluatie dient een prestatie-diagnose uitgevoerd te worden. Dit 
moet niet alleen gebeuren in situaties waarin een verschil is geconstateerd tussen de 
voorspelde en gerealiseerde prestatie, maar ook in situaties waarin geen verschillen zijn 
geconstateerd. Het kan namelijk toeval zijn dat voorspelling en werkelijkheid samenvallen.

In de diagnose wordt de gerealiseerde prestatie op een objectieve en kwantitatieve manier 
getracht te verklaren. Hiertoe wordt opnieuw het voorspellingsmodel gebruikt, waarmee nu 
een zogenaamde post-voorspelling wordt uitgevoerd. Dit is een voorspelling achteraf, met als 
input het oorspronkelijke onderhanden werk en de realisaties i.p.v. de verwachtingen van de 
in het model opgenomen variabelen. De post-voorspelde prestatie is dus de prestatie die 
achteraf gezien bereikt had kunnen worden. Dit betekent dat een vergelijking tussen de 
werkelijk gerealiseerde prestatie en de post-voorspelde prestatie in feite de enige juiste manier 
is om een objectieve prestatie-evaluatie uit te voeren. Het mogelijke verschil tussen de pre-
voorspelde prestatie en de werkelijke prestatie is dan te beschouwen als een maat voor de 
kwaliteit van het voorspellingsmodel, en het verschil tussen de pre-voorspelde prestatie en de 
post-voorspelde prestatie geeft weer wat de kwaliteit is van de schattingen van de input-
variabelen. Een eventueel waargenomen resterende verschil tussen de post-voorspelde en 
gerealiseerde prestatie vormt het uiteindelijke onverklaarbare deel van de prestatie. Om dit 
verschil te kunnen verklaren, zal een verdere diagnose uitgevoerd moeten worden. In 
tegenstelling tot de eerste, kwantitatieve diagnose, is deze verdere diagnose in eerste instantie 
kwalitatief van aard. Op deze kwalitatieve diagnose wordt in dit onderzoek verder niet ingegaan.

Samenvatting

155
De ontwikkelde methode voor de evaluatie en diagnose van de prestatie is in twee praktijksituaties getoetst op de praktische bruikbaarheid. De voorspellingsregels in het bij de methode behorende voorspellingsmodel zijn daartoe aangepast aan de specifieke afdelingskarakteristieken. In de praktijksituaties bleek de methode een duidelijke toegevoegde waarde te hebben. Ten eerste kon met behulp van het model afwegingen maken bij het alloceren van de beschikbare operator-capaciteit, hetgeen het inzicht verhoogde in de relaties tussen het onderhouden werk, de capaciteit, het werk aanbod, de output en de ordergereedheidsdata. Ten tweede werd de te realiseren prestatie voor de korte termijn beter bespreekbaar in termen van hoe de zaken georganiseerd dienden te worden om deze prevoorspelde prestatie waar te kunnen maken. Tenslotte kwam uit de diagnoses naar voren welke werkplekken in de afdelingen structureel problemen ondervinden bij het realiseren van de norm-output. Met behulp van het voorspellingsmodel kon deze invloed voor het eerst kwantitatief worden gemaakt. Door de diagnose kon de gerealiseerde prestatie voor een groot deel objectief verklaard worden en konden discussies zich richten op het zoeken naar andere verklaringen dan verschillen tussen verwachte en werkelijke capaciteit en werk aanbod. De verwachte verbeteringen ten aanzien van de voorspelskwaliteit werden alleen waargenomen bij de output voorspellingen; de toestandsafhankelijke output-prentellingen waren beduidend beter dan de op basis van MRP bepaalde output-voorspellingen. De toestandsafhankelijke voorspellingen van de ordergereedheidsdata bleken beter te zijn dan de door MRP geplande gereedheidsdata, maar ze presteerden "slechts" even goed als de door de planners gehanteerde vuistregels.

Ten aanzien van de implementatie van de ontwikkelde methode voor prestatie-evaluatie en diagnose moet aan een aantal belangrijke voorwaarden voldaan zijn om er zoveel mogelijk profijt van te hebben. De meest belangrijke voorwaarden zijn:

- Een produktie-afdeling moet redelijk onafhankelijk zijn van andere produktie-afdelingen met betrekking tot echterevenvolgende uit te voeren bewerkingen in de routingen van de produktie-orders.
- Doelstellingen voor afdelings- en goederenstroombeheersing dienen helder te zijn en dienen een duidelijke prioriteitsvolgorde te hebben. De verantwoordelijkheden en bevoegdheden van de personen die invloed op deze prestatie kunnen hebben dienen hierbij aan te sluiten.
- Degenen die een bepaalde prestatie kunnen beïnvloeden, dienen in groepen zowel de verwachte als de gerealiseerde prestatie te bespreken om de effecten van elkaars beslissingen beter te leren kennen en om het zoeken naar mogelijkheden voor verdere prestatie-verbeteringen effectiever te maken.
De prestatie-doelen moeten betrekking hebben op de termijn waarop de prestatie-evaluatie betrekking heeft.

Algemeen gesteld biedt de ontwikkelde evaluatie- en diagnose-methode ten aanzien van afdelingsbeheersing de volgende mogelijkheden. Ten eerste kunnen what-if analyses met betrekking tot werkordervrijgave-beslissingen en capaciteitsallocatie worden uitgevoerd. De voorspelde output voor de korte termijn dient daarbij tot discussiemateriaal voor de invulling van hoe men deze prestatie kan realiseren. Ten tweede kan met behulp van de methode aangegeven worden op welke plaatsen in de afdeling er relatief minder output geleverd wordt en hoe groot deze afwijking is. Hiermee kan het afdelingsmanagement zich direct concentreren op de gevoelige plekken. Voor goederenstroombeheersing biedt het voorspellen van de gereedheidsdata van orders de mogelijkheid te anticiperen op orders die naar verwachting te laat komen. Deze informatie kan leiden tot een andere prioriteitsstelling (in overleg met afdelingsmanagement en/of klant), en kan tenslotte gebruikt worden als communicatie-middel met de (interne of externe) klant.

Samenvatting
The author of this dissertation was born on December 10, 1966 in Breda. In 1985 he received his high school diploma from the "Sint Joriscollege" in Eindhoven, after which he started his study Industrial Engineering and Management Science at Eindhoven University of Technology. He received his Master's Degree in 1990 after a research project at Philips Nijmegen concerning the development of a decision support system for the operator allocation decision function. After his graduation he developed a new concept for production control for Ramaer Connection Technology in Helmond. Then he fulfilled his military service. Since 1991, he conducted research at the Graduate School of Industrial Engineering and Management Science, Eindhoven University of Technology, concerning the development of instruments for performance evaluation and diagnosis in complex production departments. The project was supervised by prof.dr. Will Bertrand and prof.dr. Jacques Theeuwes. This dissertation concludes his research. During his research he was involved with projects in industry. From January 1996 the author is employed at GPT Axxicon as project leader production control and goods flow control.
STELLINGEN

behorende bij het proefschrift

PERFORMANCE MANAGEMENT
IN MANUFACTURING

A method for short term
performance evaluation and diagnosis

van

Paul Stoop
I

De uitdrukking "Meten is weten" is feitelijk onjuist; er is pas sprake van "weten" als er na het meten een evaluatie en diagnose zijn gevolgd.

II

Voor het bepalen van prestatie-normen voor de korte termijn in complexe produktie-afdelingen moet toestandsinformatie gebruikt worden.

_Dit proefschrift; hoofdstuk 3_

III

Naarmate de bezettingsgraad van een produktie-afdeling hoger is, wordt het effect van het gebruik van toestandsinformatie op de voorspelkwaliteit van de output-voorspellingen kleiner. Het effect van het gebruik van toestandsinformatie op de voorspelkwaliteit van ordergereedheidsdata wordt daarentegen groter.

_Dit proefschrift; hoofdstuk 5_

IV

Naarmate het aantal werkplekken in een produktie-afdeling groter is, wordt het effect van het gebruik van toestandsinformatie op de voorspelkwaliteit van voorspellingen van ordergereedheidsdata kleiner. Het effect van het gebruik van toestandsinformatie op de voorspelkwaliteit van de output-voorspellingen wordt daarentegen groter.

_Dit proefschrift; hoofdstuk 5_

V

In een dynamische omgeving moet niet alleen toestandsinformatie gebruikt worden bij het beoordelen van de logistieke afdelingsprestatie op de korte termijn, maar ook de werkelijke omgevingsinvloeden in die periode.

_Dit proefschrift; hoofdstuk 6_
VI

Indien wetenschappers aan de implementatie en het daadwerkelijk gebruik van modellen evenveel aandacht zouden besteden als aan de ontwikkeling, zou niet alleen de praktijk meer baat vinden bij deze modellen maar ook de modelbouw.

VII

Een cursus "didactische vaardigheden" dient niet alleen voor aio's/oio's verplicht te zijn, maar ook voor onderwijzend personeel met een vast dienstverband.

VIII

De hoge snelheid van nieuwe generaties computers leidt tot langere doorlooptijden van onderzoeken die uitgevoerd worden met behulp van simulatie-experimenten.

IX

Televisie-programma's over klussen in en rondom huis schetsen voor de doehet-zelver een veel te positief beeld over de aard en duur van de werkzaamheden.

X

Om ergernis van bellers te voorkomen, moet het gebruik van muziek in de telefoon-wachtstand bij bedrijven en instellingen sterk afgeraden worden.

XI

Het gebruik van puntschattingen door gynaecologen leidt meestal tot verrassingen bij ouders.