Flexible versus robust lot-scheduling subject to random production yield and deterministic dynamic demand

Stefan Helber, Karl Inderfurth, Florian Sahling and Katja Schimmelpfeng

Department of Production Management, Leibniz Universität Hannover, Hannover, Germany; Faculty of Economics and Management, Otto-von-Guericke Universität Magdeburg, Magdeburg, Germany; Department of Production and Industrial Management, Technische Universität Chemnitz, Chemnitz, Germany; Department of Procurement and Production, Universität Hohenheim, Stuttgart, Germany

ABSTRACT

We consider the problem of scheduling production lots for multiple products competing for a common production resource that processes the product units serially. The demand for each product and period is assumed to be known with certainty, but the yield per production lot is random as the production process can reach an out-of-control state while processing each single product unit of a lot. A service-level constraint is used to limit the backlog in the presence of this yield uncertainty. We address the question of how to determine static production lots and how to schedule these lots over the discrete periods of a finite planning horizon. The scheduling problem is characterized by a trade-off between the cost of holding inventory and the cost of overtime, whereas the production output is uncertain. For this purpose, we develop a rigid and robust planning approach and two flexible heuristic scheduling approaches. In an extensive numerical study, we compare the different approaches to assess the cost of operating according to a robust plan as opposed to a flexible policy.

1. Introduction

In this article, we consider a multi-product lot-sizing and scheduling problem for a single production resource with limited capacity that can be extended by overtime. The different products compete for the scarce capacity of this resource. Furthermore, they face a given deterministic demand over multiple discrete periods. Each single production run causes a fixed setup cost and time. The different product units within a production lot of a given product are processed serially. While processing a product unit, the production process can randomly enter an out-of-control state, for example, due to tool wear or pipe congestion. If this occurs, the currently processed product unit and all subsequently processed units of that lot will be defective. Furthermore, we assume that quality inspection does not occur before the end of a production run; consequently, such defects can only be detected after production of the entire lot is complete. As a result, the yield of the lot is random, and it can be advisable to limit the lot size to avoid a high and costly scrap rate. For this reason, it might be advantageous to produce multiple distinct lots of a product during a given period. The respective lot size primarily must balance the fixed setup costs and variable scrap and holding costs for each product. In addition to a lot-sizing problem, this situation obviously leads to a difficult lot-scheduling problem, in which the number of production lots per product and period must be determined. This problem is characterized by a trade-off between the cost of holding inventory and the cost of using overtime.

We can achieve this balance in a fully flexible manner such that the decision to produce a lot of a particular product depends on the yield of all previously produced lots of that product. Such an approach exhibits an extremely high degree of flexibility. It also leads (on average) to the lowest possible costs. The downside of this approach is that the capacity requirements per period are unpredictable, as a stable schedule does not exist. The alternative solution approach is to determine a fixed and robust schedule ex ante. This approach avoids planning uncertainty but has the downside of potentially higher costs and failing to meet a required service level with certainty. The first approach can, in principle, be formulated as a Stochastic Dynamic Program (SDP). The solution of the SDP leads to an optimal and fully flexible policy, in which the development of the schedule reflects the realized yield outcomes. However, such an SDP suffers from the curse of dimensionality for all but the smallest problem instances. Thus, it might be advisable to develop heuristic lot-scheduling policies that react flexibly to yield realizations and are easier to implement than a full-blown SDP approach.

The objective of this article is to develop, compare, and evaluate an inflexible robust planning approach and flexible scheduling methods based on exact SDP or heuristic approaches. To this end, we report results from an extensive and systematic numerical study. The remainder of this article is structured as follows. In Section 2, we describe the problem in more detail, outline the different approaches to deciding about lot-scheduling, and comment on the related literature. For the different lot-scheduling
approaches, we present formal models and/or algorithms in Section 3. The different approaches are compared in an extensive numerical study in Section 4. Section 5 summarizes the main results and suggests directions for further research.

2. Problem description

2.1. Separation of lot sizing and scheduling

Due to the specific production yield model treated in this article, we employ a two-step separation approach in which the lot-sizing problem is first solved individually for each product before these predetermined lots are then used as input for solving the multi-product lot-scheduling problem. For practical reasons, we use a static lot-sizing approach under average demand conditions similar to the Economic Order Quantity (EOQ) model to determine the constant lot sizes. In this context, the exact properties of the yield process must be considered when the lot sizes are determined.

2.2. Production process, random yield, and lot-sizing

The yield of a production process is sometimes not entirely predictable. Hence, many authors have proposed to explicitly consider yield uncertainty in production planning and scheduling approaches; see, e.g., the surveys presented in Yano and Lee (1995) and Grosfeld-Nir and Gerchak (2004). In this article, we address production processes in which product units are processed serially by a machine or, more generally, a production resource. As the resource requires an initial setup operation, the production is organized in lots of size \( q \). Setups are assumed to be sequence independent to facilitate the analysis. In such a setting, there can be different mechanisms that cause yield uncertainty. These different mechanisms are reflected in three fundamentally different product-specific types of yield models that are also described in the above-mentioned survey papers. We generally assume that yield processes are independent among products if we consider multiple product types.

The first of these yield models, the Binomial Yield Model (BIYM), considers the case that processing a single product unit within a production lot leads to a conforming unit with probability \( p \), independent of the other units within that lot. This model hence treats the case in which there is some randomness inherent in the individual production operation itself, given that the production attempts for the different product units are probabilistically independent of each other. It also reflects situations where product defects occur due to independently distributed material faults. A single operation on a single product unit can therefore be modeled as a Bernoulli experiment with success probability \( p \). Hence, the random yield \( Y = \sum_{n=1}^{d} V_n \) of a lot of size \( q \) is the sum of \( q \) independent and identically distributed (i.i.d.) Bernoulli distributed random variables \( V_n \) with \( \text{Prob}[V_n = 1] = p \) and \( \text{Prob}[V_n = 0] = 1 - p \); i.e., the random yield \( Y \) follows a binomial distribution with parameters \( p \) and \( q \). In this model, the expected yield \( E[Y] = p \times q \) is proportional to the lot size \( q \). Thus, the fraction of the conforming parts is not affected by the lot size.

The Stochastically Proportional Yield Model (SPYM), in contrast, reflects a different mechanism in which a common cause, such as an important environmental condition, determines the fraction \( 0 \leq Z \leq 1 \) of conforming product units within a production lot, irrespective of the size of the lot. This random fraction \( Z \) obeys a probability distribution with mean \( \mu_Z \) and variance \( \sigma_Z^2 \). This yield model appears to be very relevant in the context of parallel (batch) production; for example, in farming (see Kazaz (2004)). Here, the common cause(s) could be weather conditions, vermin, or animal infections. Other examples relate to food processing or biochemical production. As in the binomial yield model, the expected yield \( E[Y] = \mu_Z \times q \) is proportional to the lot or batch size \( q \), and the fraction of conforming parts is again unaffected by the lot size.

In this article, however, we specifically consider the third and very different case of a so-called Geometric Yield Model (GYM; see Anily et al. (2002) for a detailed description), which is only relevant to serial production processes. The basic idea is that non-conforming product units are produced because a production process randomly enters an “out-of-control” state due, for example, to tool wear and tear, pipe congestion, or other technical problems that require intervention. In such a situation, it is, of course, desirable to monitor the process closely to observe whether it is still in control and to stop it otherwise. However, as described in detail in Porteus (1986), there may be cases in which—due to organizational, economic, or technological reasons—the yield can only be determined after the production of the entire lot has been completed. In this situation, the expected yield \( E[Y] \) after production is not proportional to the lot size, and the marginal production cost is no longer constant. Thus, both lot-sizing and lot-scheduling become difficult and important problems.

To study this type of situation in more detail, we assume that the production process is always initially in control after setting up the machine. While processing product unit \( n \) of the current lot, the process remains in control with probability \( p_n \), such that the \( n \)th processed product unit meets the specifications. With probability \( 1 - p_n \), however, the process changes from the “in-control” to the “out-of-control” state while product unit \( n \) of the lot is being processed such that the \( n \)th and all following product units of this particular lot are defective. According to Anily et al. (2002), the yield probabilities of such a generalized truncated geometric distribution are given by

\[
\text{Prob}[Y = y] = \begin{cases} 
(1 - \alpha + 1) \prod_{n=1}^{y-1} p_n, & y = 0, 1, \ldots, q - 1, \\
\prod_{n=1}^{y} p_n, & y = q.
\end{cases}
\]

(1)

Such a situation is depicted in Figure 1 for three consecutive lots of size \( q = 7 \) for a single given product.

The process enters an out-of-control state while processing the fifth product unit of the first lot and the third product unit of the third lot. It remains in control during the entire production of the second lot. The yields of the first, second, and third lots are four, seven, and two product units, respectively. There may be cases where the probability \( p_n \) of the process remaining in control depends on the number \( n \) of the currently being processed product unit of the current lot. The reason might be progressive tool wear. To model this case, we consider a geometrically decreasing success probability,

\[
p_n = \alpha \times p_{n-1} = \alpha^{n-1} p,
\]

(2)
we immediately find

\[ \prod_{n=1}^{r} p_n = \prod_{n=1}^{y-1} \alpha^n \times \prod_{n=1}^{y} p = \alpha^{\sum_{n=1}^{r} n} \times p^r = \alpha^{y(y-1)/2} \times p^r, \]

which, inserted into Equation (1), leads to the following probability function:

\[
\text{Prob}[Y = y] = \begin{cases} 
(1 - \alpha^y p) \times \alpha^{y(y-1)/2} \times p^r, & y = 0, 1, \ldots, q - 1, \\
\alpha^{y(q-1)/2} \times p^r, & y = q.
\end{cases}
\]

Note that for the case of \( \alpha = 1 \) this reduces to the well-known probability function of the standard Interrupted Geometric Yield (IGY) model; see, e.g., Grosfeld-Nir and Gerchak (2004). In Figure 2, we present as an example the Interrupted Geometric (IG) yield distribution for a success parameter \( p = 0.8 \), \( \alpha = 1 \), and lot size \( q = 7 \).

If the yield \( Y \) of a lot of size \( q \) is a random quantity and follows this generalized IG distribution with initial success parameter \( p \) and decay parameter \( \alpha \), we find after a few algebraic steps for the first two moments of the yield:

\[
E[Y] = \sum_{y=1}^{q} y \times \text{Prob}[Y = y] = \sum_{y=1}^{q} \alpha^{y(y-1)/2} \times p^r \tag{5}
\]

and

\[
E[Y^2] = \sum_{y=1}^{q} y^2 \times \text{Prob}[Y = y] = \sum_{y=1}^{q} (2y - 1)\alpha^{y(y-1)/2} \times p^r. \tag{6}
\]

Thus, the expected yield \( E[Y] \) increases with the lot size \( q \) but the marginal increase decreases with \( q \), as shown in the examples given in Figure 3 for different combinations of initial probability \( p \) of production success and decay factor \( \alpha \). We can conclude that the expected unit production cost per conforming part increases with \( q \). It is also obvious that for large lot sizes, a small decrease in the success probability \( p \) leads to a remarkable increase in the fraction of defective parts.

To study the impact of IG yield on the lot-sizing problem, the EOQ model provides a good starting point. As in the traditional EOQ model, we assume that demand is continuous and production is instantaneous, but the yield now obeys an IG distribution.

Using the renewal reward theorem, the standard EOQ approach has already been extended to random yield environments in Silver (1976), where it is shown that for the expected cost per unit of time \( acr \), the following equation holds (using the notation given in Table 1):

\[
acr = \left(\frac{sc + pc \times q + \frac{hc}{2E[Y^2]}E[Y^2]}{E[Y]}\right) \times \frac{d}{E[Y]} \tag{7}
\]

Here, the first moment \( E[Y] \) depends on the respective random yield model, and the same holds for the second moment, which can simply be determined via the definition of the variance from \( E[Y^2] = (E[Y])^2 + \text{Var}[Y] \). By inserting the yield model’s specific moments, the objective function in Equation (7) can easily be exploited to determine the optimal lot size \( q^* \). For

| Notation for the stationary model. |
|-----------------------------------|
| \( \alpha \) | decay factor for the production success probability |
| \( acr \) | average (expected) cost rate (monetary units per time unit) |
| \( d \) | demand rate (quantity units per time unit) |
| \( hc \) | holding cost (monetary units per quantity unit of inventory and time unit) |
| \( p \) | probability of production success for the initial (first) operation of a lot |
| \( pc \) | production cost (monetary units per quantity unit of production) |
| \( sc \) | setup cost (monetary units) |
| \( q \) | lot size (quantity units) |
| \( Y \) | random yield of a lot (quantity units) |
| \( Z \) | random fraction of conforming parts |
the SPYM and BIYM, this results in closed-form expressions for \( q^* \), which are simple extensions of the EOQ formula:

\[
EOQ = \sqrt{\frac{2 \times sc \times d}{hc}}.
\]

From Silver (1976) and Shih (1980), we find that for a stochastically proportional yield with yield rate parameters \( \mu_Z \) and \( \sigma_Z^2 \), the optimal lot size is

\[
q^* = \frac{EOQ}{\mu_Z + \sigma_Z^2}.
\]

From Mazzola et al. (1987), it can be observed that in the case of a binomially distributed yield with success parameter \( p \), the optimal lot size is

\[
q^* = \frac{EOQ}{p}.
\]

For the Interrupted Geometric Yield Model (IGYM), on which we focus in this article, a closed-form expression cannot be derived. Therefore, a numerical search procedure must be used to calculate the cost-minimizing lot size from the cost function in Equation (7) after inserting \( E[Y] \) and \( E[Y^2] \) using expressions from Equations (5) and (6). We provide the proof that the cost function \( acr(p) \) is quasi-convex for our generalized IGYM and, hence, that this search procedure always finds a global cost minimum in Section 2 of the electronic supplement to this article.

Unlike the cases of proportional (8) and binomial (9) yield, in the case of an IG yield, the unit production cost \( pc \) is relevant for the optimal lot size \( q^* \), as the expected unit cost of conforming products depends on the lot size \( q \). In Figure 4, we show examples of IGYM cost functions from Equation (7) for setup cost \( sc = 500 \) monetary units, production cost \( pc = 10 \) monetary units per product unit, holding cost \( hc = 1 \) monetary units per product unit and time unit, and demand rate \( d = 100 \) product units per time unit for different combinations of the initial probability \( p \) of production success and the success probability decay factor \( \alpha \).

The graphs in Figure 4 clearly indicate that under an IG yield, the success probability \( p \) has a significant effect on both the cost per period (as well as the cost per conforming product unit) and the cost-minimizing lot size. To avoid high costs of producing scrap, it may be economically efficient to operate with lot sizes \( q \) that are substantially smaller than those in a situation with perfect yield.

2.3. Discrete periods, deterministic demand, limited capacity, backlog, and the \( \delta \)-service level

In many real-world situations, planners operate with discrete time periods \( t \in T = \{1, 2, \ldots, T\} \), such as days, weeks, or months, and both demand and production quantities are assigned to these discrete time periods. Frequently, different product types compete for a capacity-restricted production resource, such as a machine. Then, the production schedules for the different products must be coordinated, leading to a multi-product, multi-period lot-sizing and scheduling problem. If the demand is dynamic and the production process cannot go “out of control,” a cost-minimizing production schedule will typically exhibit time-varying lot sizes and at most one lot of any given product will be scheduled for a given period.

However, in a random yield case, especially under the IGYM conditions introduced in Section 2.2, it may be beneficial to schedule multiple (small) lots of a given product during a period to avoid the risk of producing too much scrap when operating with a larger (joint) lot. Remember that according to Figure 3, the expected yield is a sub-linear function of the lot size. In this situation, the cost of producing scrap can have a higher impact on the cost-minimizing lot size than the typical trade-off between setup and holding costs. In this context, as mentioned above, we assume that a static lot size \( q_k \) for each product \( k \in K = \{1, 2, \ldots, K\} \) is determined in a first step to minimize the expected average cost rate (7) in a stationary demand setting based on long-term average demand rates.

We further assume that the demand \( d_{kt} \) for product \( k \) in period \( t \) is deterministic and given. The production requires a single machine with regular period capacity \( c_t \) and overtime \( o_t \). Each lot requires a setup time \( ts_k \) per lot of product \( k \) and a processing time \( tp_k \times q_k \) that is proportional to the lot size \( q_k \), which is assumed to be given, as explained above. If we denote the number of lots of product \( k \) scheduled in period \( t \) as \( x_{kt} \in \{0, 1, 2, 3, \ldots\} \), the capacity restriction

\[
\sum_{k=1}^{K} (ts_k + tp_k \times q_k) \times x_{kt} \leq c_t + o_t, \quad t \in T,
\]

implicitly determines the required amount of overtime \( o_t \). If regular capacity is limited, overtime capacity is costly, and yield is uncertain, it may be unavoidable or economically advisable to not meet the demand for \( d_{kt} \) for each product in each period. If we consider a backorder case, the backlog may be carried over from one period to the next. We denote by \( L_k \) the number of lots of product \( k \) produced in period \( t \), by \( y_{kt} \) the yield realization of lot \( l = 1, \ldots, L_k \), by \( ph_{kt} \) the physical inventory, and by \( bl_{kt} \) the backlog at the end of period \( t \). As it is not optimal to have both physical inventory and backlog for a given combination of product and period simultaneously—i.e., \( ph_{kt} \times bl_{kt} = 0 \)—the following balance constraints must hold for any given realization of the random yield \( y_{kt} \):

\[
ph_{kt-1} - bl_{kt-1} + \sum_{l=1}^{L_k} y_{kt} - d_{kt} = ph_{kt} - bl_{kt}, \quad k \in K, t \in T.
\]
In the literature, several approaches have been proposed to limit or avoid this backlog and the corresponding delay in demand fulfillment. Frequently, in an $\alpha$-service level constraint, the probability of backordering (i.e., creating additional backlog) is limited for each period. In other approaches, the amount of backordering relative to the demand in a production cycle is limited via a $\beta$-service level; see, e.g., Tempelmeier and Hilger (2015). Other authors punish backlogs by charging a shortage cost, which is difficult to determine in practice.

Our approach in this article is to limit backlogs directly and in a joint constraint over the entire planning horizon. For this purpose, we enforce a hard constraint on backlogs to guarantee an (expected) service level. In the planning environment under consideration, it is likely that backlogs last for multiple periods, such that the length of an out-of-stock period should be captured by the service measure. According to Little’s law (see Little (1961)), the amount of backlogged demand is proportional to the waiting time. Therefore, we use the backlog-oriented $\delta$-service-level measure (for a justification and more detailed description of this service measure, see Helber et al. (2013a)) to limit the relative overall backlog via a constraint

$$\delta_k = 1 - \frac{\sum_{t=1}^{T} b_{kt}}{\sum_{t=1}^{T} (T - t + 1) \times d_{kt}} \geq \delta_{k}^{\min}, \quad k \in K,$$

(12)

over periods $t = 1, \ldots, T$. Note that $\delta_k = 1$ indicates that no backlog occurs for product $k$ at all, whereas $\delta_k = 0$ implies that the maximum possible backlog is built up because no demand for product $k$ is satisfied until the end of period $T$. A striking advantage of this measure is that unlike other commonly used time- and quantity-oriented service measures, such as the so-called $\gamma$-service level (see Schneider (1981)), this measure is normalized to the range $[0, 1]$.

### 2.4. Fundamental trade-off and alternative decision approaches

The scheduling problem treated in this article is to determine the number of lots of each product and period to be produced in the presence of dynamic demand, limited production capacity, yield uncertainty, and a backlog-oriented service level constraint. Both production costs and setup costs are, at least in the long run, driven by demand, which must be met, and by the respective lot size $q_k$ of product $k$, which is assumed to be predetermined, as described in Section 2.2. Thus, if in this setting, capacity constraints are tight and demand is highly variable, it may be advisable to build up inventory in the current period to avoid overtime in a later period. This is the fundamental trade-off that characterizes the lot-sizing problem studied in the remainder of this article. To solve this multi-product, multi-period lot-sizing problem subject to random yield, various approaches can be used to determine how and when the production schedule should be modified to react to production risks.

One approach is to develop an SDP (Puterman (2005)) that yields a situation-specific decision rule. This rule minimizes the expected cost related to setup, production, inventory, and overtime subject to a service-level constraint that limits the permissible backlog. For a given combination of inventory level, (cumulative) backlog, and remaining regular capacity, this decision rule determines which product (if any) must be produced next in the currently considered period, given the already known yield outcome of the previously scheduled and produced lots. Such an SDP approach leads (in principle) to the optimal decisions. We actually developed and coded such an approach, using essentially a standard backward induction. This approach is, for the sake of completeness, documented in Section 1 of the electronic supplement to this article. However, as expected, it suffers from the curse of dimensionality often encountered in SDPs (see, e.g., Powell (2011, pp. 112-113)); i.e., an explosion of the state space such that only the tiniest problem instances can be solved using this method.

A second and opposing approach is to develop an inflexible Robust Production Schedule (RPS), in which the yield uncertainty is explicitly considered and the expected backlog values are anticipated. The schedule is inflexible in the sense that it is executed irrespective of the yield realization of any particular lot. Such a schedule must be robust with respect to yield uncertainty; i.e., it has to meet a predefined service-level requirement. However, this is not performed for each sequence of yield realizations but only in an ex ante and on-average sense. In addition, such an inflexible and robust schedule is more costly than the flexible schedule yielded by the SDP approach mentioned above and documented in Section 1 of the electronic supplement to this article. However, a perfectly stable and foreseeable schedule could be extremely desirable from the perspective of material procurement and workforce scheduling. We develop such an approach in Section 3.1.

This robust planning approach can also serve as the foundation for flexible and adaptive heuristic scheduling policies. We developed and tested two such policies. The basic idea of the first policy is to interpret the outcome of the robust planning approach as setting minimum limits on net inventory and maximum limits on backlog and, in any period, to schedule further lots until all such limits are firmly respected. The basic idea of the second policy is to solve the planning model repeatedly in a rolling-planning sense, given the already realized yield outcome of the already produced lots. We describe the different approaches in more detail below.

### 2.5. Related literature

The literature regarding production planning under random yield and deterministic demand mainly concentrates on lot-sizing decisions. Concerning the general problem studied in this article, contributions closely related to the static EOQ lot-sizing problem were previously discussed in Section 2.2. In addition, research regarding the problem of Multiple Lot sizing in Production-to-Order (MLPO) environments is relevant. MLPO problems refer to the combined lot-sizing and lot-scheduling question of how often under yield uncertainty production runs should be initiated and how large the sizes of the respective lots should be to meet the demand of a specific period at the minimum expected cost. In this situation, lots that are too large can lead to unnecessary overproduction costs, whereas lots that are too small might result in multiple production runs with high setup costs. Based on the realized yields from past runs, the remaining demand is determined when the decision regarding a further production run must be made such that the optimal lot size might dynamically change from run to run. In Grosfeld-Nir
and Gerchak (2004), an excellent overview of relevant research papers is given. Typically, an SDP approach is used as the solution methodology for the MLPO problem, which enables analysis of the policy properties of different yield models and development of specific numerical solution approaches.

However, MLPO models differ from our problem class, in that they only refer to a single decision period in which the demand is usually assumed to be rigid; i.e., it has to be fulfilled completely. However, concerning our problem, we plan for multiple periods and can either produce temporarily ahead of demand to build up inventory or postpone demand fulfillment to later periods to balance the cost of overtime and holding inventory. In Hsu et al. (2009), a specific two-period MLPO with variable production time and lost sales is considered. In this approach, the demand is non-rigid, as in MLPO models with an externally limited number of production runs (see Guu and Zhang (2003)) or in models that address stochastic demand (see Wu et al. (2011) for one of the few examples of stochastic MLPO approaches). All of these contributions are restricted insofar as they refer only to one or two planning periods and do not account for multiple products in a single production stage. The multi-product MLPO approach presented in Grasman et al. (2008) is also restricted to a single period and only refers to the case of a binomially distributed yield.

Only a very limited number of publications have addressed general multi-period problems with random production yield. In Mazzola et al. (1987), several heuristics for dynamic lot-sizing were developed and compared, but only single-product cases with binomial yield and a single production run per period were covered. A multi-product problem with capacity limitation was investigated in Taleizadeh et al. (2010). However, in that contribution, the demand is assumed to be constant in time, only a stochastically proportional yield was considered, and production follows a simple common cycle policy. The problem structure closest to ours was investigated in Rajaram and Karmarkar (2002). Those authors considered a multi-product multi-period production problem with dynamic deterministic demand and stochastically proportional yield. However, production decisions were restricted in such a manner that in each period, only a single product type could be manufactured and only a single production run was feasible; consequently, lot-scheduling is not referred to in that problem context. The planning task was formulated as a stochastic dynamic optimization problem. To solve this problem, a decomposition approach and several heuristics were proposed.

Other contributions in the field of stochastic dynamic multi-product lot-sizing and scheduling problems refer to situations where the production yield is known in advance but the demand is uncertain. We refer to Aloulou et al. (2014) for a recent overview of stochastic lot-sizing models. Bookbinder and Tan (1988) presented a stochastic single-product uncapacitated lot-sizing problem subject to an $\alpha$-service level. Three fundamental strategies were discussed: first, the "static uncertainty" approach, in which production decisions—i.e., setup operations and production quantities—are fixed in advance before demand realizations are known; second, the "dynamic uncertainty" approach, in which both decisions for future periods are made once the period's demand has been realized; and third, the "static-dynamic uncertainty" approach, which combines these two approaches; i.e., in which setup decisions are determined in advance, whereas production quantities are determined based on known demand realizations.

The following overview is limited to stochastic capacitated lot-sizing problems (CLSPs). Brandimarte (2006) suggested a model formulation for a stochastic CLSP using scenario trees. This stochastic CLSP was reformulated as a simple plant location problem, and a fix-and-relax heuristic was applied. Tempelmeier and Herpers (2010) presented a stochastic variant of the CLSP subject to a cyclic $\beta$-service level. For the generation of a robust production plan, the ABC heuristic proposed by Maes and van Wassenhove (1986) was adapted. For the same problem, Tempelmeier (2011) proposed a column generation approach that significantly outperformed the so-called ABC$_{\beta}$ heuristic presented by Tempelmeier and Herpers (2010).

In Helber et al. (2013a), a generic nonlinear model formulation was presented for a stochastic CLSP subject to a backlog-oriented $\delta$-service level. To overcome nonlinear behavior, two approximation techniques, a sample average approach and piecewise linear functions, were presented. A fix-and-optimize heuristic was applied to solve both linearized stochastic CLSPs. In their numerical investigation, Helber et al. (2013a) showed that the approach based on piecewise linear functions substantially outperformed the sample average approach. Tempelmeier and Hilger (2015) investigated a stochastic CLSP subject to a $\beta$-service level. The authors also utilized piecewise linear functions suggested by Helber et al. (2013a) to approximate existing nonlinear behavior, and a variant of the fix-and-optimize heuristic was used. In Helber et al. (2013b), first ideas on robust—i.e., non-reactive—lot-sizing and scheduling are presented for the special case of the IGYM with non-decreasing success probability.

3. Lot-scheduling models and heuristic scheduling algorithms

3.1. Robust planning approach

We now describe the optimization model for the rigid and robust planning approach. It is intended for situations where it is desirable to operate with a stable (or rigid) production schedule that is executed as planned, irrespective of the yield outcome of the different lots. In a situation of yield uncertainty, such a schedule should be robust, as it anticipates the yield uncertainty and leads, at least on average, to a requested service level that limits the expected backlog. As mentioned before, we assume that for each product $k \in K$, a lot size $q_{kt}$ that minimizes the long-term average cost rate given in Equation (7) has already been computed. The remaining problem is then to determine the number $x_{kt}$ of lots of product $k$ to be scheduled in period $t$ such that the expected costs are minimized while a limit on the backlog is guaranteed. Note that this problem setting specifically reflects the generalized IGY model treated in Section 2.2. If a BIYM or a SPYM were used, one could reduce setup costs and/or times by only producing at most one single lot of a given product per period without reducing the expected combined yield, so that it would not be advisable to deliberately produce multiple lots as proposed by our approach for the IGY case. The number of lots per period reflects the demand dynamics.
We assume that from the *ex ante* perspective, for each product, the expected yield of the scheduled lots over the planning horizon $t = 1, \ldots, T$ must not fall below the total demand. In other words, the expected net inventory of the final period $T$ (expected physical inventory minus expected backlog) must not be negative as, in the long run, the demand must be met. If the demand must be met in the long run, and if the lot sizes are determined via a stationary model based on average demand rates, then both the production and setup costs cannot be affected in the long run by the lot-scheduling decisions studied in this section. However, both the cost of holding inventory and the overtime cost are affected by scheduling decisions and are considered in the objective function of the robust planning model developed below and in the SDP approach documented in the electronic supplement to this article.

We now explain how the expected backlog $ebl_{ktn}$ and the expected physical inventory $eph_{ktn}$ can be computed, given that $n$ lots of product $k$ have been scheduled and produced until period $t$. For this purpose, we first must determine the probability function $\text{Prob}[Y_k = y_k]$ for the combined yield of $n$ consecutive lots. As before, we assume that for product $k$, a discrete probability function $\text{Prob}[Y_k = y_k]$ is given for the yield of a *single* lot of size $q_k$. The respective probabilities depend on the specific yield model under consideration and are given in Equation (4) for the case of a generalized IG yield. We denote with $\text{Prob}[Y_k^{(n)} = y_k^{(n)}]$ the yield of $n$ consecutive lots with i.i.d. yield each. Then, we have $\text{Prob}[Y_k^{(1)} = y_k^{(1)}] = \text{Prob}[Y_k = y_k]$,

and the following recursive convolution equation holds for $n > 1$:

$$\text{Prob}
\left[
Y_k^{(n)} = y_k^{(n)}
\right]
= \sum_{y_k=0}^{q_k} \text{Prob}[Y_k = y_k] \times \text{Prob}
\left[
Y_k^{(n-1)} = y_k^{(n-1)} - y_k
\right].$$

(13)

This recursive convolution equation relates the probability function for the yield of $n$ consecutive lots to the probability function for the yield of $n - 1$ consecutive lots. It is hence (numerically) straightforward to determine the probability function $\text{Prob}[Y_k^{(n)} = y_k^{(n)}]$ of the yield of $n$ consecutive lots via a numerical convolution for $n = 2, 3, \ldots$. Given those values of the yield distribution for $n$ consecutive lots in addition to the cumulative demand $cd_t = \sum_{t=1}^{t} d_{kt}$ until period $t$, the expected value of the physical inventory of product $k$ at the end of period $t$ after the production of $n$ lots is

$$eph_{ktn} = \sum_{y_k^{(n)}=cd_t}^{y_k^{(n)}+q_k} \left(y_k^{(n)} - cd_t\right) \times \text{Prob}
\left[
Y_k^{(n)} = y_k^{(n)}
\right].$$

(14)

and the corresponding expected backlog is

$$ebl_{ktn} = \sum_{y_k^{(n)}=0}^{y_k^{(n)}+cd_t} \left(cd_t - y_k^{(n)}\right) \times \text{Prob}
\left[
Y_k^{(n)} = y_k^{(n)}
\right].$$

(15)

The expected values $ebl_{ktn}$ and $eph_{ktn}$ are parameters of the optimization model, to be determined in a preprocessing step.

### Table 2. Notation of the robust planning model.

| Indices | Deterministic parameters | Decision variables |
|---------|--------------------------|-------------------|
| $k=1, \ldots, K$ | products | $n=0, \ldots, N$ | produced lots, $N$ sufficiently large to meet the demand |
| $t=1, \ldots, T$ | periods | $\delta_{k}^{\text{min}}$ | required $\delta$-service level of product $k$ |

Using the notation presented in Table 2, the problem of determining a cost-minimizing RPS can now be stated as follows:

**RPS model:**

Minimize

$$\sum_{k=1}^{K} \sum_{t=1}^{T} \sum_{n=0}^{N} h_k \times eph_{ktn} \times v_{ktn} + \sum_{t=1}^{T} oc \times \alpha_t,$$

subject to

$$\sum_{k=1}^{K} (ts_k + t ps_k \times q_k) \times x_{kt} \leq c_t + \alpha_t, \quad t \in T,$$

$$\sum_{t=1}^{T} x_{kt} = \sum_{n=0}^{N} n \times v_{ktn}, \quad k \in K, \quad t \in T,$$

$$\sum_{n=0}^{N} v_{ktn} = 1, \quad k \in K, \quad t \in T,$$

$$\sum_{t=1}^{T} \sum_{n=0}^{N} ebl_{ktn} \times v_{ktn} \leq (1 - \delta_{k}^{\text{min}}) \times \sum_{t=1}^{T} (T - t + 1) \times d_{kt}, \quad k \in K,$$

$$\sum_{t=1}^{T} \sum_{n=0}^{N} (eph_{ktn} - ebl_{ktn}) \times v_{ktn} \geq 0, \quad k \in K.$$
for the expected net inventory and, furthermore, a (required) level for the expected backlog. Note that the latter is consistent with the backlog-oriented service-level constraint only in the sense that this service restriction is fulfilled a priori on an average basis. In a practical planning situation, however, the realized service level can exhibit major deviations from its prescribed value as, with the RPS approach, no adjustments to production are made if major deviations between the expected and realized yields occur. In practice, major deviations might not be acceptable; thus, some type of rescheduling flexibility may be desired to guarantee that the required service is actually met, irrespective of the yield realizations. To develop flexible scheduling policies, we can use the optimal RPS solution as a starting point and correct it with additional lot-scheduling decisions in response to yield realizations.

### 3.2. Semi-flexible scheduling policy

The basic idea of the Semi-Flexible Scheduling Policy (SFSP) is to observe, in each period and for each product, the current level of net inventory or backlog, respectively. Denote by \( v_{ktn}^* \) the optimal values of the binary variables \( b_{ktn} \) from the RPS solution. In the SFSP, additional lots for each product are produced, and the yield outcome is determined until the following two conditions are met: First, the current net inventory does not fall below the required net inventory \( rni_{kt} = \sum_{n=0}^{N} (e_{kt} - d_{kt}) \times v_{ktn}^* \) determined ex ante in the robust planning model. Second, the current backlog does not exceed the corresponding admissible backlog level \( abl_{kt} = \sum_{n=0}^{N} e_{ktn} \times v_{ktn}^* \). Algorithm 1 describes the approach formally. The major advantage of the SFSP approach is that it guarantees meeting the required service-level constraint, as the actual number of lots produced reflects the realized yield outcomes and can hence differ from the ex ante planned number of lots and the corresponding \( v_{ktn}^* \) values. Furthermore, in contrast with the SDP approach, the computational expense is limited, as the solution procedure relies on application of the RPS model, which has to be solved only once. The disadvantage of this approach is that it operates based on the predetermined fixed limits of required net inventory and admissible backlog from the RPS that are not updated to reflect the yield realizations.

### 3.3. Re-optimization-oriented scheduling policies

To overcome the disadvantage of the SFSP approach and to reduce the expected cost, it is possible to update the parameters of the RPS in each period and to solve it again with an improved information basis reflecting information regarding the yield realization of the already produced lots. These Re-optimization-Oriented Scheduling Policies (ROSPs) are hence more flexible and can be expected to result in more flexible and less costly schedules. The major drawback of this approach is the necessity of solving the model frequently, possibly whenever a lot has been produced and its yield outcome has been observed. To avoid such an extreme computational cost, we propose two policy variants ROSP-1 and ROSP-2, in which per product and period 1 or 2 re-optimizations are performed, respectively.

The underlying rationale is that when determining the number of lots to be produced in period \( t \), the backlog-oriented \( \delta \)-service-level constraint and the already accrued backlog from periods \( t - 1 \) to \( t - 1 \) set an upper limit on the permissible backlog in period \( t \). As long as the current backlog \( bl_{kt} \) of any product \( k \) still exceeds this level for the current period \( t \), an additional lot of that product must be scheduled. For this reason, in our ROSP approaches, we first schedule and produce for each product the required number of lots to ensure meeting the respective \( \delta \)-service-level constraints in the following periods \( t + 1 \) to \( T \). When this situation has been reached in period \( t \), in ROSP-1 a single re-optimization is performed by solving the Linear Programming (LP) relaxation of the respective RPS model. Given the current inventory level and the information regarding periods \( t + 1 \) to \( T \), the solution of the re-optimization attempt can suggest scheduling a rounded number of \( x_{kt} \) additional lots of product \( k \) in period \( t \), especially to avoid expensive overtime in periods \( t + 1 \) to \( T \). For each product \( k \), this number of additional lots \( x_{kt} \) in period \( t \) is produced, the yield outcome is determined, the backlog and net inventory levels are updated, and we finally proceed to period \( t + 1 \), in which the same approach is repeated in a rolling manner. In the ROSP-2 policy, by contrast, we first only produce \( x_{kt} - 1 \) lots from the first re-optimization for the current product and period. We then perform a second and final re-optimization based on the yield outcome of the first \( x_{kt} - 1 \) already produced lots. Note that this approach is more flexible than the ROSP-1 approach and hence promises lower costs at the expense of additional optimizations.

/* Phase I: Determine admissible backlog and required net inventory values */
Solve RPS model and determine optimal values \( v_{ktn}^* \)
Determine admissible backlog values
\[ abl_{kt} := \sum_{n=0}^{N} e_{ktn} \times v_{ktn}^* \]
Determine required net inventory values
\[ rni_{kt} := \sum_{n=0}^{N} (e_{kt} - d_{kt}) \times v_{ktn}^* \]

/* Phase II: Schedule lots based on yield realizations \( y_{kl} \) for lot \( l \) of product \( k \) */
Set initial net inventory \( n_{k1} := 0, k \in K \)
Set initial backlog \( bl_{k1} := 0, k \in K \)

for period \( t = 1 \) to \( T \) do

/* Consider previous net inventory and current demand */
\[ n_{kt} := n_{kt-1} \]
\[ n_{kt} := n_{kt} - d_{kt} \]
for product \( k = 1 \) to \( K \) do

while \( \sum_{l=1}^{t} bl_{kl} + \min(0, n_{kt}) > \sum_{l=1}^{t} abl_{kl} \) or \( n_{kt} - rni_{kt} < 0 \) do

Produce an additional lot \( l \) of product \( k \)
Determine the yield \( y_{kl} \) of that lot
Update the net inventory \( n_{kt} := n_{kt} + y_{kl} \)
end
Update the current backlog \( bl_{kt} := \max(0, -n_{kt}) \)
end

Algorithm 1: Algorithmic outline of the heuristic SFSP
To use the RPS model as part of the ROSP approaches without changing the planning horizon \( t = 1, \ldots, T \), some minor modifications are necessary for the starting period \( t \) of a re-optimization with respect to the planning periods between \( t \) and \( T \):

1. Additional production for the already treated past periods \( 1 \) to \( t - 1 \) has to be prohibited by imposing the following additional constraint:

\[
x_{kt} = 0, \quad t = 1, \ldots, t - 1.
\]

2. The yield outcome of the already produced lots in the past and current periods \( 1 \) to \( t \) must be netted against the original demand. If period \( t - 1 \) is closed with a backlog \( b_{kt-1} \geq 0 \) of product \( k \), i.e., a negative net inventory—then this backlog is carried over to the original demand \( d_{kt} \) of period \( t \) to constitute the updated demand \( \hat{d}_{kt} = \hat{d}_{kt} + b_{kt-1} \) relevant for the decisions in period \( t \). If period \( t - 1 \) should be closed with a physical inventory \( p_{ht-1} > 0 \), i.e., a positive net inventory—then this physical inventory has to be subtracted from the original demand \( \hat{d}_{kt} \) to determine the relevant demand \( d_{kt} \) for periods \( t, t + 1, \ldots \), subject to the constraint that the demand cannot be negative. Note that this is compatible with setting \( d_{k1} = d_{k2} = \cdots = d_{kt-1} = 0 \) for each product \( k \) when solving the problem for period \( t \).

3. The parameters of expected backlog and physical inventory \( \text{eb}_{kt} \) and \( \text{ep}_{kt} \) must be updated based on this remaining (updated) demand.

4. The already accrued backlog \( b_{kt} \) during the past periods \( 1 \) to \( t - 1 \) must be considered in the \( \delta \)-service-level constraint, given the updated values of \( d_{kt}, \text{eb}_{kt}, \) and \( \text{ep}_{kt} \):

\[
\begin{align*}
\sum_{t=1}^{T-1} b_{kt} + \sum_{t=1}^{T} \sum_{n=0}^{N} \text{eb}_{kt} \times u_{kt} & \leq (1 - \delta_{k}^{\text{min}}) \\
\times \sum_{t=1}^{T} (T - t + 1) \times d_{kt}, & \quad k \in K.
\end{align*}
\]

Note that there is no double-counting of the backlog in periods \( 1 \) to \( t - 1 \), because the expected backlog \( \sum_{t=1}^{T} \sum_{n=0}^{N} \text{eb}_{kt} \times u_{kt} \) on the left-hand side of Constraint (23) is based on the updated demand \( d_{k1} = d_{k2} = \cdots = d_{kt-1} = 0 \) for each product \( k \).

5. The capacity \( c_t \) has to be updated to represent only the remaining regular capacity of the currently considered period \( t \).

In our ROSP approach, we therefore minimize Function (16) subject to Constraints (17), (18), (19), (21), (22), and (23) based on the updated demand, backlog, and capacity parameters. The computational effort to solve an RPS instance is very limited, in particular, when only the LP relaxation is considered in fractions of a second or a few seconds and its results are rounded.

4. Numerical results

4.1. Objectives and outline of the numerical study

In our numerical study, we wanted to address two topics. First, we wanted to compare the different approaches for solving the lot-scheduling problem presented in this article. In particular, we wanted to analyze the benefit of being able to react to the yield outcome of the produced lots. Second, we wanted to study how the relative performance of the different approaches and the cost figures of the resulting schedules are affected by the characteristics of the underlying problem instances. For this purpose, we treated four systematically generated problem classes using synthetic data. The first problem class, studied in Section 4.2, consisted of only two products over only six periods. This problem class is so small that we were able to use the SDP approach presented in the electronic supplement to this article to determine the optimal decision rules for each state of the system and the corresponding cost values as a benchmark. In the other three problem classes, analyzed in Section 4.3, the numbers of products and periods were larger. As a result, the state space of the underlying SDP became so large that we were not able to determine the optimal decision rules for the states of the system. In this situation, we were therefore only able to compare the different heuristic approaches presented in this article.

4.2. Small test cases

In Table 3, we present both the systematically varying and common parameters of the small problems studied in this section. Each problem is characterized by a specific demand pattern for two products over six periods with identical means of four units and either a lower or higher (L/H) variability, a probability of success \( p_k \) for a single initial operation, a success probability decay factor \( \alpha \), an available period capacity of the production system \( c_t \), and, finally, the required \( \delta \)-service level. The common parameter values of all of the small problems are listed in the lower part of Table 3. Following the approach described in Section 2.2 for the example shown in Figure 4, a straightforward numerical search procedure was used to determine, e.g., the cost-minimizing lot size of \( \alpha_{k} = 5 \), \( k = 1, 2 \) product units for the case of initial success probability \( p_k = 0.7 \), decay factor \( \alpha = 0.9 \), and initial period capacity \( c_{t0} = 10 \) units.

| Feature | Value or parameter setting |
|---------|---------------------------|
| Demand time series | \( L = \begin{bmatrix} 4 & 5 & 3 & 4 & 3 & 5 \\ 3 & 5 & 4 & 5 & 3 & 4 \end{bmatrix} \) |
| Probability of success \( p_k \) | (0.7: 0.8) |
| Decay factor \( \alpha \) | (1.0: 0.8) |
| Capacity per period \( c_t \) | (20: 30) time units |
| Required \( \delta \)-service level | (0.90: 0.95: 0.99) |
| Setup cost \( sc_k \) | 100 monetary units |
| Unit production cost \( pc_k \) | 10 monetary units |
| Holding cost \( hc_k \) | One monetary unit per period and product unit |
| Overtime cost \( oc \) | Two monetary units per time unit |
| Setup time \( ts_k \) | Five time units |
| Unit processing time \( tp_k \) | One time unit |
the four varying problem aspects specified in the upper part of Table 3 resulted in a total of 48 different problems. Note that a particular instance of a problem is characterized by a specific set of trajectories (or sequences) of yield realizations for the different products. It is conceivable that a particular scheduling heuristic performs well on one such set of trajectories but poorly on a different set. To eliminate these random effects, we created 500 different instances characterized by specific sets of yield trajectories for each of these 48 problems. To each of these 24 000 different problem instances, we applied the decision rules obtained using the SDP approach, the flexible SFSP, ROSP-1, and ROSP-2 scheduling heuristics and the robust RPS approach. We used GAMS 24.5.3 and CPLEX 12.6.2 on an office desktop computer to perform the computations. The aggregated results are reported in Table 4.

The column labeled SDP reports the average objective function or cost values from the exact SDP approach presented in the electronic supplement to this article. This SDP approach defines the optimal decision rule for each possible state of the system in addition to the optimal (expected) cost value for the initial state. The next column in the table, labeled SIM, presents the averages over all of the objective function values for the 500 problem instances of each problem when we applied the decision rules from this SDP approach to the particular set of problem instances. The next column, labeled SFSP, represents the average objective function values obtained by applying the SFSP heuristic presented in Section 3.2, and the columns labeled ROSP-1/2 represent the results obtained by applying the ROSP-1 and ROSP-2 heuristics introduced in Section 3.3. The second to last column reports the average costs of the original rigid and robust RPS schedule; see Section 3.1. Note that for this inflexible approach, the production quantities and capacity requirements are known in advance, whereas the resulting δ-service level is determined by the yield realization of the different lots and is hence a realization of a random variable. Therefore, in each particular problem instance, which is characterized by a specific set of trajectories of yield realizations for the different products, the required δ-service level may be met for some products and violated for others. Hence, the last column in Table 4 reports the averages of the violations max(0; δ(required) − δ(achieved)) of the required δ-service level over the products, problems, and instances.

The numbers presented in Table 4 reveal several interesting insights. First, it should be noted that the values in columns SDP and SIM are very similar. This suggests simultaneously that the value function in the SDP approach is computed correctly and that averaging over 500 independent replications for each problem instance leads to a reliable picture of the cost values. It should further be noted that the average cost values resulting from the re-optimization approaches (ROSP-1/2) are substantially lower than those resulting from the semi-flexible approach (SFSP), which only considers the initially determined limits on the backlog and inventory. However, the average ROSP values clearly exceed those yielded by the exact SDP approach. This result is expected, due to the heuristic nature of the ROSP-1/2 policies. It should also be noted that the ROSP-2 policy with two re-optimizations leads to substantially better results than the less flexible ROSP-1 policy.

Further examination of the results presented in Table 4 leads to the conclusion that demand variability does not substantially affect the relative performance of the different heuristics or the average cost of the resulting schedules. Apparently, the immense yield variability from the random production process dominates any effect of the demand variability. The same conclusion holds for the average violation of the required δ-service level by the rigid RPS approach. With respect to the probability of success pk, we observe that an increase in that probability can lead to a decrease in the costs of the resulting schedules, whereas the relative performance of the different scheduling approaches is still the same, an effect that we study in more detail below. If the period capacity ct increases, the average costs of the schedules decrease, which can also be explained by the decreasing need to use expensive overtime. Finally, we observe that an increase in the required δ-service level leads, as expected, to a substantial increase in the costs of the resulting schedules when the problem is solved optimally via the SDP approach. However, and interestingly, it also leads to a decrease in the average violation of the demanded δ-service level for the rigid RPS approach.

The effects shown in the numerical results are partly due to underlying properties of the optimization problem at hand and partly due to the (heuristic) solution approaches presented in this article. To study these effects in more detail, we now analyze a single-product problem based on the low-variability demand time series (4, 5, 3, 4, 3, 5) for the first product from Table 3. (Studying only a single product reduces the SDP state space along the product dimension so that we are able to increase it along other dimensions, as required for this particular analysis.) We consider the case of a success probability decay factor α = 1, a period capacity of 20 time units, and a required δ-service level of 90%. The other parameters remain as given in Table 3 with the initial success parameter p being systematically varied from 0.7 to 0.99 in steps of 0.01. The resulting expected costs of the optimal SDP approach, as well as the respective lot size, are shown in Figure 5. The graph reveals that as the initial success probability increases, the cost-minimizing lot size q from the underlying EOQ model increases as expected. As indicated in Figure 4, we expect the long-run total costs (including production and setup costs) to decrease as the initial success probability

| αk = 1 or qpk = 7 product units for the case of success probability pk = 0.8 and decay factor αk = 1. | The combination of the four varying problem aspects specified in the upper part of Table 3 resulted in a total of 48 different problems. Note that a particular instance of a problem is characterized by a specific set of trajectories (or sequences) of yield realizations for the different products. It is conceivable that a particular scheduling heuristic performs well on one such set of trajectories but poorly on a different set. To eliminate these random effects, we created 500 different instances characterized by specific sets of yield trajectories for each of these 48 problems. To each of these 24 000 different problem instances, we applied the decision rules obtained using the SDP approach, the flexible SFSP, ROSP-1, and ROSP-2 scheduling heuristics and the robust RPS approach. We used GAMS 24.5.3 and CPLEX 12.6.2 on an office desktop computer to perform the computations. The aggregated results are reported in Table 4.

The column labeled SDP reports the average objective function or cost values from the exact SDP approach presented in the electronic supplement to this article. This SDP approach defines the optimal decision rule for each possible state of the system in addition to the optimal (expected) cost value for the initial state. The next column in the table, labeled SIM, presents the averages over all of the objective function values for the 500 problem instances of each problem when we applied the decision rules from this SDP approach to the particular set of problem instances. The next column, labeled SFSP, represents the average objective function values obtained by applying the SFSP heuristic presented in Section 3.2, and the columns labeled ROSP-1/2 represent the results obtained by applying the ROSP-1 and ROSP-2 heuristics introduced in Section 3.3. The second to last column reports the average costs of the original rigid and robust RPS schedule; see Section 3.1. Note that for this inflexible approach, the production quantities and capacity requirements are known in advance, whereas the resulting δ-service level is determined by the yield realization of the different lots and is hence a realization of a random variable. Therefore, in each particular problem instance, which is characterized by a specific set of trajectories of yield realizations for the different products, the required δ-service level may be met for some products and violated for others. Hence, the last column in Table 4 reports the averages of the violations max(0; δ(required) − δ(achieved)) of the required δ-service level over the products, problems, and instances.

The numbers presented in Table 4 reveal several interesting insights. First, it should be noted that the values in columns SDP and SIM are very similar. This suggests simultaneously that the value function in the SDP approach is computed correctly and that averaging over 500 independent replications for each problem instance leads to a reliable picture of the cost values. It should further be noted that the average cost values resulting from the re-optimization approaches (ROSP-1/2) are substantially lower than those resulting from the semi-flexible approach (SFSP), which only considers the initially determined limits on the backlog and inventory. However, the average ROSP values clearly exceed those yielded by the exact SDP approach. This result is expected, due to the heuristic nature of the ROSP-1/2 policies. It should also be noted that the ROSP-2 policy with two re-optimizations leads to substantially better results than the less flexible ROSP-1 policy.

Further examination of the results presented in Table 4 leads to the conclusion that demand variability does not substantially affect the relative performance of the different heuristics or the average cost of the resulting schedules. Apparently, the immense yield variability from the random production process dominates any effect of the demand variability. The same conclusion holds for the average violation of the required δ-service level by the rigid RPS approach. With respect to the probability of success pk, we observe that an increase in that probability can lead to a decrease in the costs of the resulting schedules, whereas the relative performance of the different scheduling approaches is still the same, an effect that we study in more detail below. If the period capacity ct increases, the average costs of the schedules decrease, which can also be explained by the decreasing need to use expensive overtime. Finally, we observe that an increase in the required δ-service level leads, as expected, to a substantial increase in the costs of the resulting schedules when the problem is solved optimally via the SDP approach. However, and interestingly, it also leads to a decrease in the average violation of the demanded δ-service level for the rigid RPS approach.

The effects shown in the numerical results are partly due to underlying properties of the optimization problem at hand and partly due to the (heuristic) solution approaches presented in this article. To study these effects in more detail, we now analyze a single-product problem based on the low-variability demand time series (4, 5, 3, 4, 3, 5) for the first product from Table 3. (Studying only a single product reduces the SDP state space along the product dimension so that we are able to increase it along other dimensions, as required for this particular analysis.) We consider the case of a success probability decay factor α = 1, a period capacity of 20 time units, and a required δ-service level of 90%. The other parameters remain as given in Table 3 with the initial success parameter p being systematically varied from 0.7 to 0.99 in steps of 0.01. The resulting expected costs of the optimal SDP approach, as well as the respective lot size, are shown in Figure 5. The graph reveals that as the initial success probability increases, the cost-minimizing lot size q from the underlying EOQ model increases as expected. As indicated in Figure 4, we expect the long-run total costs (including production and setup costs) to decrease as the initial success probability

\[ \alpha_k = 1 \text{ or } q_k = 7 \text{ product units for the case of success probability } p_k = 0.8 \text{ and decay factor } \alpha_k = 1. \]
p approaches a value of one. However, in the context of the lot-scheduling problem underlying our SDP and our RPS model at the heart of this article, we only consider the remaining trade-off between using overtime and holding inventory. For this trade-off, Figure 5 shows a more complex behavior. As the initial success probability p increases from 0.7 to 0.85, the expected costs increase whenever the lot size increases. We conjecture that this is due to the larger capacity requirements for larger lots. However, if an increase of the success probability does not lead to an increase of the lot size, then the expected costs of the optimal policy from the SDP approach mostly tend to decrease. This might partly be due to the fact that the higher yield for a given lot size reduces the number of required lots and, hence, the usage of costly overtime. The highest costs occur when the success probability is close to one. In this case, a relatively large lot is produced quite infrequently, potentially using costly overtime during the respective production period and also leading to a high level of inventory over several successive periods. We therefore conclude that an increase of the initial success probability p may lead to both a decrease and an increase of the expected overtime and inventory holding costs of the optimal policy from the SDP approach through a complex interaction via the lot size from the underlying EOQ model variant.

In our next experiment, we explore the relative performance of the different heuristic approaches for different required δ-service levels. We use the same single-product data set as before, an initial success probability of p = 0.8, and now vary the required δ-service level from 0.7 to 0.99 in steps of 0.01. The results over 300 different yield scenarios for the different algorithms are presented in Figure 6.

The line “SDP-EV” shows the expected cost value of the SDP approach. It is very close to the line “SDP-300-Scn,” which represents the average cost when the optimal SDP policy is applied to the 300 different yield scenarios. The rather primitive SFSP shows a quite poor and erratic performance, with costs increasing dramatically as the required δ-service level increases. For this particular example, the performance of the ROSP-1 and ROSP-2 heuristics is very similar. Note that the costs of the ROSP-1/2 heuristics do not necessarily increase as the required δ-service level increases. Surprisingly, the performance of the two reactive heuristics seems to improve (relative to the exact SDP approach) as the required δ-service level increases. We conjecture that a very high required δ-service level essentially prohibits backlog so that finding good solutions gets relatively simple. It is also instructive to consider a single (and common) yield trajectory as shown in Figure 7 for different required δ-service levels. The cost of the optimal decision for that single scenario is substantially below the expected costs as indicated by line “SDP-1-Scn.” The reason for this behavior is that the yield realizations of the first lots of that particular scenario are larger than expected. Interestingly, for this particular scenario, the ROSP-1/2 heuristics even outperform the SDP policy, which is optimal only in expectation but not necessarily for each single scenario. Again we see cases where an increase of the required δ-service level can lead to decreasing costs of the ROSP-1/2 solution. We conjecture that this is due to the myopic approach of the ROSP-1/2 heuristics in combination with the fact that by scheduling an integer number of lots of given size it may lead to solutions in which the service-level constraint is not binding but is in fact exceeded.

4.3. Larger test cases
As the problems studied in Section 4.2 were quite small, we analyzed additional problem classes with three, six, and nine products and 5, 10, and 15 periods. The details of those problem classes are reported in Appendix A. As mentioned above, we were unable to use the exact SDP approach here due to the explosion of the state space. Table 5 presents an aggregated analysis of the results. The relative performances of the flexible scheduling approaches SFSP, ROSP-1, and ROSP-2 and the inflexible RPS planning approach are similar to those reported for the small instances in Section 4.2. Note how the costs decrease from SFSP over ROSP-1 to ROSP-2 as the schedule is updated more frequently in response to yield realizations.

Essentially the same structural results were found for the impact of the systematically changed parameters of the problems, with the single exception of the influence of the success probability p_k and success decay factor α. In Table 4, we
5. Conclusions and further research

In this article, we studied the problem of determining production schedules that meet a predetermined and backlog-oriented service-level requirement in a production environment with limited capacity and yield uncertainty that is modeled via an IG distribution. Initially, we considered the ex ante perspective such that the backlog-oriented δ-service-level requirement is met in expectation. The outcome is a robust and stable production schedule that will typically violate the service-level requirements in the ex post perspective, given a specific set of trajectories of yield realizations. However, this robust scheduling model can be used as a basis to develop flexible heuristic scheduling policies that react to the yield realizations such that the desired service level can be guaranteed. For very small problem instances, it is even possible to determine the optimal decision rules via an SDP approach. A numerical study revealed that the flexibility to react in response to the yield realizations within an SDP approach or frequent re-optimization of the scheduling model can indeed be very advantageous from a cost perspective. We also discovered complex and opposing effects of the success probability of a single operation on the capacity requirements for setup and processing operations, inventory level, and, eventually, costs. These effects might be due to the separation of lot-sizing and lot-scheduling in our approach. This aspect should be investigated in more detail in future work.

In our problem formulation, we focused on yield uncertainty due to production processes that can randomly reach an “out-of-control” state and generate IG yields. As mentioned in Section 3.1, any yield type for which the probability mass function for the yield realizations is known can be incorporated into our modeling approach. Thus, it is, in principle, possible to apply our lot-scheduling methods for stochastically proportional or binomial yields. To determine the lot sizes, the simple formulas in Equations (8) and (9) can be used in the SPYM or BIYM cases. However, it should be noted that the scheduling part of the problem is much more relevant for situations with an IG yield. From MLPO research, we know that for the IGYM case, the optimal lot sizes are always below the demand level, whereas they exceed the demand under the SPYM and BIYM conditions. Therefore, multiple production runs will be less likely in the latter yield cases. This might also result in fewer adjustments under flexible scheduling approaches. For these reasons, the separation of the lot-sizing and the lot-scheduling problems proposed in this article appears to be beneficial if, as in the IGYM, the average fraction of conforming product units decreases with the lot size. Otherwise, lot sizing and scheduling should likely be treated simultaneously, following, for the case of a δ-service level, the basic approach presented in Helber et al. (2013a) for the case of uncertain demand.

An interesting option for future research would be to develop a policy based on an approximate dynamic programming approach for our problem. Alternatively, one could use different service-level measures and compare their long-term performance in a rolling-horizon environment.

Notes on contributors

Stefan Helber holds the Chair of Production Management at the Faculty of Economics and Management, Leibniz Universität Hannover, Germany. He holds a doctoral degree from Ludwig-Maximilians-Universität München. His research interests focus on production planning and scheduling.

Karl Inderfurth is (the retired) Professor of Production Management and Logistics at the Otto-von-Guericke-Universität Magdeburg, Germany. He holds a doctoral degree from the Freie Universität Berlin. His main research interests are related to supply chain management and stochastic inventory control.

Florian Sahling holds the Chair of Production and Industrial Management at the Faculty of Economics and Business Administration, Chemnitz University of Technology, Germany. He holds a doctoral degree from Leibniz Universität Halle-Wittenberg.

Table 5. Aggregated results for the larger cases.

| SFSP | ROSP-1 | ROSP-2 | RPS | VioSl |
|------|--------|--------|-----|-------|
| 3 × 5 | 767    | 284    | 270 | 502   | 0.019 |
| 6 × 10 | 2326   | 804    | 699 | 1830  | 0.016 |
| 9 × 15 | 4486   | 1457   | 1127| 4123  | 0.013 |
| By demand variability |        |        |     |       |       |
| L     | 2594   | 874    | 723 | 2189  | 0.017 |
| H     | 2472   | 822    | 674 | 2114  | 0.015 |
| By probability p_k of operation success |        |        |     |       |       |
| 0.9800 | 2329   | 793    | 628 | 2083  | 0.016 |
| 0.9900 | 2737   | 903    | 769 | 2220  | 0.016 |
| By success decay factor α_k |        |        |     |       |       |
| 1.0000 | 3242   | 1074   | 894 | 2706  | 0.018 |
| 0.9990 | 1822   | 622    | 503 | 1598  | 0.014 |
| By period capacity utilization |        |        |     |       |       |
| 0.95  | 3059   | 1119   | 941 | 2413  | 0.013 |
| 0.90  | 577    | 457    | 1891|       |       |
| By required δ-service level |        |        |     |       |       |
| 0.90  | 880    | 636    | 568 | 823   | 0.020 |
| 0.95  | 1110   | 642    | 551 | 1203  | 0.021 |
| 0.99  | 5609   | 1267   | 977 | 4430  | 0.007 |
Universität Hannover. His research interests focus on production planning and scheduling.

Katja Schimmelpfeng holds the Chair of Procurement and Production at the Faculty of Business, Economics and Social Sciences, University of Hohenheim in Germany. Her research interests focus on production planning and scheduling for material goods and services (especially in the health care sector).

ORCID

Katja Schimmelpfeng [http://orcid.org/0000-0001-5415-8231]

References

Aloulou, M.A., Dolgui, A. and Kovalyov, M.Y. (2014) A bibliography of non-deterministic lot-sizing models. International Journal of Production Research, 52(8), 2293–2310.

Anily, S., Beja, A. and Mendel, A. (2002) Optimal lot sizes with geometric production yield and rigid demand. Operations Research, 50(3), 424–432.

Bookbinder, J.H. and Tan, J.-Y. (1988) Strategies for the probabilistic lot-sizing problem with service-level constraints. Management Science, 34(9), 1096–1108.

Brandimarte, P. (2006) Multi-item capacitated lot-sizing with demand uncertainty. International Journal of Production Research, 44(15), 2997–3022.

Grassman, S.E., Olsen, T.L. and Birge, J.R. (2008) Setting basestock levels in multi-product systems with setups and random yield. IIE Transactions, 40(12), 1158–1170.

Grosfeld-Nir, A. and Gerchak, Y. (2004) Multiple lotsizing in production to order with random yields: Review of recent advances. Annals of Operations Research, 126(1–4), 43–69.

Guu, S.M. and Zhang, A.X. (2003) The finite multiple lot sizing problem with interrupted yield and holding costs. European Journal of Operational Research, 145(3), 635–644.

Helber, S., Sahling, F. and Schimmelpfeng, K. (2013a) Dynamic capacitated lot sizing with random demand and dynamic safety stocks. OR Spectrum, 35(1), 75–105.

Helber, S., Sahling, F. and Schimmelpfeng, K. (2013b) First results on robust dynamic lot sizing with capacity constraints and random yield, in H. Tempelmeier, and H. Kuhn (eds), Proceedings of the Ninth International Conference on Stochastic Models of Manufacturing and Service Operations, pp. 55–62. Catholic University of Eichstätt-Ingolstadt, Ingolstadt.

Hsu, H.M., Su, T.S., Wu, M.C. and Huang, L.C. (2009) Multiple lot-sizing decisions with an interrupted geometric yield and variable production time. Computers and Industrial Engineering, 57(3), 699–706.

Kazaz, B. (2004) Production planning under yield and demand uncertainty with yield-dependent cost and price. Manufacturing and Service Operations Management, 6(3), 209–224.

Little, J.D.C. (1961) A proof for the queuing formula: L = λ W. Operations Research, 9(3), 383–387.

Maes, J. and van Wassenhove, L.N. (1986) A simple heuristic for the multi item single level capacitated lotsizing problem. Operations Research Letters, 4(6), 265–273.

Mazzola, J.B., McCoy, W.F. and Wagner, H.M. (1987) Algorithms and heuristics for variable-yield lot sizing, Naval Research Logistics, 34(1), 67–86.

Porteus, E.L. (1986) Optimal lot sizing, process quality improvement and setup cost reduction. Operations Research, 34(1), 137–144.

Powell, W.B. (2011) Approximate Dynamic Programming: Solving the Curses of Dimensionality, second edition, Wiley, Hoboken, NJ.

Puterman, M.L. (2005) Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley-Interscience, Hoboken, NJ.

Rajaram, K. and Karmarkar, U.S. (2002) Product cycling with uncertain yields: Analysis and application to the process industry. Operations Research, 50(4), 680–691.

Schneider, H. (1981) Effect of service-levels on order-points or order-levels in inventory models. International Journal of Production Research, 19(6), 615–631.

Shih, W. (1980) Optimal inventory policies when stockouts result from defective products. International Journal of Production Research, 18(6), 677–686.

Silver, E.A. (1976) Establishing the reorder quantity when the amount received is uncertain. INFOR, 14(1), 32–39.

Taleizadeh, A.A., Niaki, S.T. and Najafi, A.A. (2010) Multiproduct single-machine production system with stochastic scrapped production rate, partial backordering and service level constraint. Journal of Computational and Applied Mathematics, 233(8), 1834–1849.

Tempelmeier, H. (2011) A column generation heuristic for dynamic capacitated lot sizing with random demand under a fill rate constraint. Omega, 39(6), 627–633.

Tempelmeier, H. and Herpers, S. (2010) ABCP—a heuristic for dynamic capacitated lot sizing with random demand under a fill rate constraint. International Journal of Production Research, 48(17), 5181–5193.

Tempelmeier, H. and Hilger, T. (2015) Linear programming models for a stochastic dynamic capacitated lot sizing problem. Computers & Operations Research, 59, 119–125.

Wu, M.C., Huang, L.C., Hsu, H.M. and Su, T.S. (2011) Multiple lot-sizing decisions in a two-stage production with an interrupted geometric yield and non-rigid demand. Journal of the Operational Research Society, 62(6), 1075–1084.

Yano, C.A. and Lee, H.L. (1995) Lot sizing with random yields: A review. Operations Research, 43(2), 311–334.

Appendix A. Details of the larger problem instances

Table A1 presents the varying and common aspects of the larger problem instances. The demand time series were taken from Helber et al. (2013a, p. 102 and p. 103) for the cases of demand coefficients of variation of 0.2 and 0.3, respectively. For example, for the problems with P = 9 products and T = 15 periods and a coefficient of variation of the demand of 0.3, we took the first nine rows and 15 columns of the table in Helber et al. (2013a, p. 103). We determined setup costs corresponding to a given cycle length (or time between orders) of the classic EOQ model for the case of a perfect yield. Similarly, we computed a capacity ci leading to long-term capacity utilizations of 95% and 70%, respectively. For each of the 144 problems, we created 100 different instances with specific sets of yield trajectories. We used GAMS 24.4.3 and CPLEX 12.6.1 to perform the computations on an office desktop computer and terminated the computations when an optimality gap of 0.05 was reached.

| Feature | Value or parameter setting |
|---------|---------------------------|
| Problem size (P × T) | (3 × 5), (6 × 10), (9 × 15) |
| Demand coefficient of variation | (0.2, 0.3) |
| Probability of success prk | (0.98, 0.99) |
| Success decay factor uk | (1.0, 0.999) |
| Capacity utilization | (95%; 70%) |
| Required δ - service level | (0.90; 0.95; 0.99) |
| Setup cost sck | corresponding to an EOQ TBO of two periods |
| Unit production cost cck | 10 monetary units |
| Holding cost hck | One monetary unit per period and product unit |
| Overtime cost ac | Two monetary units per time unit |
| Setup time fck | Five time units |
| Unit processing time tpk | One time unit |