Abstract
3D printing services are appearing, allowing consumers or businesses to order custom parts for delivery within a quoted lead-time, regardless of design specification. The manufacturer’s problem is to ensure that customers receive their goods on time, while capacity costs (including overtime costs) are kept low. We introduce a performance measure called the service rate, analogous to the fill rate of inventory control. Using design science, we identify and evaluate order book control policies to manage the service rate – capacity cost trade-off. Our results show that production smoothing is compatible with high service rates for customized products with short delivery times.

Keywords: Service levels, Make to order (MTO), Production smoothing

Introduction
Additive manufacturing or 3D printing is predicted to revolutionize supply chains, as a single manufacturer now can make customer-specific parts for the mass market. The variety of products is so large, that a service operation is necessary, producing to specific customer orders. This means that a 3D-printing operation cannot use inventory to decouple demand variability from production – any demand variability must be absorbed by capacity, or by the delivery time. Take Shapeways as an example; with physical production times of minutes or hours, they offer a manufacturing throughput time of six days (or two days with expediting) for white plastic products with no dimension exceeding 250mm (Shapeways, 2016). This leaves some slack time for decoupling demand variability from production.

For a delivery to occur on time, the promised delivery time must be longer than the total time spent in the order book (not yet released to production) plus the physical production time (Kingsman et al., 1989), as Figure 1 illustrates. Comparing the promised delivery time of 3D printing to other service operations, we find it relatively short, leaving
only moderate slack time for buffering (Table 1).

Most of the make-to-order theory focuses on scheduling and workload control, often requiring coordination between multiple machines in job shops (Stevenson et al., 2005). Additive manufacturing differs, because a single machine can achieve complex part geometries, and because each customer may want a unique product, making the sequencing of products less of an issue than in traditional job-shop environments (Holmström et al., 2016). For additive manufacturing, production sequencing may thus be less critical than production smoothing. Nevertheless, the smoothing problem in service operations is both practically important and academically understudied.

The first reference about smoothing an order-based-operation may be Forrester (1961, p. 144), who implemented a proportional policy in a system dynamics model, where a fixed fraction of the order book was shipped in every period. The same policy was later used by Sterman (2000, pp. 723–725) in a similar context. An alternative proportional policy was suggested by Wikner et al. (2007), who also considered an agile policy for capacity adaption. In relation to this, Anderson et al. (2005) used control theory and a system dynamics model to show that bullwhip can occur in multi-stage service operations where the capacity is adapted over time. This suggests that the smoothing problem is as important in service operations as it is in inventory settings. Note that none of these studies investigated the resulting delivery performance.

Our contribution is the development of the service rate, a metric comparable to the fill rate popular in inventory theory, which measures customer satisfaction against promised delivery times. In addition, we investigate how production-smoothing policies influence service levels in the additive manufacturing context. We show that high service levels are compatible with production smoothing, and present strategies for its implementation.

While our motivation for this paper came from the 3D printing problem highlighted above, our modelling methodology and model solution could be applied to other service operations settings, see Table 1 for examples.

**Methodology**

Operations management often finds it useful to adopt a critical realist line of thinking, as it assumes an objective world to be understood or improved (Mingers, 2015). To achieve this improvement, we select appropriate models or frameworks: Here, we use the CIMO framework, explained in Denyer et al. (2008) as well as in Pawson and Tilley (1997), to guide the development of an analytical model. The main steps are illustrated in Figure 2.

First, every situation or problem that we seek to improve is embedded in a context, or a set of circumstances. Although the overall context is simply service operations, a more
precise description includes customer requirements, the cost structure of the operation, and the characteristics of the product and the process. Depending on the context, we may entertain a set of feasible interventions, which are modifications expected to improve the situation or to rectify the problem. Based on the context, there is a systems mechanism by which the intervention produces outcomes. The mechanism may loosely be regarded as the physics of the situation, while the outcomes are the results of interest. These can be divided further into expected outcomes, which is a change in the parameters, and unexpected outcomes, reflecting those outcomes that cannot be anticipated prior to implementation and testing. As this is a theoretical piece, we can only investigate expected outcomes.

**Context – Order book management in the 3D printing industry**

The introduction highlighted the context of our study. We do not repeat it here for brevity. Similarly, the intervention is production smoothing via the order book, which we have specified now, and will revisit after defining the mechanism and the expected outcomes.

**Table 1 – Examples of service operations with promised delivery times.**

<table>
<thead>
<tr>
<th>Company</th>
<th>Product or service</th>
<th>Promised delivery time</th>
</tr>
</thead>
<tbody>
<tr>
<td>British Gas</td>
<td>Heating system repair</td>
<td>Same-day</td>
</tr>
<tr>
<td>Moonpig</td>
<td>Greeting cards</td>
<td>24h</td>
</tr>
<tr>
<td>Shapeways</td>
<td>3D Printing</td>
<td>2–16d</td>
</tr>
<tr>
<td>Amazon</td>
<td>Supersaver delivery</td>
<td>5d</td>
</tr>
<tr>
<td>Dell</td>
<td>Customized computers</td>
<td>7d</td>
</tr>
<tr>
<td>Lenovo</td>
<td>Customized computers</td>
<td>14d</td>
</tr>
<tr>
<td>Nationwide</td>
<td>Loan application processing</td>
<td>14d</td>
</tr>
<tr>
<td>Anonymous</td>
<td>Industrial equipment</td>
<td>1–6w</td>
</tr>
<tr>
<td>Anonymous</td>
<td>Material handling equipment</td>
<td>Several weeks</td>
</tr>
</tbody>
</table>
Mechanism – Dynamics of the order book
Consider a production system where temporally independent periodic demand, \( d_t \), is drawn from a normal distribution, \( d_t \in N(\mu_d, \sigma_d) \). The variable \( d_t \) represents the total work content (in hours) ordered by customers in period \( t \). The demand must be released as production orders within \( Q \) periods to be delivered on time (immediate releases are necessary when \( Q = 0 \)). Let \( o_t \) denote the production orders released at time \( t \). The order book \( b_t \) contains all received customer orders that have not yet been released to production. It has the difference equation

\[
b_t = b_{t-1} + d_t - o_t. \tag{1}
\]

Since we cannot release orders that we have not yet received, the order book will never turn negative. This means that \( \sum_{t=1}^{T} o_t \leq b_0 + \sum_{t=1}^{T} d_t \) must hold. We assume that orders are released according to a First-In-First-Out (FIFO) policy implying all tardy orders must be processed before any orders that are not yet late. The schedule adherence \( a_t \) in a given period describes the difference between the cumulative actual deliveries and required deliveries according to

\[
a_t = a_{t-1} - d_{t-Q} + o_t. \tag{2}
\]

The schedule adherence behaves much like the order book, with the special property that it is positive when all deliveries are on time, and negative when there are tardy deliveries. Thus, \( a_t \) can be understood as the amount of orders released ahead of their due date and is of most interest when negative as it then quantifies tardiness.

Outcome – Service delivery performance
Let availability, \( S_t \), denote the fraction of periods in which all expiring demand has been satisfied on time. This is analogous to inventory systems where \( S_t \) measures the probability of not experiencing a stockout in an order cycle (Axssäter, 2006, p. 94). According to this definition, in a given period \( t \), \( S_t(t) = 1 \) if \( a_t \geq 0 \), otherwise \( S_t(t) = 0 \). The expectation of \( S_t(t) \) is simply,

\[
S_t = P(a_t \geq 0) = \Phi\left(\frac{\mu_a}{\sigma_a}\right), \tag{3}
\]

where \( \Phi(\cdot) \) is the cumulative density function of the standard normal distribution and \( \mu_a \) and \( \sigma_a \) are the mean and standard deviation of \( a_t \), respectively.

In inventory control, the fill rate, \( S_2 \), denotes the long-run fraction of demand that can be filled immediately from stock (Sobel, 2004). The equivalent for service operations, which we term the service rate, measures the fraction of demand satisfied on or before the promised delivery date. Therefore, we define the service rate as \( S_2 = E[DOT]/E[\text{Delivered On Time}] \), where DOT is the demand Delivered On Time in period \( t \). DOT depends on both the
expiring demand $d_{i-Q}$ and the schedule adherence $a_i$. Let us introduce the variable *modified schedule adherence* $w_i = a_i + d_{i-Q}$, which describes DOT when $0 \leq w_i \leq d_{i-Q}$.

Should $w_i < 0$, none of the demand that expires in $t$ is satisfied, and if $a_i > 0$ all demand that expires in period $t$ is satisfied. The on-time deliveries can then be expressed as,

$$
\text{DOT} = \begin{cases} 
0 & w_i < 0 \\
 w_i & 0 \leq w_i \leq d_{i-Q} \\
 w_i - a_i & 0 < a_i
\end{cases}
$$

(4)

The expectation of DOT is provided by Hedenstierna (2016, pp. 120–122) as $E[\text{DOT}] = E[(w)'] - E[(a)']$ for arbitrary nonnegative demand distributions. This can be used as an approximation when demand is normally distributed with a low probability of negative demand; then it takes the form (Disney et al., 2015):

$$
S_2 = \frac{E[\text{DOT}]}{E[d_{i-Q}]} = \frac{\sigma_w (-\mu_w/\sigma_w) - \sigma_a G(-\mu_a/\sigma_a)}{\mu_d},
$$

(5)

where $G(x) = \int_x^\infty (v-x)\phi(v)dv = \phi(x) - x[1-\Phi(x)]$ is the unit normal loss function (Axsäter, 2006). A useful property of the service rate is that when expressed as a function of $Q$ it reflects the cumulative density function of the time spent in the order book. Before identifying expressions for the variables required in (5), we shall now introduce a capacity cost model.

**Outcome – Capacity costs**

We assume that we pay for a fixed amount of regular capacity from labour, even if the actual production requirements are sometimes less than this. Overtime charges are incurred if order releases exceed the normal capacity in any single period. The capacity cost $C_i$ in a period can be expressed as $C_i = c_1 \cdot z + c_2 \cdot (a_i - z)^+$, where $c_2$ is the labour cost per hour under overtime hours, $c_1$ is the hourly labour cost during normal hours (with $c_i < c_2$), and $z$ is the nominal (guaranteed) capacity per period, also expressed in hours, $\mu_o$ is the average production rate, and $\sigma_o^2$ the corresponding variance. The cost function we use follows Hosoda and Disney (2012), and its expectation is

$$
E(C) = c_2 \sigma_o \Phi^{-1}\left(\frac{\mu_o}{c_2}\right) + c_1 \mu_o,
$$

(6)

when using the optimal capacity is $z^* = \mu_o + \sigma_o \Phi^{-1}\left(\frac{\mu_o}{\sigma_o}\right)$, where $\Phi^{-1}\left(\cdot\right)$ is the inverse of the standard normal cumulative density function (Hosoda and Disney, 2012). Note that the capacity costs are a linear function of both the mean $\mu_o$ and the standard deviation $\sigma_o$. When $\sigma_o \to 0$, $z^* \to \mu_o$, furthermore, $z^*$ is an increasing function of $\sigma_o$ when $c_2 > 2c_1$ and a decreasing function of $\sigma_o$ when $c_1 < c_2 < 2c_1$. 


**Intervention – Order book control**

As we have a certain time to fulfil demand, we will entertain a policy that pools demand in the order book, to enable a steady release rate. We choose the general linear policy of Balakrishnan et al. (2004):

\[
o_t = \sum_{n=0}^{\infty} \theta_n d_{t-n},
\]

(7)

where \( \theta_n \) is simply the covariance function between orders and demand, and \( \sum_{n=0}^{\infty} \theta_n = 1 \).

As \( d_t \) is a sequence of independent random variables, the expectation and variance can be taken directly, providing (8) and (9) in Table 2. For the schedule adherence, we may use (2) to express \( a_t \) as a weighted sum of past demands \( (\ldots, d_{t-1}, d_t) \) and then take the expectation and the variance, providing (10) and (11). Here \( H(\cdot) \) is the Heaviside step function. The modified schedule adherence \( w_t \) is directly obtainable from \( a_t \), leading to (12) and (13). These results are general, but not concrete. We shall now consider two pragmatic policies that are special cases of (7): The moving-average policy and the proportional policy.

**Table 2 – Properties of the general linear policy**

<table>
<thead>
<tr>
<th></th>
<th>General linear policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_o )</td>
<td>( \mu_d \sum_{n=0}^{\infty} \left( \sum_{j=0}^{n} \theta_j \right) - H(n-Q) )</td>
</tr>
<tr>
<td>( \sigma_o^2 )</td>
<td>( \sigma_d^2 \sum_{n=0}^{\infty} \theta_n^2 )</td>
</tr>
<tr>
<td>( \mu_w )</td>
<td>( \mu_d \sum_{n=0}^{\infty} \left( \sum_{j=0}^{n} \theta_j \right) - H(n-Q-1) )</td>
</tr>
<tr>
<td>( \sigma_w^2 )</td>
<td>( \sigma_d^2 \sum_{n=0}^{\infty} \left[ H(n-Q-1) - \sum_{j=0}^{n} \theta_j \right]^2 )</td>
</tr>
</tbody>
</table>

Under the moving-average (MA) policy, orders are simply a moving average of the last \( \beta \in \mathbb{N}^+ \) periods demand (including the current). One benefit of this policy is that all demand will be satisfied if \( \beta \leq Q+1 \), but if \( \beta \) is greater there is a risk for tardiness. The proportional policy (ES, for exponential smoothing) releases a fixed fraction of the order book each period, i.e. \( o_t = \alpha (b_{t-1} + d_t) \), where \( \{\alpha \in \mathbb{R}, 0 < \alpha \leq 1\} \). Table 3 shows the properties of these policies, following directly from the general linear policy in Table 2.

For the moving-average policy with perfect service, \( \beta = Q+1 \), the optimal capacity level is increasing in \( Q \), and tends to \( \mu_d \) as \( Q \to \infty \). For both policies (MA and ES), assuming that only \( \alpha \) or \( \beta \) are altered, \( S_1 \) and \( S_2 \) are decreasing functions of \( \sigma_o \).
Table 3 – Properties of two popular policies from the literature

<table>
<thead>
<tr>
<th></th>
<th>Moving average (MA)</th>
<th>Proportional (ES)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta_n )</td>
<td>( 1 - H(\beta - n) )</td>
<td>( \alpha(1 - \alpha)^\gamma )</td>
</tr>
<tr>
<td>( \mu_n )</td>
<td>( \mu_d )</td>
<td>( \mu_d )</td>
</tr>
<tr>
<td>( \sigma_d^2 )</td>
<td>( \beta^{-1} )</td>
<td>( \alpha(1 - \alpha)^{-1} )</td>
</tr>
<tr>
<td>( \mu_d )</td>
<td>( \mu_d \left[ Q - (\beta - 1)/2 \right] )</td>
<td>( \mu_d \left[ Q - (1 - \alpha)/\alpha \right] )</td>
</tr>
<tr>
<td>( \sigma_d^2 )</td>
<td>( \frac{1 + 2\beta^2}{6\beta} + Q \left( \frac{1 + Q}{\beta} - 1 \right) - \frac{1}{2} )</td>
<td>( 1 + Q + \frac{3 - 2\alpha}{\alpha(\alpha - 2)} + \frac{2(1 - \alpha)^{\gamma + 1}}{\alpha} )</td>
</tr>
<tr>
<td>( \mu_d )</td>
<td>( \mu_d \left[ Q + 1 - (\beta - 1)/2 \right] )</td>
<td>( \mu_d \left[ Q + 1 - (1 - \alpha)/\alpha \right] )</td>
</tr>
<tr>
<td>( \sigma_d^2 )</td>
<td>( \frac{13 + 2\beta^2}{6\beta} + Q \left( \frac{3 + Q}{\beta} - 1 \right) - \frac{3}{2} )</td>
<td>( 2 + Q + \frac{3 - 2\alpha}{\alpha(\alpha - 2)} + \frac{2(1 - \alpha)^{2\gamma + 1}}{\alpha} )</td>
</tr>
</tbody>
</table>

**Numerical analysis**

Consider normally distributed demand with \( \mu_d = 10 \) and \( \sigma_d = 1 \). We shall study the effect of \( Q \), so the physical lead-time is of no import, as is the case with the cost factors \( c_1 \) and \( c_2 \), because the capacity cost is proportional to the standard deviation of orders.

The first trade-off to consider is that between availability and capacity costs, illustrated in Figure 3. A prominent feature is the minuscule difference in cost between some near-zero and near-perfect service configurations. Similar results hold for the service rate, as Figure 4 illustrates. Both of these figures suggest that for a fixed service level, the variable \( Q \) offers diminishing returns in terms of capacity cost reduction. We can show this analytically for the moving average policy with perfect service, as the cost reduction achieved by incrementing \( Q \) is \( \sqrt{(Q+1)/(Q+2)} \), which tends to zero as \( Q \) increases.

This is made clear in Figure 5, which illustrates the capacity cost required to maintain very high service levels for both policies. In this setting, the proportional smoothing policy offers a significant cost advantage over the moving-average policy.

**Managerial implications**

The evidence so far suggests three possible configurations for the 3D-printing operation: plan for perfect service with the moving-average policy, at a considerable cost; plan for near-perfect service with the proportional policy, at significantly lower cost than perfect service; or set a completely level schedule with \( \alpha \) close to zero, accepting poor service levels. The proportional alternative seems to be the most realistic, as it ensures high service with significant production smoothing. If, as in Shapeways case, there is an option for expediting, separate order book and control policies (with different \( \alpha \)-values) must be maintained for each delivery mode. As the same capacity can be used for different promised lead times, a capacity pooling effect can be exploited, see Hedenstierna and Disney (2012).
Conclusion
At first sight, it would appear that service operations require an agile production system. However, an order book can release work orders to the production system smoothly, allowing one to make better use of capacity in a lean production mode. We have provided expressions for the first and second order moments of a general smoothing policy for order book management. We have also provided the moments for the moving average and proportional order book management policies.
We have developed a unique metric for measuring on-time delivery called the service rate. Furthermore, we have provided a measure of capacity costs based on guaranteeing workers a nominal wage each week as well as the opportunity to gain overtime during peaks of high demand. Our analysis reveals that as $Q$ decreases, the cost to maintain a given service rate increases. However, as the target service rate decreases, the capacity costs decrease. Interestingly, the optimal capacity level can be an increasing or decreasing function of $\sigma_o$ depending on whether $c_o > 2c_i$.

While the CIMO framework guided this research, we had to revisit the intervention stage after having defined the mechanism and the output in unambiguous terms. In this way, the mechanism and the intervention gave rise to an integrated system for providing the intended outcomes. To identify unintended outcomes, we would have to test the policy on actual order book data from a service operation. This might reveal new insights, not predicted by the model, that could aid in the development of a refined intervention.

Acknowledgements
Carl Philip Hedenstierna gratefully acknowledges the support of the Norwegian Research Council and the BIA program for financial support of the RSC2020 project, which enabled this study.

References
Disney, S.M., Gaalman, G.J.C., Hedenstierna, C.P.T. and Hosoda, T. (2015), “Fill rate in a periodic review order-up-to policy under auto-correlated normally distributed,


