*Proceedings of the 1990 Winter Simulation Conference*
Osman Balci, Randall P. Sadowski, Richard E. Nance (eds.)

# SINGLE-REPLICATION SIMULATION

David Goldsman
James J. Swain

School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

David Withers

CIM Application Architecture
Industrial Sector Division (ISD)
IBM
Atlanta, Georgia 30328

## ABSTRACT

Simulation is often used to study and make predictions about dynamic stochastic processes. We consider the specialized problem of analyzing simulations in a single (short) replication. This type of analysis would facilitate efficient examination of a number of simulated process control strategies over a short horizon. For instance, a manager might want to know whether or not expediting a certain "hot" order will badly upset the schedule over the next few hours. For such problems, a detailed simulation study would not be feasible, and ideally a single simulation replication per alternative would be desired. We ultimately seek to establish a methodology for single-replication simulation using control or alteration of the underlying stochastic processes.

## 1. INTRODUCTION

In a modern computer-controlled manufacturing environment, a wealth of information about current and past behavior of the operation is available. It is often possible to use this information to make predictions about future behavior of the system under various alternative control strategies. These predictions could be made on the basis of a simulation model that uses (i) the current state and schedule of the system, (ii) parameter estimates derived from past behavior, and (iii) the structure of the system. For short-term predictions, particularly when several alternatives are to be compared, the premium is on short computing times, ideally one simulation replication for each alternative.

We investigate methodologies for designing single-replication simulation experiments which provide information about the reasonable behavior of the system. The methodologies consist of two approaches for altering the simulation model while adequately modeling the likely (or reasonable) behavior of the real system. The first approach eliminates certain sources of variation; for instance, events of low probability that have substantial effect on the system output (e.g., machine failures) are ignored. (These rare but important events will be treated as initial conditions.) The second approach simply alters other sources of variation; for example, the experimenter might restrict the variation of some probability distributions. An additional goal is to provide statistical analysis from single-replication experiments; here, several procedures are examined.

The classical *direct simulation* approach would be to perform multiple simulation replications for each alternative control strategy. Standard output analysis approaches almost always require some form of multiple replication, either directly or implicitly, as in the method of batch means. The reason one uses replications is to obtain meaningful variance estimates of the point estimator for some parameter of interest (e.g., the mean or the variance of a distribution, or even the entire distribution itself). Such an approach is asymptotically exact (as the number of replications increases) and provides arbitrarily accurate views of the statistical parameters of a process. However, the direct simulation approach is not always efficient and is therefore somewhat naive. One could also use specialized variance reduction techniques to reduce the number of required simulation replications to achieve a specified sample variance. Nevertheless, for the environment this article addresses, multiple replications may be *too costly or take too long to complete.*

In addition, the standard approaches are heavily weighted towards marginal information, that is, information about all possible cases. But from a manager's perspective, conditional information might be of much greater importance. For example, we might seek information conditioned on the absence of a machine breakdown and given a specific number of customers in the queue, or perhaps conditional on the current situation plus some specified disturbance.

Finally, the conventional simulation literature usually focuses on statistical measures such as average waiting time or mean number of customers in a queue, together with methods for forming confidence intervals about the appropriate point estimator used. The manager may well be interested in different measures, the choice of which will be particularly critical in his decisions. For instance, Wu and Wysk [1989] suggest study of primary measures of interest such as maximum completion time (make span), mean product flow time, maximum flow time, number of tardy jobs, mean tardiness, and maximum lateness of jobs. Some of these measures (e.g., the maximum lateness of jobs) have distributions which are highly skewed and variable. However, since confidence intervals are usually not desired, less sophisticated estimates of variability than those which cope with the skewness may suffice.

The organization of the remainder of the article is as follows. §2 describes two typical manufacturing systems and the output measures that are of interest in these systems. §3 examines how input distributions can be altered to reduce variance in the output, while §4 discusses methods for interval estimation when only one replication is run. Conclusions and further work are outlined in the final section.

## 2. REPRESENTATIVE SYSTEMS

Single-replication simulations differ from other simulations significantly with respect to the kinds of information that are desired from the program execution. For our purposes, a typical environment is a manufacturing production line, the user

is the production supervisor, the simulation horizon is a few hours, and the performance metrics of interest are often exception conditions from normal operation. We have chosen to classify the problem domain according to production dispatch strategy since the metrics of interest vary considerably between so-called "pull" and "push" systems. Key metrics for the two types of systems are described below. A simple example consisting of three operations (kitting, soldering, and testing), together with some rework, is used to illustrate the two production strategies. See Figure 1.

### 2.1 Pull Systems

When it is desired to reduce cycle time and work in process, an attractive control strategy is to pull work through the system as production and materials handling resources become available. Pull systems are frequently used where modern manufacturing concepts and objectives are the subjects of management focus. Small or no buffers are placed before and after each resource. The metrics of particular interest for short-horizon simulations of pull systems include:

- **Starvation.** Production managers need to know if critical (bottleneck) equipment will be idle due to lack of materials or work to process.
- **Blockage.** The possibility that output is blocked from critical production equipment is another consideration to production management. This condition is common in pull systems since output buffers are typically small or non-existent. Also, if input buffers at the next work station are small, then minor disturbances at that station in operations might cause massive disturbances in overall system performance.
- **Number of completions.** An obvious short-term metric is the number of units of work that will be completed, particularly as output from a line or sector. This metric is not unique to the pull strategy.
- **Lateness.** Similarly, this metric is not unique to the pull strategy, but is an important measure of the facility's ability to meet customer requirements and commitments. The maximum tardiness and the number of units that are late are two examples.

Figure 2 illustrates the pull version of the simple three-operation system. In this production strategy, "tokens" are used to pull units forward into a server or buffer when a unit is required by a succeeding station. For example, pulling a unit forward from the soldering operation causes a unit from the soldering buffer to be pulled forward, which pulls a unit from the kitting operation, and so forth.

### 2.2 Push Systems

Push systems use the more typical dispatch strategies for existing manufacturing systems. Work is loaded into the system at a defined rate or schedule, and pushed as far forward as possible. Traditional measures for queueing systems are of interest, but estimates of average performance are not very valuable. These statistics are usually already known from experience. The important metrics for short-horizon simulations of these systems include:

- **Maximum queue size.** Even with large buffers in the manufacturing line, there are always physical limits that cannot be exceeded. Additionally, production management may be trying to lower in-process work levels with an existing planning and dispatch system in order to simultaneously capture some of the advantages of the pull strategy.
- **Idle resources.** If a production resource is predicted to be idle due to lack of material, work, or operator, the production manager can take steps to reallocate jobs or people as necessary.
- **Number of completions.** As in pull systems.
- **Lateness.** As in pull systems.

The push version of the three-operation line is illustrated in Figure 3. In this production strategy, units proceed forward as they are completed unless the succeeding buffer is full or the succeeding server is busy. In contrast to the pull system, more than one unit can accumulate in the buffers between stations.

While the above lists of output metrics are not complete, they point out some differences and similarities in terms of output requirements for short-horizon, single-replication simulations. Traditional queueing and simulation literature does not seem to provide adequate treatment of these output statistics, particularly for transient, short-horizon models.

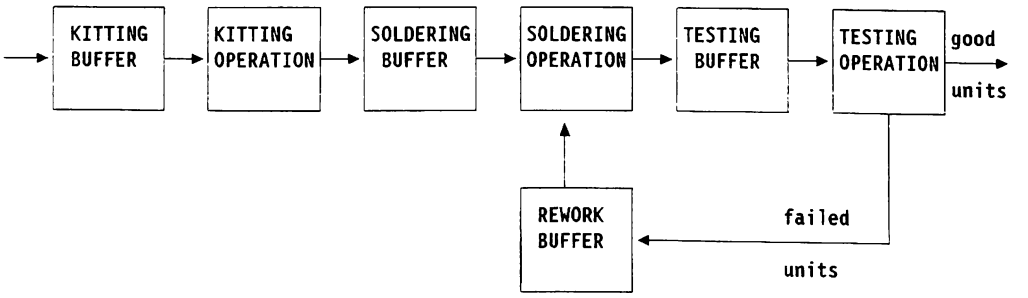


**Figure 1.** A Simple Production Process

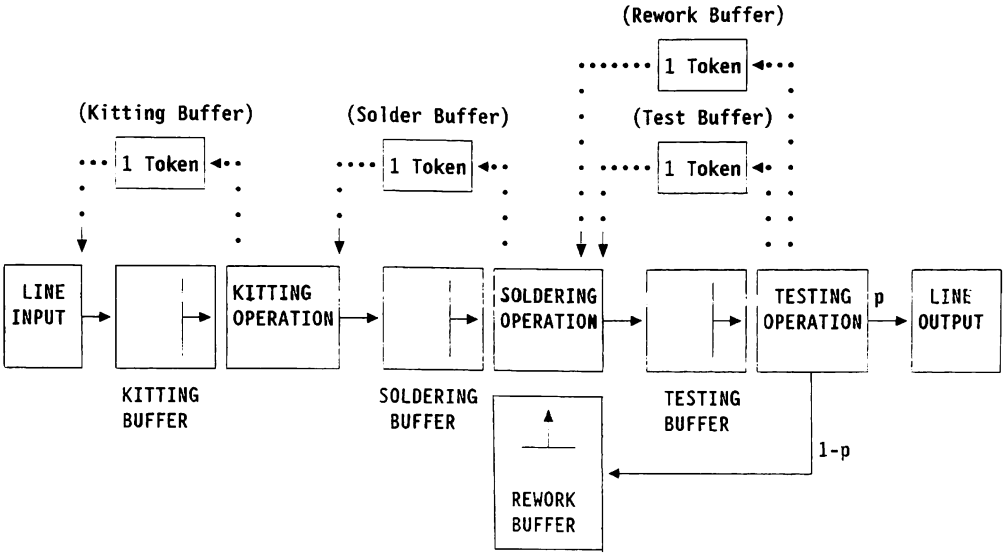


**Figure 2.** Pull Strategy for Production System. Buffers Have Single Unit Capacity (Controlled by a Passive Queue), Work Stations Have Unit Capacity, and $p$ is the Probability of a Unit Being Acceptable

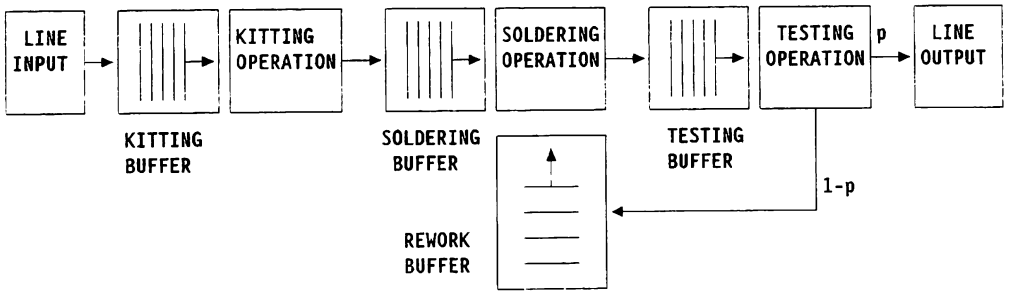


**Figure 3.** Push Strategy for Production System. Buffers Have Unlimited Capacity, Work Stations Have Unit Capacity, and $p$ is the Probability of a Unit Being Acceptable

## 3. ALTERATION OF INPUT DISTRIBUTIONS

An inherent problem of the direct simulation approach is that the statistics garnered from this approach attempt to reflect all possible system behavior; however, the decision-maker might only be interested in a "likely" subset of all possible system realizations. To concentrate on likely realizations, the analyst faces two general classes of rare behavior – those events that have relatively large influence, and those that have relatively small influence on the system output. We argue that the former be excluded from the model, since their occurrence would probably force the manager to change the set of alternatives under study. For example, when a key machine breaks down, the manager would very likely re-evaluate the situation from that point. The machine breakdown now becomes the initial condition for a subsequent simulation run. So in effect, our approach treats infrequent but influential events as initial conditions; in contrast, such events are usually sampled in direct simulation. Similarily, the interarrival times at the source can be assumed known if either the dispatch algorithm or the scheduled release times are known. In this case, a variance reduction occurs since a marginal variance in the direct approach is replaced by a conditional variance in our approach (e.g., conditional on the occurrence of the rare event). We mention that, when they are included, the effects of rare but influential events are "averaged out" over multiple replications. This averaging is ruled out when performing only a single replication of the model per alternative.

To make inferences in a single replication, random variation must be carefully controlled. We control events that are infrequent but significant in their contribution to system response variation. But even after these events are controlled, there may still be significant variation present due to the remaining "reasonable" randomness. Of course it is not desirable to eliminate all of the system's randomness, since variation is often a factor in the response of the system. For instance, the maximum tardiness is dependent upon both the mean and variance of the service time distributions. So we seek methods that

reduce the variability of the response without unduly altering the general behavior of the response.

We have considered several methods to alter the random sampling in a simulation.

- Truncate the sampling distributions, especially the upper tails for service times. A particular distribution may have to be altered in other ways in order to preserve the mean of the original distribution.
- Another way to control the random variation in a simulation response is through the use of conditional sampling. For instance, we can form a random sample conditional upon some specified mean. Such a conditional sample has less variability than an unrestricted sample. Cheng [1981] describes how to form conditional samples for a variety of popular distributions.
- Selective alteration of the distributional parameters. For example, in a queueing problem, one might increase the mean of the service-time distribution in order to compensate for decreases in the variance of the distribution.

In the remainder of this section, we present some empirical results concerning how the output of a single-server queue varies as the variance of the service distribution is reduced. During the course of our experimentation, we assumed that the service distribution is chosen so that the output mean is preserved.

For example, suppose we want to preserve the mean waiting time while reducing variation in the output. Let $s_0$ and $a_0$ denote the mean service and interarrival times, respectively, for an M/M/1 queue; let $s$ and $\sigma^2$ denote the service time mean and variance, respectively, for the M/G/1 queue. Then if the steady-state mean waiting time is to be preserved between the two systems, we must set

$$s = \frac{-s_0^2}{a_0 - s_0} + \left[\frac{s_0^2}{(a_0 - s_0)^2}(2a_0^2 + s_0^2 - 2a_0 s_0) - \sigma^2\right]^{1/2}.$$

For the cases of interest (i.e., $\sigma \leq s_0$), these solutions have the property that $s \geq s_0$, with equality in the case that $\sigma = s_0$ (which occurs for exponential services). The mean service time is maximized for the M/D/1 case, for which $\sigma = 0$. Obviously, the increase in the mean service time compensates for the decrease in the service variance.

We have conducted experiments initialized empty and idle, in which we sample a fixed number of completed customers. For instance, suppose that the interarrival rate is 1.0 per unit time, and let the mean service time be 0.8. For the M/M/1 queue, the steady-state mean waiting time can be shown to equal 3.2 units. We compared the results from the M/M/1 queue to those of various M/E$_k$/1 (Erlang-$k$) queues, taking the $k = 12$ system to be approximately normal, and the $k = \infty$ to be the M/D/1 system. Each of the Erlang systems's parameters are adjusted so that the expected waiting time matches that of the M/M/1 system (3.2). See Table 1, where we denote the waiting time of the $i$th customer as $W_i$. By direct observation, we found the transient response function for the M/M/1 system to be close to 3.2 for the 100th completed customer (over 400 replications, $E[W_{100}]$ was estimated to be 3.2, with a standard error of 0.2). Table 1 compares the match between the transient response and the variance of the output.

We see that the transient responses of the non-M/M/1 systems are slower than that of the M/M/1 system. The less variable systems take longer to reach the 90% point; that is, the point $i$ at which the system obtained $E[W_i] = .9E[W_{100}]$. On the other hand, the variance of the response is reduced. This suggests that at least in the single-server case the mean response pattern was approximated with decreased variance by the substitution of a lower-variance service distribution. It remains to be seen whether this effect can be produced in a more complicated queueing network.

**Table 1.** Response for $W_Q$-Equivalent Systems

| Case | Time to 50% $E[W_{100}]$ | Time to 90% $E[W_{100}]$ | $V[W_{20}]$ | $V[W_{80}]$ |
|---|---|---|---|---|
| M/M/1 | 12.5 | 59.5 | 5.5 | 13.7 |
| M/E$_2$/1 | 15.5 | 61 | 3.5 | 12.6 |
| M/E$_4$/1 | 16 | 76 | 3.2 | 9.3 |
| M/E$_8$/1 | 16.5 | 61 | 2.8 | 9.6 |
| M/Nor/1 | 16 | 80 | 2.9 | 8.6 |
| M/D/1 | 18 | 96 | 2.4 | 8.2 |

## 4. INTERVAL ESTIMATION

Having obtained an estimate of an output measure from a single replication of a simulation, we might desire an interval estimate bounding our uncertainty of that measure. When there is only one replication, the method of independent replications cannot be used (as illustrated below), but Bayesian methods and the method of Machol and Rosenblatt [1966] can be used for this purpose. For now, suppose that we denote the sample (replicate) means from $b$ independent replications of the same simulation by $Y_1, \ldots, Y_b$. We assume that the run lengths of the replications are long enough so that the $Y_i$'s are approximately independent and identically distributed (i.i.d.) normal random variables.

### 4.1 Method of Independent Replications

Suppose the $Y_i$'s are approximately i.i.d. normal with unknown mean $\mu$ and variance $\sigma^2$. One of the most popular ways to estimate $\mu$ is to use a confidence interval estimator of the form

$$P(\mu \in \overline{Y} \pm t_{b-1,\alpha/2}(S^2/b)^{1/2}) \doteq 1 - \alpha,$$

where $\overline{Y}$ and $S^2$ are the sample mean and variance, respectively, of the $b$ $Y_i$'s, $1 - \alpha$ is the confidence level, and $t_{b-1,\alpha/2}$ is the $1 - \alpha/2$ quantile of the $t$-distribution with $b$ degrees of freedom. Of course, this method requires $b > 1$ in order to estimate the variance of the sample mean; so we have temporarily begged the question of single-replication simulation. To overcome this problem, we can use Bayesian techniques, as described below.

### 4.2 Normal Bayesian Method

Now suppose that the $Y_i$'s are approximately i.i.d. normal$(\mu, \sigma^2)$, where we assume that $\sigma^2$ is *known* and $\mu$ is itself a random variable with a normal$(\theta_0, \tau_0^2)$ *prior* distribution and $\theta_0$ and $\tau_0^2$ are *known*. (In a real application, an experi-

enced analyst might choose "reasonable" $\theta_0$ and $\tau_0^2$ based on prior knowledge.) Then (cf. Degroot [1975, p. 269]), the *posterior* distribution of $\mu$ given that $Y_i = y_i$, $i = 1, \ldots, b$, is normal($\theta_1$,$\tau_1^2$), where

$$\theta_1 = \frac{\sigma^2\theta_0 + b\bar{y}\tau_0^2}{\sigma^2 + b\tau_0^2},$$

$$\tau_1^2 = \frac{\sigma^2\tau_0^2}{\sigma^2 + b\tau_0^2},$$

and $\bar{y} = \frac{1}{b}\sum_{i=1}^{b} y_i$. Notice that the posterior mean $\theta_1$ is simply a weighted average of the prior mean $\theta_0$ and the sample mean $\bar{y}$; further, the special case $b = 1$ is allowed. Let $\mu_1 \equiv \mu \mid Y_i = y_i$, $i = 1, \ldots, b$. We can derive the following probability statement for $\mu$. (This statement is functionally similar to a confidence interval, but philosophically quite different, since $\mu$ is a random variable.)

$$P(\mu_1 \in \theta_1 \pm z_{\alpha/2}\tau_1) \doteq 1 - \alpha,$$

where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the normal(0,1) distribution. One problem with this Bayesian formulation is that $\sigma^2$ is usually not known in practice. This problem is addressed in a more complicated Bayesian technique, which we present next.

### 4.3 Normal-Gamma Bayesian Method

We consider here the more realistic case in which neither the mean nor the variance of the replicate means are known. Suppose that the $Y_i$'s are approximately i.i.d. normal($\mu$,$1/\tau$), where $\mu$ and the *precision* $\tau$ are *unknown*. Also, suppose that the conditional distribution of $\mu \mid \tau$ is normal($\mu_0, 1/\lambda_0\tau$), where $\lambda_0 > 0$ is a constant, and the marginal distribution of $\tau$ is gamma($\alpha_0$,$\beta_0$), where $\alpha_0$ and $\beta_0$ are positive constants, i.e., the probability density function of $\tau$ is $f(t) = \beta_0(\beta_0 t)^{\alpha_0 - 1}e^{-\beta_0 t}/\Gamma(\alpha_0)$ for $t > 0$. (The experimenter must supply $\lambda_0$, $\alpha_0$, and $\beta_0$ based on prior knowledge.) Then (cf. Degroot [1975, p. 341]), the *posterior* distribution of $\mu$ given $\tau$ and $Y_i = y_i$, $i = 1, \ldots, b$, is normal($\mu_1$,$1/\lambda_1\tau$), where

$$\mu_1 = \frac{\lambda_0\mu_0 + b\bar{y}}{\lambda_0 + b}$$

and

$$\lambda_1 = \lambda_0 + b.$$

Further, the distribution of $\tau$ given $Y_i = y_i$, $i = 1, \ldots, b$, is gamma($\alpha_1$,$\beta_1$), where

$$\alpha_1 = \alpha_0 + b/2$$

and

$$\beta_1 = \beta_0 + \frac{1}{2}\sum_{i=1}^{b}(y_i - \bar{y})^2 + \frac{b\lambda_0(\bar{y} - \mu_0)^2}{2(\lambda_0 + b)}.$$

One can then show that $(\lambda_1\alpha_1/\beta_1)^{1/2}(\mu - \mu_0)$ has a $t$-distribution with $2\alpha_1$ degrees of freedom. This results in the following probability statement for $\mu_2 \equiv \mu \mid Y_i = y_i$, $i = 1, \ldots, b$.

$$P(\mu_2 \in \mu_1 \pm t_{2\alpha_1,\alpha/2}(\beta_1/\lambda_1\alpha_1)^{1/2}) \doteq 1 - \alpha.$$

### 4.4 Method of Machol and Rosenblatt

Machol and Rosenblatt [1966] give a surprisingly simple technique for estimating confidence intervals for $\mu = E[Y_1]$, where the single-replicate mean $Y_1$ is normal($\mu$,$\sigma^2$), with $\sigma^2$ unknown. The confidence intervals are of the form

$$P(\mu \in Y_1 \pm c \mid Y_1 - a \mid) \dot{\geq} 1 - \alpha,$$

where $c > 1$ solves

$$\int_{c/(c+1)}^{c/(c-1)} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \alpha < 0.5,$$

and $a$ represents the user's "best guess" for $\mu$. For instance, to obtain approximate coverage of at least 90%, we require $c \doteq 5$. Notice that $\sigma^2$ is never explicitly estimated. However, the price that must be paid is that the confidence intervals are very wide unless the user's guess $a$ happens to be close to the realization of $Y_1$.

## 5. CONCLUSIONS AND FURTHER WORK

To obtain useful outputs from manufacturing simulations quickly, single replications of a simulation providing "representative" outputs (rather than steady-state means, for instance) are proposed. Our approaches alter the model of the original system and replace it with an approximation that might be more interesting for the restricted problem at hand. A research question is to determine the extent to which such strategies are useful, in particular, the extent to which they are more informative (or less variable) than the uncontrolled, unaltered model used with the direct approach. A related research question is to determine how to alter the sampling in the best way, if sample modification turns out to be a useful strategy for analysis.

Because only a single simulation replication will be run per alternative control strategy, Bayesian methods or the method of Machol and Rosenblatt can be used to provide interval estimates of a parameter. We illustrate how this can be done when estimating the mean of a normal output process. Further research will determine how these procedures can be applied to nonnormal output metrics, such as the maximum lateness (which is likely skewed).

## ACKNOWLEDGEMENT

This work was partially supported by IBM grant #A89071-00.

## REFERENCES

Cheng, R.C.H. (1981), "The Use of Antithetic Control Variates in Computer Simulations," In *Proceedings of the 1981 Winter Simulation Conference*, T.I. Ören, C.M. Delfosse, and C.M. Shub, Eds. IEEE, Piscataway, NJ, 313–318.

DeGroot, M.H. (1975), *Probability and Statistics*, Addison-Wesley, Reading, MA.

Machol, R.E., and J. Rosenblatt (1966), "Confidence Interval Based on Single Observation," In *Proceedings of the IEEE*, IEEE, Piscataway, NJ, 1087-1088.

Wu, S.-Y.D., and R.A. Wysk (1989), "An Application of Discrete-Event Simulation to On-Line Control and Scheduling in Flexible Manufacturing," *International Journal of Production Research 27*, 1603–1623.