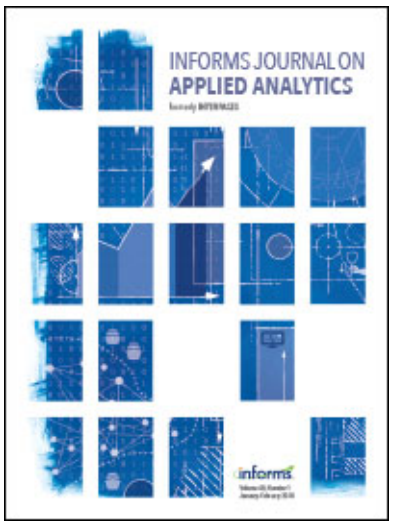

Interfaces

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Comments on "The Limits of Japanese Production Theory" by Willard I. Zangwill

To cite this article:
(1994) Comments on "The Limits of Japanese Production Theory" by Willard I. Zangwill. Interfaces 24(5):77-94.
https://doi.org/10.1287/inte.24.5.77

Comments on "The Limits of Japanese Production Theory" by Willard I. Zangwill

In 1992, we published the article "The limits on Japanese production theory" by Willard I. Zangwill. This article alone drew more response than all other articles combined during my editorship. The three comments we have chosen to publish reflect the views of all of the letters and comments that we received and were reviewed for their content. Two are technical and challenge the nature of Zangwill's results. The third is from a practitioner who asks, "Who cares?" Zangwill replies to all three, defending his point of view. Each of these authors has something useful to say, and I look forward to the debate continuing, since this is what science is about.

Frederic H. Murphy

The Limitations of Suboptimal Policies, a comment from Izak Duenyas, Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, Michigan 48109-2117.

In a recent paper, Zangwill [1992] constructed an example in which reduction of setup times led to an increase in inventory costs. In fact, in Zangwill's example, reduction of setup times causes inventory, and therefore the costs of carrying inventory, to grow without bound. Zangwill argued that these examples expose a flaw in the current Japanese production theory.

Zangwill's example involves a machine that makes n different items. The machine makes an item from kits of supply parts that arrive following a Poisson distribution. The time to make each item of a given type from a kit is an independent and identically distributed random variable. If the machine is set up to produce items of one type and management decides to switch to producing items of another type, a setup is required. The duration of this setup is an independent, identically distributed random variable.

In Zangwill's example, the machine produces items in the following manner: First, the machine is set up to make item 1, then the machine keeps making item 1 until all kits for item 1 (including those that arrived after the machine was set up) are exhausted. Then, the machine is set up to produce item 2 (regardless of whether a kit for item 2 is available at that point). Then, all kits for item 2 are exhausted, and the machine is set up for item 3 and so on. After finishing all kits for the last item, the cycle starts again. Thus, the machine uses an exhaustive polling policy to process the different items.

If the purpose of the facility is to provide quick response times to customers and to minimize inventory carrying costs for kits, the exhaustive polling policy is far from being an optimal policy. It is only a heuristic policy since there is no reason why the machine should be controlled in this particular manner. Furthermore, it is not even a reasonably good heuristic policy. To see this, assume that somehow the facility is able to reduce all setup times to zero. In this case, it has been shown that

0091-2102/94/2405/0077$01 25
This paper was refereed

INVENTORY PRODUCTION—APPLICATIONS

INTERFACES **24:** 5 September–October 1994 (pp. 77–94)

the policy that minimizes inventory costs is the weighted shortest mean processing time rule. The weighted mean processing times for each item are obtained by dividing the mean processing time for an item by the holding cost rate for that item. The optimal policy is to produce the item that has the shortest weighted mean processing time [Buyukkoc, Varaiya, and Walrand 1985]. If the inventory costs are the same for all the products, then this rule results in the production of the item with the shortest mean processing time. Hence, if the mean processing time of one product is shorter than the mean processing time for another product, the machine should not keep producing items of the second type when a kit for the first is available. However, it is clear that the exhaustive polling policy will keep on producing items of the second type, until the kits for this item are exhausted. Hence, this policy can result in large inventory costs, even if the facility is able to reduce setup times to zero.

The fact that using the exhaustive polling policy in this problem is a suboptimal heuristic policy provides an explanation for the "paradoxes" described in Zangwill [1992]. Suboptimal heuristic solutions can display unexpected behavior. This behavior does not necessarily indicate a flaw in the underlying theory. In fact, it often indicates that there is something wrong with the suboptimal heuristic policy being used. Consider a practitioner who has reduced setup times and found that inventory has grown. Should she think that there is something wrong with reducing setup times and that maybe she should be increasing setup times or should she perhaps be paying more attention to the particular heuristic production policy giving rise to this behavior?

Clearly, if the optimal production policy (that is, the policy that results in production with the minimum possible cost, given the setup times) were known, one could implement the optimal policy. Unfortunately, for the problem described above, the optimal policy is not known in the case where setup times are nonzero. However, recently researchers have partially characterized the optimal policy [Hofri and Ross 1987; Liu, Nain, and Towsley 1992; Van Oyen and Duenyas 1992]. For example, Van Oyen and Duenyas [1992] showed that when the machine is set up to produce the item with the shortest weighted mean processing time, it is indeed optimal to keep producing until all kits for that item are exhausted. However, it is clear that if the machine is set up to produce other items, it may sometimes be optimal to switch to the production of an item with a shorter weighted mean processing time. For example, in the case where the inventory carrying costs for all kits are identical, it may make sense to switch to producing the product with the shortest processing time if the setup times are low enough, and enough kits for that product have accumulated. This is due to the fact that even though a setup will be required at first, a high number of products can be produced quickly after the setup. Hence, the large decrease in the total amount of inventory that occurs after the setup is

It is not even a reasonably good heuristic policy.

completed and production begins might more than compensate for the initial time lost to the setup.

Even though the optimal policy for a problem may not have been completely characterized, it may still be possible to derive an effective heuristic solution by making use of some of the properties that the optimal solution has been shown to have. In fact, for the problem described above, Van Oyen and Duenyas [1992] recently derived an effective heuristic policy in precisely this manner by making use of known results about the properties of the optimal policy. In a comprehensive simulation study, they found that this heuristic outperformed other heuristics from the literature, including the exhaustive polling policy. I will describe this heuristic for a facility that produces two different items, and then show, using the example by Zangwill, that this heuristic results in significantly better performance than the exhaustive polling policy described above. (Van Oyen and Duenyas also derived a heuristic for systems with more than two items; however, for simplicity, I will focus on the two-item case. The heuristic Van Oyen and Duenyas describe is certainly not the only heuristic that outperforms the exhaustive policy.)

We should first question the particular suboptimal heuristic in use.

The heuristic policy Van Oyen and Duenyas [1992] derive is a simple policy that is very easy to implement. Without loss of generality, let the item that has the shortest weighted mean processing time be item 1. As I noted above, if the machine is set up for production of item 1 and kits for producing product 1 are available, it is optimal to produce another item of product 1. Therefore, suppose that the machine is set up to produce product 2, and kits for producing both product 1 and product 2 are available. In this case, the policy is to set up the machine for producing item 1 only if the number of kits for item 1 is greater than or equal to a certain threshold. The reasoning behind this heuristic is that valuable production time is lost by setting up the machine for item 1 production. Hence, the number of kits available for item 1 production should be above a certain threshold so that once the setup is completed, production of type 1 items can be continued long enough to make the switch worthwhile.

We also need to specify what the operator should do when she runs out of kits for the type of product that she is producing. If the machine is set up for item 1 production and no more kits for item 1 are available, the policy allows the operator to set the machine up to produce item 2 only if the number of kits for item 2 is above a certain threshold (which is not necessarily equal to the previous threshold described above). Otherwise, the machine is kept idle until another kit of either type 1 or type 2 arrives. Similarly, if the machine is set up for production of item 2 and there are no kits for item 2 production, the machine is set up for item 1 production only if the number of kits available for item 1 at that time is above another threshold. Finally, the policy does not allow for a setup unless at least one item has been completed

since the last setup. The appendix contains further details of this heuristic policy, and the expressions for the three thresholds that completely characterize the policy. Van Oyen and Duenyas [1992] describe a comprehensive study of the performance of this heuristic.

Having described an alternative heuristic policy to the exhaustive polling policy, we now return to the class of examples constructed by Zangwill where the exhaustive polling policy results in increased costs when setup times are decreased. We consider a machine that produces two items. We assume that the inventory holding costs for the kits are identical, so that inventory costs are proportional to the average number of kits in the system (or to the expected waiting time of a kit). Our heuristic can also be used if holding costs are different; this assumption is for expository purposes only. The setup time for item 1 is deterministic and 0.01 hours. The set-up time for item 2 is $15/u^2$ hours with probability $1 - 1/u^3$, and $15u$ hours with probability $1/u^3$. (It is easy to check that as u goes to infinity, the variance of the setup time over the mean setup time diverges to infinity, which, as Zangwill demonstrated is a sufficient condition for inventory to grow without bound when the exhaustive polling policy is used.) The Poisson arrival rate of kits for item 1 (item 2) is 0.08 (0.025) items per hour. The processing times are also deterministic: four hours for item 1 and 10 hours for item 2.

Exact results for waiting times under the exhaustive polling policy used in Zangwill [1992] were derived by Takagi [1986], and Sarkar and Zangwill [1991]. We can use those results to compute the expected waiting time for an item of type 1 or type 2. To estimate the average waiting times under our heuristic, we used simulation. For each value of u, we made 20 independent runs, each consisting of 100,000 item completions. We computed the average waiting times for items 1 and 2 separately.

Table 1 shows the results we obtained from our simulation study, as well as the results for the exhaustive polling policy. As u is increased from 2 to 7, the average setup time and the variance of the setup time for item 2 decrease. However, as Zangwill predicted, under the exhaustive polling policy, the average waiting times

			Exhaustive Polling Heuristic			New Heuristic			
u	Average Setup Time	Variance of Setup Time	Average Wait for Item 1	Average Wait for Item 2	Average Wait per Job	Average Wait for Item 1	Average Wait for Item 2	Average Wait per Job	s
2	7.031	75 4	16 26	14 67	15.88	10 58 (0.09)	21.64 (0.24)	13.21	4
3	3.272	67 0	18.88	15.30	18.02	8.38 (0 06)	20.41 (0.35)	11 24	2
4	1.860	53.7	22.53	17 41	21 31	7 50 (0 07)	19.74 (0.38)	10.41	1
7	0.611	31 8	34.56	25 40	32.38	7.12 (0.04)	17.76 (0.29)	9 65	1

Table 1: Comparison of the exhaustive polling heuristic used in Zangwill [1992] against the new heuristic.

for item 1 and item 2 increase (which implies, by Little's law, that the average inventory is also increasing.) Notice that the average waiting time per job is not the average of the average waits since kits for item 1 and item 2 arrive at different rates. The mean waiting times for item 1, item 2, and all items, under our heuristic policy (along with the standard deviation for the 20 runs in parenthesis) are also displayed in Table 1. In all cases, the average waiting time is significantly lower than that under the exhaustive polling policy. Furthermore, for higher values of u, the mean and the variance of setup times for item 2 decrease, and this results in the average waiting time for an item decreasing as well. Hence, under our heuristic, average inventory, and inventory costs decrease even though the variance over the mean is diverging.

I also considered the case where setup times for both items are zero. As previously noted, in this case, the weighted shortest mean processing time rule is optimal. Since the average waiting time when setup times are zero is a lower bound on the performance of our heuristic, I simulated this case and obtained a value of 9.11 for the average waiting time of a job. Hence, as u is increased, and thereby average setup times are decreased, the waiting times under our heuristic are getting very close to the theoretically minimum possible average waiting time that can be obtained when all setup times are zero.

The last column of Table 1 displays an interesting result. Japanese inventory theory would predict that, in this example, it is better for the machine to process item 1 when kits for item 1 are available, since item 1 can be processed much faster and wasteful inventory can be decreased more rapidly. However, when setup times for item 2 are high, it is not possible to switch over to production of item 1 very rapidly. In Table 1, s denotes the minimum number of kits of type 1 that must be available before the machine is allowed to leave unfinished work of type 2 and switch to production of item 1 under our heuristic policy. (The values for s can be easily calculated from equation (2) in the appendix.) When $u = 2$, set-up times are high, and at least four kits for item 1 are required before switching is allowed. (Setup times for item 2 affect whether we switch from producing item 2 to producing item 1, since once all kits for product 1 are exhausted, the machine has to be set up for item 2 production once again.) As u is increased, and thereby, average setup times are decreased, the number of kits of type 1 required to allow switching decrease. Finally, when $u = 4$, the setup times are low enough that the machine is allowed to switch to production of item 1 as soon as only one kit for item 1 is available.

It is interesting to note that in describing the guiding principles of Japanese inventory theory, Zangwill writes (p. 15)

> (. .) Inventory is a reflection of waste, this theory proclaims, and the more the inventory, the more the underlying waste. (. . .)
>
> (.) Reducing setups allows small production batches and low inventory, which cuts cost. Much more important perhaps, it facilitates mixed-model production. Fast setups enable a firm to make a small quantity of one model and then quickly switch over to make another model.

In our example, processing jobs of type 1 when kits for item 1 are available is pref-

erable because item 1 can be processed more rapidly and hence in a given interval, more "wasteful" inventory can be transformed into "product" than when processing item 2. It is not possible to switch to production of item 1 very quickly when setup times are high. However, when setup times are reduced, the firm gets the opportunity to switch to production of item 1 and to lower "wasteful" inventory more rapidly. As the s values indicate, our heuristic makes use of this increased capability to switch more rapidly and decreases inventory and inventory costs when setup times are reduced. However, processing jobs according to an exhaustive polling policy even when setup times are lower does not use this capability of mixed-model production to reduce wasteful inventory and results in large amounts of inventory. Hence, Japanese production theory would indeed predict that the exhaustive polling policy is not a very good heuristic policy for this problem.

Conclusions

Paradoxes and logical inconsistencies observed when using a suboptimal heuristic solution should not lead to the immediate conclusion that the underlying theory is flawed. Rather, we should first question the particular suboptimal heuristic in use, since that process may lead us to find the "optimal" solution or at least enable us to characterize some properties of the optimal solution so that we can obtain a more effective heuristic. For the particular examples of paradoxes described by Zangwill [1992], Japanese production theory indicates that the particular suboptimal heuristic policy (the exhaustive polling policy) being used in these examples is not a good policy. Another heuristic policy for the same problem outperforms the exhaustive polling policy, and it has the behavior that Japanese production theory predicts. However, this policy is still a suboptimal heuristic policy, and therefore it is entirely possible that it might display some type of behavior that is contradictory to that expected by the theory. Such behavior would only indicate the limitations of this suboptimal policy and the need for even better heuristics and more work on characterizing the optimal solution.

Appendix 1

Let λ_i denote the arrival rate of kits for product i, μ_i denote the service rate for product i, c_i denote the holding cost rate, d_i denote the (random) setup time with mean D_i, and x_i denote the number of kits of type i available for production. Assume $c_1\mu_1 \geq c_2\mu_2$, and let $\rho = \lambda_1/\mu_1 + \lambda_2/\mu_2$ denote the utilization.

The heuristic is based on the notion of "reward rates" developed in Van Oyen and Duenyas [1992]. The machine is said to be earning rewards at a rate of $c_i\mu_i$ when processing jobs of type i. The quantity $c_i\mu_i$ can be regarded as the (reward) rate at which inventory carrying costs are reduced by processing a job of type i. No rewards are earned during setups or when the machine is idle.

Since rewards are earned at the highest rate when the machine is producing product 1, if the machine is currently setup for type 1 production, and $x_1 > 0$, it is optimal to produce another item of type 1. In calculating an index for switching to type 1 production from type 2 production, I assume that when kits of type 1 are exhausted, the machine will be immediately set up for type 2 production,

even though the eventual decision may be to keep the machine idle. Now, when the machine is being set up for item 1 production, no rewards are earned for a (random) duration d_1. Then, production begins and kits for item 1 are exhausted. When the machine is set up for item 2 production, no rewards will be earned again for a (random) duration d_2. Hence, by switching to production of item 1, exhausting all kits of type 1, and switching again to item 2 production, the machine will have spent a total expected amount of time $D_1 + D_2 + \dfrac{x_1 + \lambda_1 D_1}{\mu_1 - \lambda_1}$, where the last term is the expected amount of time that the machine will be spending processing units of type 1. On average, however, rewards will be earned only for the expected duration of time equal to $\dfrac{x_1 + \lambda_1 D_1}{\mu_1 - \lambda_1}$. Therefore, we define the reward rate of switching to production of part 1, exhausting the kits for part 1 production, and setting up for production of part 2 again as

$$\varphi = c_1\mu_1 \frac{\dfrac{x_1 + \lambda_1 D_1}{\mu_1 - \lambda_1}}{D_1 + \dfrac{x_1 + \lambda_1 D_1}{\mu_1 - \lambda_1} + D_2}. \qquad (1)$$

The heuristic allows switching to production of item 1 only when the index φ is high enough that, in the time interval in which the machine is set up for item 1, kits for item 1 are exhausted, and the machine is again set up for item 2 production, the proportion of time that the machine is doing useful work (actually processing jobs and not being set up) is at least ρ. In fact, it is easy to check that requiring $\varphi > \rho c_1\mu_1 + (1 - \rho)c_2\mu_2$ ensures this. Furthermore, Van Oyen and Duenyas [1992] show that this condition ensures that the system will be stable. Hence, the heuristic allows switching to production of item 1 when $x_2 > 0$ and the machine is currently set up for item 2 production only if $x_1 > 0$ and x_1 is large enough that

$$\varphi > \rho c_1\mu_1 + (1 - \rho)c_2\mu_2. \qquad (2)$$

Furthermore, a switch is not allowed unless at least one item of type 2 has been completed since the last setup.

The derivation of the rule on whether to idle or to set up when kits are exhausted is based on a similar idea. Assume that the server is set up to process item 2 and $x_2 = 0$. In this case, we do not necessarily require that the machine perform useful work at least ρ fraction of the time when it switches, since the machine is not performing useful work by remaining idle. An immediate switch will still result in a reward rate of φ. However, consider the policy where the machine idles until the next arrival of a kit of type 1 and then is set up for item 1 production. Of course, before the next arrival of a type 1 kit, type 2 kits might arrive, and rewards may be earned by processing units of item 2. However, we assume no rewards are earned while waiting for another kit of type 1 and compute the reward rate of the inadmissable policy that idles until the next arrival of a type 1 item, then switches to production of item 1, exhausts all type 1 kits, and again switches to item 2 production. This results in the following reward rate:

$$\varphi' = c_1\mu_1 \frac{\dfrac{x_1 + 1 + \lambda_1 ED_1}{\mu_1 - \lambda_1}}{\dfrac{x_1 + 1 + \lambda_1 ED_1}{\mu_1 - \lambda_1} + \dfrac{1}{\lambda_1} + ED_1 + ED_2}. \qquad (3)$$

The condition for switching from 2 to 1 when $x_2 = 0$ is then given by

$$\varphi' < \varphi \tag{4}$$

which implies that the server earns rewards at a higher rate by switching now than by waiting for one more arrival of a kit of type 1. Simplifying (4) leads to a very simple formula for the number of kits of type 1 required so that the machine will switch to type 1 production from type 2 production without idling:

$$x_1 > \lambda_1 D_2. \tag{5}$$

Similarly, when the machine is set up for item 1 production and $x_1 = 0$, the decision is to set up if $x_2 > \lambda_2 D_1$ and to idle otherwise.

References

Buyukkoc, C.; Varaiya, P.; and Walrand, J. 1985, "The cμ-rule revisited," *Advances in Applied Probability*, Vol. 17, No. 2, pp. 237–238.

Hofri, M. and Ross, K. W. 1987, "On the optimal control of two queues with server set-up times and its analysis," *SIAM Journal of Computing*, Vol. 16, No. 2, pp. 399–420.

Liu, Z.; Nain, P.; and Towsley, D. 1992, "On optimal polling policies," *Queueing Systems*, Vol. 11, No. 1, pp. 59–83.

Sarkar, D. and Zangwill, W. I. 1991, "Variance effects on cyclic production systems," *Management Science*, Vol. 37, No. 4, pp. 444–453.

Takagi, H. 1986, *Analysis of Polling Systems*, MIT Press, Cambridge, Massachusetts.

Van Oyen, M. P. and Duenyas, I. 1992, "Control of a queueing system with sequence dependent set-ups," Technical Report 92-60, Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, Michigan.

Zangwill, W. I. 1992, "The limits of Japanese production theory," *Interfaces*, Vol. 22, No. 5, pp. 14–25.

The Cheaper/Faster-Yet-More-Expensive Phenomenon: Are Zangwill's "Paradoxes" Indeed Paradoxical? A Comment from Yigal Gerchak and Zhe Zhang, Department of Management Sciences, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.

It is well known that the performance of systems that are managed or that behave in a suboptimal manner can actually become worse when the conditions under which they operate improve. The total cost of heuristically "optimized" systems can go up when certain cost parameters are reduced or some constraints are relaxed; perhaps the best known MS/OR examples of this type are the multi-processor scheduling anomalies observed by Graham [1969, 1976]. Decentralized organizations and social systems in which each unit acts according to its own objectives, and hence not in a globally optimal manner, are also known to be susceptible to such consequences. Inasmuch as a globally optimal behavior of such systems is practically unattainable (due to the computational burden or to coordination problems), one has to accept the possibility of such cheaper/faster-yet-more-expensive phenomenon. The possibility, usually remote, of its occurrence should clearly not be used as a reason for not trying to reduce cost parameters or relax constraints in real systems, even if those systems cannot be expected to act optimally.

One of the major intellectual challenges facing production management researchers in the last decade is to precisely understand the consequences of a major part of the Japanese just-in-time philosophy—

sharp reduction in setup or changeover times and costs. While many researchers have attempted to model the benefits of such reductions so they can be traded off against the costs of achieving them, one prominent researcher, Zangwill [1987, 1992], has of late repeatedly questioned the very notion that reductions in setup time and cost are always beneficial. Whatever the other complex trade-offs associated with the JIT philosophy are and whatever its other implications, Zangwill's examples, however creative and interesting, are merely manifestations of the possibility of cheaper-yet-more-expensive consequences of suboptimality, something that can result from any attempt to enhance productivity.

In his response to Gerchak [1988], Zangwill [1988] clearly states that he believes that real systems often behave in a nonoptimal fashion and hence acknowledges that it is this suboptimality that caused the total cost to rise when setup times were cut in his [Zangwill 1987] example. Yet neither the recent *Interfaces* paper [Zangwill 1992] nor the analytical paper [Sarkar and Zangwill 1991] on which it is based explicitly tell the reader that the "paradoxical" behavior is a consequence of the underlying (exhaustive cyclical) production schedule not being optimal: "The paradoxes are clear. If setup costs or times are reduced, inventory and costs can increase. The very actions the Japanese production theory suggest should help, cutting setups, paradoxically can have the opposite impact. They can be counterproductive and hurt" [Zangwill 1992, pp. 23–24].

Sarkar and Zangwill's [1991] analysis concerns a cyclic production system in which a single machine can process two types of products. Products of type i arrive in a random fashion at node i and form a queue waiting for processing. The machine continues processing products type i until the queue for that product is empty and then switches to the other queue. There is a setup (switchover) time when the machine switches from one type of product to another. In addition, the machine keeps switching and incurring setups even when the whole system is empty. This service policy is often called impatient-and-exhaustive service. Given this particular policy, Sarkar and Zangwill construct two examples to demonstrate that cutting switchover times can increase WIP inventories and point out that the variance of switchover times plays a crucial role in these "paradoxes." The key point is that an impatient, exhaustive service policy may not always be appropriate for a manufacturing system (it is more suitable for a telecommunication system; see Takagi [1986]). A more reasonable policy for production systems is the patient-and-exhaustive service that was first studied by Eisenberg [1971]. In a system with patient and exhaustive service, the machine will remain stationary when the system becomes empty. Actually, Liu, Nain, and Towsley [1992] have recently proved that when the processing times for both queues have the same distribution (symmetric ser-

Zangwill's observations do not seem to have any normative implications.

vice times), patient and exhaustive service is the optimal policy to minimize WIP inventories or mean waiting times. Hence, the service policy considered in Sarkar and Zangwill is not optimal and might not be appropriate for production systems. Unfortunately, there is no analytical approach to find the optimal policy in this type of cyclic production system. Zhang [1992] used a computational approach to investigate the optimal policies in general cyclic production systems with two product types. The optimal policies for a system with asymmetric service times are characterized as a delayed strict priority service, as follows. When the server attends the queue with higher μc value, where μ is the rate of service and c is the holding cost rate, he or she serves that queue exhaustively; when attending the queue with lower μc value, he or she will not switch to the other queue until either the number of customers there exceeds a certain critical number or the difference in the lengths of the queues exceeds a certain critical number.

In Sarkar and Zangwill's first example, before cutting the service time, the system has a symmetric service feature. Hence, impatient, exhaustive service is close to the optimal policy (patient, exhaustive service) because it has the exhaustive service nature. However, when the service time is cut, the system becomes one of asymmetric service, and impatient, exhaustive service is far from the optimal policy (delayed strict priority). This implies that maintaining the impatient and exhaustive service is costlier. The paradox in Sarkar and Zangwill can be explained as follows: the WIP inventories in a system with a worse policy and a short processing time are higher than in a system with a bad (not optimal) policy and a long processing time. In their second example, the traffic intensities for both queues are very low (0.1). This means that the machine can clear the queue very quickly. Hence, the probability that the system becomes empty is also high. During the period when the system is empty, the machine will keep switching until the next arrival occurs. Note that cutting switchover times may then not be beneficial since the number of useless switches during the period the system is empty can increase as the switchover times decrease. This demonstrates why an impatient policy is not optimal and why the paradox may thus happen.

This suboptimal policy, which becomes worse and worse as switchover times are cut, is also the main source of the increase in total costs, Zangwill [1992] points out.

Since virtually any local cost-cutting, relaxation, or productivity-enhancement step can turn out to be counterproductive within a system that behaves suboptimally, Zangwill's observations as far as cutting setup times or costs is concerned do not seem to have any normative implications. What they may be useful for is to point out the possible strange consequences of operating systems in a nonoptimal manner. But as that is not how the author presents them, his observations are potentially highly misleading. We hope this note clarifies the issue once and for all.

References

Eisenberg, M. 1971, "Two queues with changeover times," *Operations Research*, Vol. 19, No. 1, pp. 386–401.

Gerchak, Y. 1988, "Can a reduction in set up costs indeed increase total costs?" *Manage-*

ment Science, Vol. 34, No. 10, pp. 1271–1272.
Graham, R L. 1969, "Bounds on multiprocessing timing anomalies," *SIAM Journal of Applied Mathematics*, Vol. 17, No. 1, pp. 416–429.
Graham, R. L. 1976, "Bounds on the performance of scheduling algorithms" in *Computer and Job-Shop Scheduling Theory*, ed E. G. Coffman Jr , John Wiley and Sons, New York.
Liu, Z.; Nain, P.; and Towsley, D. 1992, "On optimal polling policies," *Queueing Systems*, Vol. 11, No. 1, pp. 59–83.
Sarkar, D. and Zangwill, W I. 1991, "Variance effects in cyclic production systems," *Management Science*, Vol 37, No. 4, pp. 444–453
Takagi, I. 1986, *Analysis of Polling Systems*, MIT Press, Cambridge, Massachusetts.
Zangwill, W. I. 1987, "From EOQ towards ZI," *Management Science*, Vol. 33, No. 10, pp. 1209–1223.
Zangwill, W I 1988, "Rejoinder to the note by Yigal Gerchak," *Management Science*, Vol. 34, No. 10, pp. 1272–1273.
Zangwill, W. I. 1992, "The limits of Japanese production theory," *Interfaces*, Vol. 22, No. 5, pp. 14–25.
Zhang, Z. 1992, "Performance evaluation and optimal control for vacation queueing models and multiclass queueing models," PhD diss., Department of Management Sciences, University of Waterloo.

A Comment on Zangwill's "The Limits of Japanese Production Theory" by Barry McIntyre, Business Development, Economic Development Edmonton, Edmonton Convention Center, 9797 Jasper Avenue, Edmonton, Alberta T5J 1N9 Canada.

I am an economic development manager for Edmonton, Alberta, a city of about 700,000. Part of my mandate is to assist local firms in efforts to become more competitive. I was hired because I had several years of management experience and had done my PhD research in the area of productivity and quality. We have many firms in Edmonton that are globally competitive, and we are always looking for new ideas and methods to improve the competitiveness of firms in our region. With our natural disadvantage, being in a remote area with severe winters, I was naturally attracted to a journal whose mandate reads:

> "*Interfaces* seeks to improve communication between managers and professionals in MS/OR and to inform the academic community about practice. The most appropriate papers are descriptions of the practice and implementation of MS/OR in commerce, industry, government, or education."

Upon reading many of the articles published in *Interfaces*, I scratch my head and wonder how they connect with this mandate. Usually, I shrug my shoulders and look for some research that has some hope of application or develop an application myself. When I read the article by Zangwill in the 1992 September-October issue, I began to wonder whether the journal was losing all contact with reality. The author shows a lack of understanding for designed systems and their functioning and compounds this by drawing conclusions that are not appropriate for reasons that are not valid. Your publishing an article that relies on contrived, unrealistic conditions also makes me question your commitment to your stated mandate.

Zangwill rather loosely equates a series of concepts that are designed into useful, efficient systems with "Japanese production theory." I see no definition of "Japanese production theory" anywhere in the article. He takes a successful working system, substitutes in components and processes for which the system was not designed and which probably don't even ex-

ist, and finds that the system does not work as well. So what? One would naturally expect that to occur. This only shows that systems do not function well when critical processes are operated outside specification limits. The following example uses the same logic:

Imagine, if you will, an auto mechanic assembling an engine with components designed for different brands and types of engines. After much effort, some semblance of an engine is complete. It could have a Chevrolet short block, Chrysler heads, a Ford intake manifold, Toyota fuel injection, and a Cummins diesel exhaust manifold. After much grief and frustration in attempting to make the engine work, the mechanic throws his tool in the air and screams "we need a new theory to describe the internal combustion engine."

As Don Cherry, former coach of the Boston Bruins and current hockey commentator, would say "it doesn't take a rocket surgeon to figure this one out." It is obvious to any engineer that using components not designed to work together will usually yield negative results. If the replacement components are either critical components or are far enough from those specified in the design, the system will not work. This does not, of course, mean that there is anything wrong with the system.

The production system described by Zangwill is designed to operate as an integrated system [Senge 1990] and thus requires specific components and processes, such as on-time delivery of parts. In the article, the author introduces, in critical areas, contrived processes for which the production system was not designed, finds they yield poor results, and concludes that a new theory is needed.

The following conditions and processes are particularly suspect in the analysis:

—Higher setup costs for a shift with lower production. Setup costs can't usually be higher for a process with more slack time [Goldratt and Cox 1984]. This is illogical and not realistic.

—Setup time reduction without setup cost reduction: Goldratt and Cox, in "The Goal," indicate that the only way setup costs are not reduced with setup times is if the process has slack. If there is no slack, the cost savings are expected to be high. In the context of JIT, setup times are reduced by improving processes.

"It doesn't take a rocket surgeon to figure this one out."

—Poisson arrival times of parts. Poisson arrivals are random. Here a critical JIT process is changed. A JIT system is the antithesis of random arrivals. If the mean time is used as the expected delivery time, something on the order of half the parts will arrive late. Companies drop suppliers with this type of delivery performance. In Edmonton, our experience with supplier development at the local Dow Chemical Company plant is that on-time deliveries (meaning delivered when promised) of 80 percent are the low end of acceptable, with 90 percent becoming the new target range. Early deliveries are considered on time. A JIT system requires known, consistent, high-percentage on-time deliveries of parts, not random arrivals! I suspect that

Toyota's requirements are more stringent than those here in Edmonton. In cases in which demand is not reliably known, the supplier facilitates JIT building in excess capacity to enable the firm to meet unstable demand.

—Greater inventory requirements with lower setup times: If setup times are reduced, one can still meet production requirements with less inventory. Just idle the machine or process until the time is used up. One doesn't have to artificially increase inventory and costs to suit the model proposed.

It is neither valid nor rational to marginalize a system by using a mix-and-match approach. Systems are designed to work with components and processes that meet design specifications. Some of these components are critical to the system's working. Taguchi [Taguchi, Elsayed, and Hsiang 1989], arguably a significant player in the development of the so-called theory in question, has shown that when processes deviate from design specification, results deteriorate. It is widely known that when the design tolerances are exceeded, systems, products, and processes will fail. The negative results from the contrived processes in Zangwill's article are predictable.

The use of extreme assumptions that would likely not occur in any controlled production process makes one question the relevance of the example to the stated mandate of the journal. The word *practice* is used in the mandate. A practical environment would include a production control system that would surely not permit these extreme conditions. We surely don't need a new theory to describe improbable circumstances.

The broader issue is the statement of purpose for *Interfaces*. As a manager in an agency, working for the development of improved and more competitive production systems, I fail to see how this type of research fulfils the mandate of the journal. Indeed, the paper talks of paradoxes, and the only paradox I see is how, given your mandate, you accepted this paper.

The issues and concerns I raise should prompt some soul searching and I hope some action, on the part of both the MS/OR community and your editorial staff.

In closing, I will quote Mintzberg [1975], "the manager is in a kind of loop, with increasingly heavy work pressures but no aid forthcoming from Management Science." The real challenge to the academic community is to develop new, better systems that are applicable. Failure to address these real and pressing issues threatens the discipline with irrelevance.

I failed to see how this type of research fulfills the mandate of the journal.

References

Goldratt, E. and Cox, J 1984, *The Goal*, North River Press, Croton-on-Hudson, New York.

Mintzberg, H. 1975, "The manager's job: Folklore and fact," *Harvard Business Review*, Vol. 53, No. 4 (July–August), p. 54.

Senge, Peter M. 1990, "The leader's new work: Building learning organizations," *Sloan Management Review*, Vol. 32, No. 1 (Fall), pp. 7–23.

Taguchi, Gen'ichi; Elsayed, Elsayed A.; and Hsiang, Thomas 1989, *Quality Engineering in Production Systems*, McGraw-Hill, New York

Response to Comments on Our Work by Duenyas, by Gerchak and Zhang, and by McIntyre, by W. I. Zangwill, Graduate School of Business, University of Chicago, Chicago, Illinois 60637, and D. Sarkar, AT&T Bell Laboratories, New Jersey.

Every theory, it has been said, evolves though three stages: first the theory is utterly disregarded; second it is disputed as faulty and wrong; and finally it is declared to be self-evident and obvious. It is not clear in which of the stages the Sarkar-Zangwill (S-Z) work fits, although depending upon the comment one reads (by Duenyas, by Gerchak and Zhang, or by McIntyre), it seems to be in all three stages simultaneously.

Relative to the comments themselves, first of all we want to stress that our work is technically and mathematically correct. No one has raised the slightest issue in that regard, and our analyses as published are accurate and valid.

As for the issues raised about our work, they concern modeling or applicability, murky realms in which shadow and substance are frequently confused. What we will show in this rejoinder is that the issues raised are not substance but shadows.

To begin, it might help to review some of the history of this work. In the United States in the late 1970s, it became apparent to us that certain Japanese firms had implemented some novel concepts of production. One particular concept was to chop the time and cost of setting up certain machines. This concept was striking as it was diametrically opposite to the accepted model of the time, the EOQ, which assumes the setups are fixed and immutable. Suddenly, because of the Japanese advancement, setups were no longer fixed; instead the goal was to reduce them and, in doing so, to produce significant improvements in productivity.

A few years later, one of us (Zangwill) decided to try to prove mathematically that by reducing setup costs, inventory was reduced. This result was true for elementary models and should easily hold for more general models, he felt, as after all, it was "obvious" that cutting setups cut inventory. After several attempts to find a proof failed, he began to feel silly, yet since the theorem was "obviously" true and believed by seemingly everyone, he kept struggling to obtain a proof. After a while, of course, he constructed a counterexample in which cutting setup costs increased inventory. That example, paradoxically opposite to the "obvious" truth, formed the basis of a paper [Zangwill 1987].

Although many people responded well to that paper, a few questioned it. After all, they claimed, Toyota did not cut setup costs; they cut setup times. Surely, they said, the paradoxical result does not hold when setup times are considered; surely when setup times are cut, inventory will decrease.

At this point the two of us commenced an effort to show that reducing setup times could increase inventory. That research produced a paper [Sarkar and Zangwill 1991 (S-Z)], in which we validated that cutting setup times could also increase inventory. The usual presumption that inventories must fall was thereby proven incorrect whether one cuts setup times or setup costs.

Following the Sarkar and Zangwill pa-

per, Zangwill [1991 (Z)] suggested that these results might imply a conceptual flaw in the Japanese production philosophy and that other flaws in that philosophy might also be found. Moreover, since historically the discovery of flaws in one theory often leads to developing the next theory, perhaps examining the flaws in Japanese production theory might lead to a new and deeper theory of production. In particular, he observed that Japanese production philosophy, when confronted with considerable changes or uncertainty, seemed vulnerable. It was hoped that these deficiencies could be analyzed and overcome, leading to the development of a new deeper theory of production.

Often the "anomalous" phenomena turns out to be fundamental.

Analysis of Critiques

Returning now to the critiques themselves, the one by Duenyas and the one by Gerchak and Zhang start from different directions but express the same general argument. They assert that the S-Z model is not optimal and that it is this lack of optimality that causes the "paradoxes" with the inventory, not any fundamental flaw in the production theory. They then provide a model that is "more" optimal, they assert, and that does not exhibit the inventory paradox. That is, in their model when setup costs or times are cut, inventory goes in the direction it is supposed to, namely down.

Optimality may be in the eye of the beholder, however, as performance that to one person is optimal might to another person be dismal. Specifically, the model they select has the unfortunate property that the waiting times in the different queues can be quite disparate. Even in the example Duenyas provides, the wait in one queue is over twice as long as in the other. Depending upon the parameters, that disparity can be much larger, with the wait in one line hundreds of times longer than in the other.

The point is that even if their model is "optimal" under some criterion, many people would not call their model optimal at all. To many people the S-Z model is really "optimal" since the different queue lengths are more balanced and equal.

In short, declaring the S-Z model not optimal is incorrect, as it might be more "optimal," depending upon the criterion. Therefore, the Duenyas and the Gerchak and Zhang argument that the inventory paradoxes occur because the S-Z model is less "optimal," cannot be true.

Suboptimality is Not the Issue

Moreover, even using the criterion for optimality Duenyas, Gerchak, and Zhang prefer, their arguments are erroneous. In particular (and under their optimality criterion), Gerchak and Zhang state that the paradoxes can occur when the S-Z model is far from optimal and will not occur when the S-Z model is close to optimal. To illustrate this, Gerchak and Zhang focus on symmetric service times. For that situation the exact form of the optimal policy is not known in general, but Gerchak and Zhang note that the S-Z model is close to optimal. And they suggest, the S-Z model should exhibit no paradoxes then.

Their reasoning, however, is in error.

The paradox can occur even when the S-Z model is approximately optimal, that is, under symmetric service times, as the following example demonstrates.

Example 1: Consider a 2-node cyclic production system with symmetric service features in which the setup time is cut. Specifically, assume $\lambda_1 = \lambda_2 = 20.0$, $\bar{S}_1 = \bar{S}_2 = 0.02$, Var (S_1) = Var (S_2) = 0.0, $d_1 = 2.0$, $d_2 = 3.0$, Var $(d_1) = 0.0$, Var $(d_2) = 3{,}092.0$.

Debate might lead to a new and more powerful theory.

Under these parameter values, $\rho = 0.8$ and average response time is 316.74 ($w_1 = 364.31$, $w_2 = 269.17$). If the setup time of station 1 (that is, d_1) is cut to 1.0, then the average response time increases to 392.54 ($w_1 = 452.00$, $w_2 = 333.08$). This example is also noteworthy because the setup time that is cut has no variance.

Countering Gerchak and Zhang's argument about symmetry more strongly, the paradox can occur even when the systems are totally symmetric, before and after the setup time reduction.

Example 2: Consider a 2-node cyclic production system with symmetric service, symmetric setup times, and symmetric arrivals. Specifically, assume $\lambda_1 = \lambda_2 = 24.5$, $\bar{S}_1 = \bar{S}_2 = 0.02$, Var (S_1) = Var (S_2) = 0.0, $d_1 = d_2 = 3.0$, Var (d_1) = Var (d_2) = 3,092. Under these parameter values, $\rho = 0.98$ and average response time is $w_1 = w_2 = 592.32$. If the setup times of both stations are cut to $d_1 = d_2 = 2.9$, with the variances remaining at Var (d_1) = Var (d_2) = 3,092, then the average response time = 607.54.

The above examples also reveal another assertion of Gerchak and Zhang to be in error. In particular, they suggest that a low utilization (about 0.1) is required for the paradoxes to occur, as then the S-Z model would be far from optimal (according to their criterion). In the above examples, however, the utilizations are 0.8 and 0.98 respectively. Moreover, it appears that the paradox can occur even if the utilization is arbitrarily close to 1.

The thesis of the commentators that it is suboptimality (according to their criterion) that is causing the paradoxes is thus incorrect. What seems to be causing the paradoxes are not the issues that Gerchak and Zhang or Duenyas assert, that is, not the supposed suboptimality nor the symmetry nor the utilization level. Rather the paradoxes are caused by a combination of factors with the variance of the switchover times playing an important part. Moreover, the precise interaction of those factors is known, as we have derived conditions under which the paradox can occur [Sarkar and Zangwill 1994].

McIntyre Analysis

McIntyre's discussion is of a totally different nature. During the cold winter nights in Edmonton, he suggests (tongue-in-cheek) that reading *Interfaces* helps keep him warm but that the Z article left him cold and a bit frosted. One author of this article survives the cold winters of Chicago and also curls up with *Interfaces* for warmth, but, unlike McIntyre, finds that the Z article kept him quite toasty.

We will respond to his specific points: McIntyre suggests that a system made from incompatible components is not likely

to work well. We certainly agree. But the system we consider is made not from incompatible components but from ones designed to operate together.

If the components were incompatible, the system would almost never work. But the systems we consider, founded upon Japanese production philosophy, do work and work quite well. The problem is that suddenly and quite unexpectedly they start working in reverse—when the inventory was supposed to fall, instead it rose. The issue is not about incompatibility but about a surprising and intriguing paradox, the investigation of which might yield insight into production systems.

McIntyre also suggests that if a shift has lower production, then the setup costs should not be higher. And to presume so, he claims, is "illogical and unrealistic." We beg to differ. Suppose that the labor on a particular shift is inept. Then the production is likely to be lower and also the setup costs higher. Since it is quite possible for labor to be inept, our assumption, contrary to McIntyre's assertion, is "logical and realistic."

> The issues raised are not substance but shadows.

McIntyre then states that if setup times are reduced, setup costs are likely to be reduced also. We concur. In fact, we demonstrated our result for both cases, when setup costs are reduced and when setup time is reduced.

Continuing his assertions, McIntyre raises the issue about Poisson arrival of parts from suppliers. He states that at a JIT plant, such as Dow in Edmonton or at Toyota, the parts do not arrive randomly, but on a fixed tight schedule. We agree. But to equate what occurs at Toyota or at Dow in Edmonton with what occurs at most plants is erroneous. Many small firms have very little control over when their suppliers deliver. Also many firms are quite distant from their suppliers and so must rely on the vagaries of a distribution system that might, for instance, have to face severe winters. Demand is by no means certain in such cases, and McIntyre's assertion is incorrect.

Furthermore and as stated in our paper, our model also depicts the situation in which the arrivals are not parts from suppliers but customers. Arrivals of customers can certainly be Poisson.

Lastly, McIntyre proclaims, "We surely don't need a theory to describe improbable circumstances." He is correct, if he can be sure that the circumstances are improbable or anomalous. But how does he know that the "improbable" event is something to be rejected out of hand? Many events judged improbable or anomalous have later turned out to be not unique at all. Initially, for instance, most people in the West dismissed the high quality of Japanese cars as an improbable anomaly, not even to be considered. Such myopic head-in-the-sand approaches can only cause later grief. Instead of rejecting anomalies out of hand, one should first investigate them, because too often the "anomalous" phenomena turns out to be fundamental.

Conclusion

In sum, we strongly differ with the commentators and believe not only that our original points remain valid but that the original purpose of our papers remains in-

tact. In particular, our papers sought to expose a possible "paradox" in Japanese production theory and, in doing so, to foster a debate that might lead to the development of a new and more powerful theory. From the vigorous response we have received, our papers may be serving their purpose, with the debate just heating up.

References

Sarkar, D. and Zangwill, W. I. 1991, "Variance effects on cyclic production systems," *Management Science*, Vol. 37, No. 4, pp. 444–453.

Sarkar, D and Zangwill, W. I. 1994, "Response to critique," working paper.

Zangwill, W. I. 1992, "The limits of Japanese production theory," *Interfaces*, Vol. 22, No. 5, pp. 14–25.

Zangwill, W I. 1987, "From EOQ to ZI," *Management Science*, Vol. 33, No. 10, pp. 1209–1223.