Notes on the Productivity Paradox
Or Why Work Waits

William J. Latzko

Operations people in both manufacturing and administration are concerned with the most effective use of labor resources. They often discuss productivity. What is almost always meant by the term *Productivity* is *Labor Efficiency*. The two terms are not necessarily equivalent.

Productivity is defined as Output/Input. A good definition of Output is, "... the number of usable, saleable, acceptable goods or services produced."[1] Input should include all the costs associated with defective goods or services. Principal of these costs is rework. Other costs such as warranty and liability costs must also be included.

In short, to talk about productivity without considering quality is nonsense. In fact, if one achieves quality, productivity follows.

> Frequently, scholars and practitioners alike refer to "Productivity" and "Quality" as if they were two separate performance measures. Yet a significant part of any productivity equation **is** quality. There is no economic value in increased output levels if the increase is offset by lower quality.[2]

Many consider productivity to be the magic solution to become more competitive. The truth is that working to achieve quality will make a firm competitive as well as productive.

One can emphasize quality or emphasize productivity. The emphasis on quality brings with it increased productivity and competitiveness. The emphasis on productivity guarantees no such result. Quality pays dividends.

What makes the excess emphasis on productivity so dangerous is that it seems so simple to achieve. If only the workers would do the job as they are supposed to do, all of managements problems would disappear. "Do it right the first time" sounds great until the concept is examined in light of reality. Dr. Deming long ago showed the fallacy of this thinking.[3]

What can workers control? At best they can control the local problems. There are many factors which are global in nature which the workers cannot control. For instance, the way work arrives at their station is generally beyond their control. Yet this has a great influence in whether they work efficiently or not. The whole concept of Short Interval Scheduling (SIS), for all of its other problems, is based is the recognition that the flow of work is related to the labor efficiency and that supervisors, not the workers, are the only ones that can control this factor.[4]

Unfortunately, a large number of managers take a superficial view of the subject. They are convinced that by using the time and motion study methods and various positive and/or negative incentives, they will achieve 100% labor efficiency, which they equate to productivity. They often wonder why it just seems not to work out this way.

For instance, the manager of an engineering design unit was wondering why work always fell behind. He can estimate with great accuracy how long it takes to do a job. He can estimate accurately the

number of jobs for the year. Yet when he staffs according to this data he falls behind with an ever increasing backlog of work. His design engineers never have enough time to do the job right. They are driven by schedules which are always too tight. They are driven by numbers not quality.

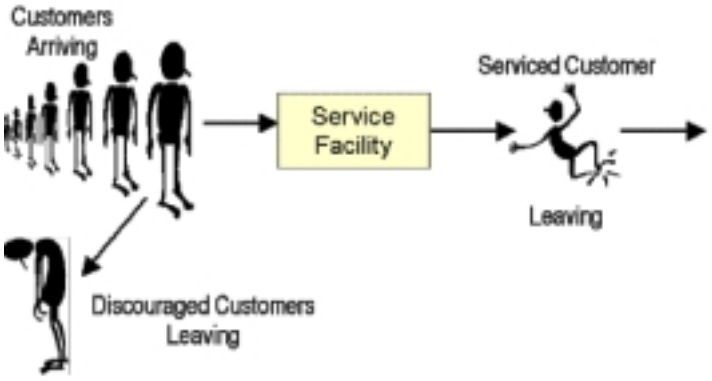
Bankers also often succumb to the industrial engineering method of achieving productivity. Indeed, there are firms that specialize in telling bankers the degree to which they are over staffed based on time study methods. Some of these firms take a percentage of the "savings" in staff that they *achieve*. These firms have long vanished from the scene when the seriousness of the problems due to the excessive cuts show up.

Managers say, "But it makes such sense. Why shouldn't we find out how long it takes to do a job and staff accordingly." The problem is that the time study methods give average data, sometimes adjusted by some factors for "fatigue" or similar things. They do not take into consideration the nature of the workflow nor the nature of the service time. What is particularly surprising is that bankers with their large number of studies of teller performance seem not to have learned from this experience.

To understand the issues we need to look at queuing systems. A queuing system is a description of how work arrives, how it is managed (or serviced), and the formation of backlogs when work cannot be serviced as soon as it appears on the scene. To begin, consider the system description below. It employs the technical term of *customer* for the arriving work.

QUEUING SYSTEM DESCRIPTION

Consider the diagram below:



The diagram shows a typical queuing system. In general, the system can be described in three stages:

1. Customers arriving for service
2. Customers waiting if service facility is busy
3. Serviced customers departing

The term customer is generic and does not necessarily apply to a human being. In this discussion, the term customer refers to work coming into the department. It can be a request for an engineering

drawing change, a requisition to purchase materials, a request to hire a new employee, baggage check at an airline, patients waiting for the availability of an operating room (and surgeon), orders in a restaurant, insurance policy claims, as well as many other types of work.

CHARACTERISTICS OF QUEUING SYSTEMS

There are five major characteristics of a queuing system. These are

1.	Arrival pattern of customers	(A)
2.	Service pattern of servers	(B)
3.	Number of service channels and service stages	(X)
4.	System capacity	(Y)
5.	Queuing discipline	(Z)

The letters in parenthesis relate to a standard notation system describing a given system using the order A/B/X/Y/Z. (This method is due mainly to D. G. Kendall)

1. Arrival patterns

The arrival of work is the input to the system. These arrivals can be either singly or in batches. The order of arrival can be of a number of distributions or be deterministic. The inter-arrival time distribution of work is mostly exponential (also known as Poisson arrivals). This will be as explained below.

The distribution has an average value. It is this average value that is often used in process planning, neglecting entirely the fact that there is a distribution around this average. Perhaps this is because most planners never learned to handle distributions.

2. Service pattern of servers

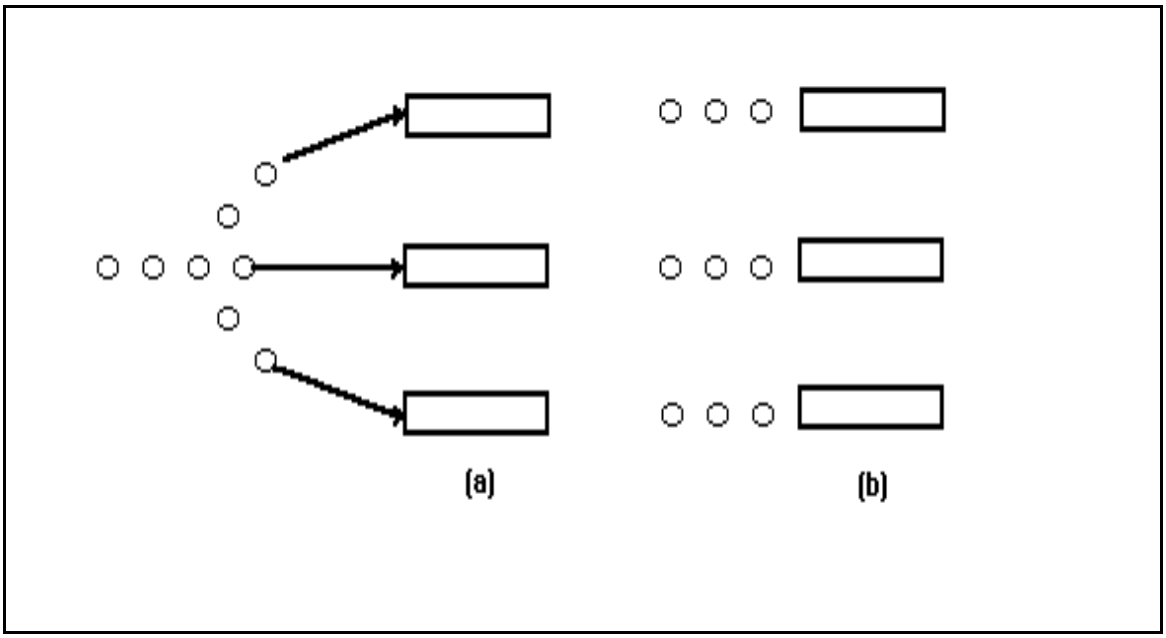
Service patterns are also described by the amount of work processed per unit of time. The condition in the case is that some work is in the system. If the system is empty, the service facility is idle. The same distributions exist for service as do for arrivals.

Again there is an average value to this distribution. Work measurement units are often the estimate of the average of the distribution. These averages are sometimes adjusted for various factors such as fatigue and the speed of the operator who was observed to get the data. These industrial engineering measures can be very misleading.

When we say that a teller services two customers per minute (30 seconds per customer) we are using an average. There are customers with some speedy transactions that take just a few seconds. There are also those customers with complex transactions that take what seems to be forever. (500 - 1000 seconds are a long time to wait behind a customer.

3. Number of service channels

The system can consist of a single server, as illustrated above or of a number of parallel servers (e. g. bank tellers.) Such multichannel servers can receive customers from (a) a single line or (b) from multiple lines.



4. System capacity

In some cases there is a physical limitation on the capacity for a queue to form. In the case of a barber shop, only so many people can be in the shop at any one time. Usually if all the chairs for waiting are filled, newly arrived customers leave again. This phenomenon is called "balking." Such a situation is rare where work is concerned. Additional work is stored and removed through the use of overtime or addition to staff.

5. Queue discipline

The most usual is first in, first out (FIFO). There are other disciplines as well. In this paper, only FIFO will be considered. There is also the opportunity to give priority to certain jobs. This is often the case with incoming work. The problem with that policy is that the lower priority work is often delayed in an excessive fashion. Unless there is a reordering of priorities on a regular basis, the use of priorities often leads to disastrous consequences and materially effects the quality of the output.

The above five characteristics of a queuing system can take the attributes shown in the following table formulated by Gross and Harris.[5]

QUEUING NOTATIONS

Characteristic	Symbol	Explanation
Inter arrival-time distribution (A)	M	Exponential
	D	Deterministic
	Ek	Erlang type k (k=1,2...)
	GI	General Independent
Service-time distribution(B)	M	Exponential
	D	Deterministic
	Ek	Erlang type k
	G	General
Number of parallel servers(X)	1,2,...infinity	
Restriction on system capacity(Y)	1,2,...infinity	
Queue discipline (Z)	FIFO	First in, first out
	LIFO	Last in, first out
	SIRO	Service in random order
	PRI	Priority
	GD	General Discipline

A model with Poisson arrivals (exponential inter arrival time), exponential service time, single server, unrestricted capacity and first in, first out discipline would be recorded as

M/M/1/infinite/FIFO.

Because of the many models where the restriction is infinity and the queue discipline is FIFO, the last two characteristics are often omitted. If omitted, assume infinity and FIFO. The above description would then read

M/M/1.

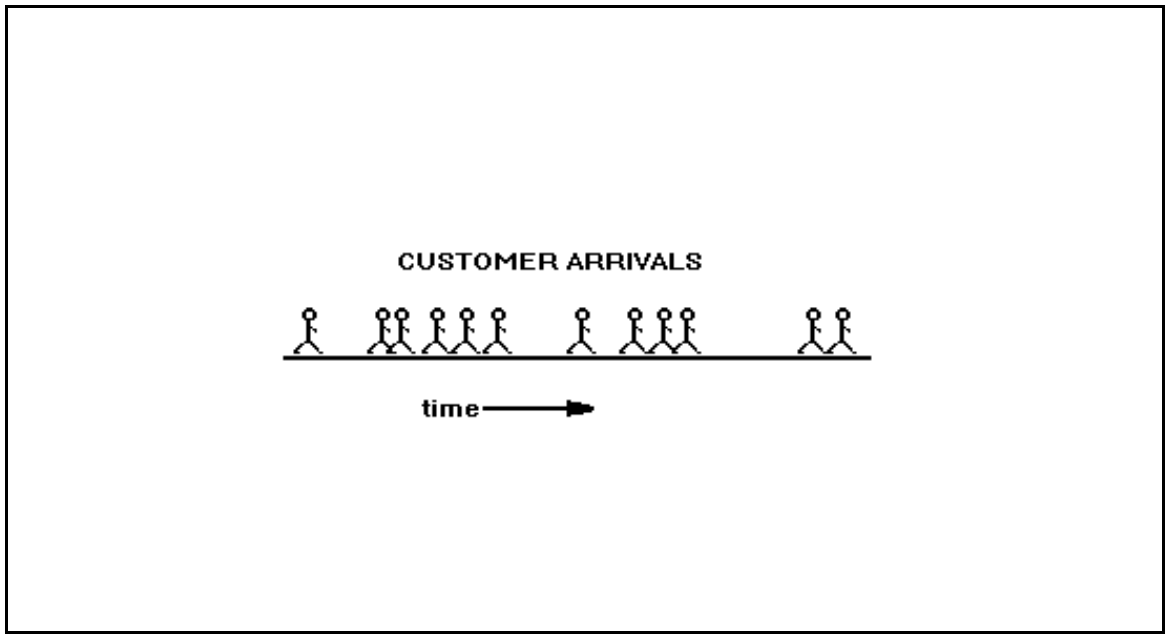
This is in effect the simplest model and one covered here.

It was stated above that the normal flow of work generally followed a pattern of a known distribution called the *Poisson Distribution.* This distribution is one commonly encountered in many applications of queuing theory. A typical example is the manner in which work arrives at teller

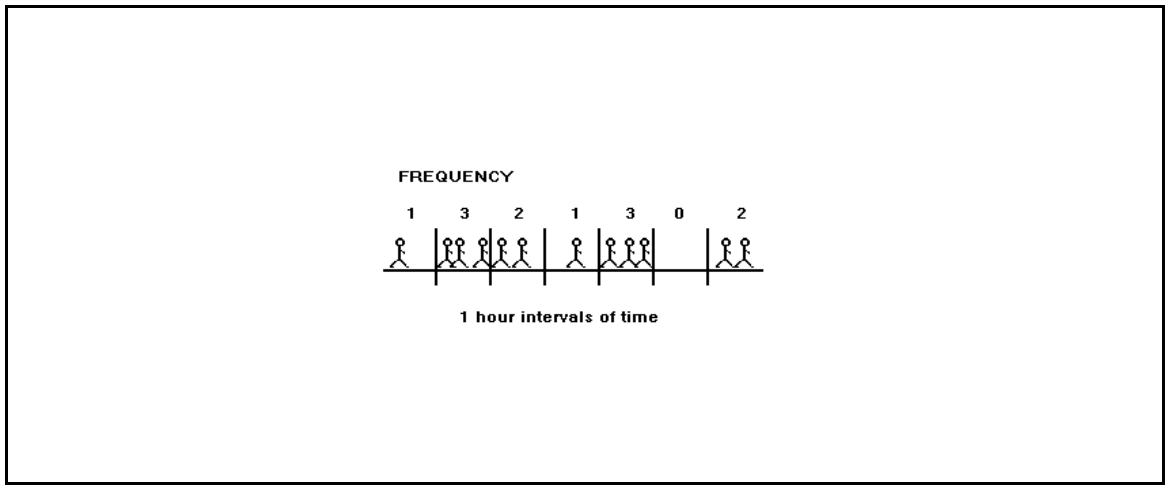
stations in a bank or supermarket checkout. The arrival of the customers (work) in these circumstances is almost always a Poisson Distribution.

POISSON ARRIVAL PATTERN

To illustrate the Poisson arrival pattern, consider the case of tellers in a bank. Customers enter the bank at random intervals as shown in the diagram below

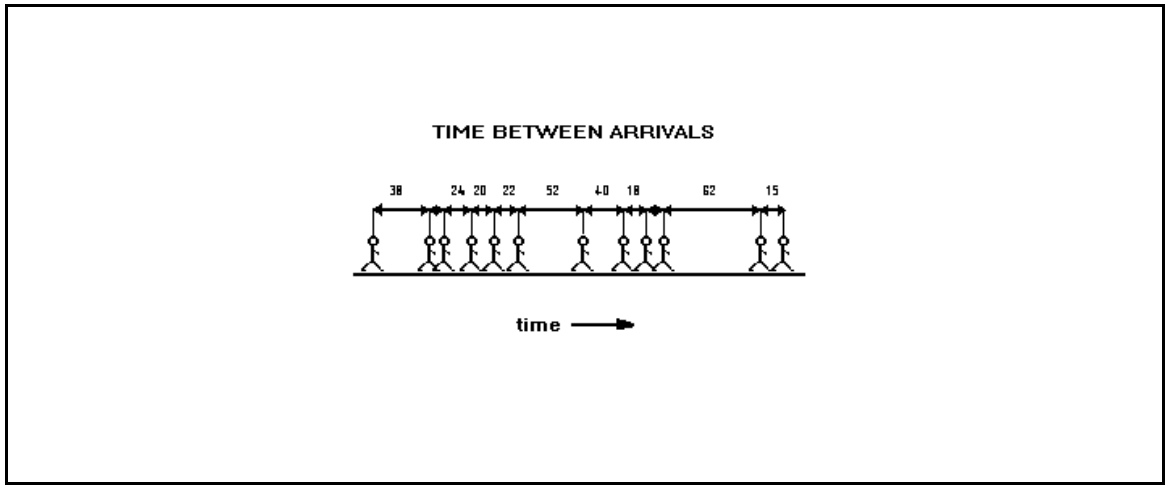


The above represents work arriving at random. Each 'O' represents an arrival. By breaking the time line up into equal intervals, a Poisson distribution is obtained. For instance, if the above time line is



broken into hours the following results:

This is a Poisson Distribution. Looking at this another way, we know from statistical theory that if the arrivals are random, then the intervals between arrivals will follow an exponential distribution.



A simple test for randomness is that

```
(Mean of the exponential) = 1 /( Mean of the Poisson)
```

If the mean of the Poisson is 3 arrivals per hour and the inter arrival time is 20 minutes (1/3 hour) the equation is satisfied.

SIMPLE QUEUES (M/M/1/infinity/FIFO)

Definitions (formulas applicable to single server only):

a =	The mean arrival rate per unit time
s =	The mean service rate per unit time
i = a/s	The intensity factor = a/s
$Lq = i^2/(1-i)$	Expected or mean length of queue
Lq(n) = 1/(1-i)	Expected number of customers in queue when queue exists
Ls = i/(1-i)	Expected number of customers in system
Wq = a/[s(s-a)]	Expected Waiting time in queue
Ws = 1/(s-a)	Expected Waiting time in system
$P(n) = P(0)(a/s)^n$	The Probability of having n customers in system

P(0) = 1 - (a/s) The Probability that the system is empty (idle)

Note that the length of queue and the waiting time are inversely proportional to the idle time.

THE CASE OF 100 % EFFICIENCY

Under industrial engineering theory, staffing is set in such a way that the average arrival of work is just covered by the average service capability. That would give 100 % efficiency of labor. For the case of a single server: a = s. Carrying this out, the intensity factor, i = 1. The theoretical idle time, P(0) = 1 - (1) = 0. No slacking off!

But what does this do to quality? The length of queue in the system when i = 1 becomes

Ls = i/(1-i) = 1/(1-1) = 1/0 = infinity.

By reducing the service capacity to remove all idle time more work enters the system than can be handled. Why? Because the work does not come in at a regular rate represented by the average. In all but the simplest cases, the amount of time to handle the work varies from item to item. Teller service is an excellent, and representative, example. The average time for transactions may be 30 seconds. Many transactions take less time, 5 seconds, 10 seconds, etc. A few transactions take the occasional 600 seconds. On average, the work takes 30 seconds, but the range is enormous.

When queues build up, the speed with which the work is performed tends to slow down. The management style of always having more work available than can be handled is a certain guarantee that less gets done and done worse at that.

In an effort to meet the work standards (a number which is at best doubtful) operators cut corners and concentrate on volume rather than how the work is performed.

References

1. Adam, Everett E., Jr.; Hershauer, James C.; and Ruch, William A., Measuring the Quality Dimension of Service Productivity, (National Science Foundation Grant No. APR 76-07140), p 2.

2. *ibid*

3. Deming, W. Edwards, Out of the Crises, (Massachusetts Institute of Technology Center for Advanced Engineering Study, Cambridge, MA 1986), See Chapters 1 and 2.

4. Smith, Martin R., Short Interval Scheduling, (New York: McGraw-Hill Book Company, 1968), p. 11.

5. Gross, Donald and Harris, Carl M., Fundamentals of Queuing Theory, (New York: McGraw-Hill Book Company, 1974), p. 9.