12-1999

The Economics of Yield-Driven Processes

Roger E. Bohn

Christian Terwiesch
University of Pennsylvania

Recommended Citation
Bohn, R. E., & Terwiesch, C. (1999). The Economics of Yield-Driven Processes. *Journal of Operations Management, 18* (1), 41-59. http://dx.doi.org/10.1016/S0272-6963(99)00014-5

The Economics of Yield-Driven Processes

Abstract
The economic performance of many modern production processes is substantially influenced by process yields. Their first effect is on product cost — in some cases, low-yields can cause costs to double or worse. Yet measuring only costs can substantially underestimate the importance of yield improvement. We show that yields are especially important in periods of constrained capacity, such as new product ramp-up. Our analysis is illustrated with numerical examples taken from hard disk drive manufacturing. A three percentage point increase in yields can be worth about 6% of gross revenue and 17% of contribution. In fact, an eight percentage point improvement in process yields can outweigh a US$20/h increase in direct labor wages. Therefore, yields, in addition to or instead of labor costs, should be a focus of attention when making decisions such as new factory siting and type of automation. The paper also provides rules for when to rework, and shows that cost minimization logic can again give wrong answers.

Keywords
Production yields, cost of quality, product cost, rework, ramp-up, location decisions, international operations

Disciplines
Business Administration, Management, and Operations | International Business | Other Economics

Title:
The Economics of Yield-Driven Processes

Author:
Roger E. Bohn; Christian Terwiesch

Publication Date:
10-02-1997

Series:
High-Technology Manufacturing

Abstract:
The economic performance of many modern production processes is substantially influenced by process yields. Their first effect is on product cost. In some cases low yields can cause costs to double or worse. Yet measuring only costs can substantially underestimate the importance of yield improvement. We show that yields are especially important in periods of constrained capacity, such as new product ramp-up. Our analysis is illustrated with numerical examples taken from hard disk-drive manufacturing. A one percentage point increase in yields can be worth about 6 percent of gross revenue and 17 percent of contribution. In fact, an eight percentage point improvement in process yields can outweigh a $20 per hour increase in direct labor wages. Therefore yields, in addition to or instead of labor costs, should be a focus of attention when making decisions such as new factory siting and type of automation. The paper also provides rules for when to rework, and shows that cost minimization logic can again give wrong answers.

The Economics of Yield-Driven Processes

Roger E. Bohn*
University of California, San Diego

Christian Terwiesch
The Wharton School

September 1, 1998

Abstract

The economic performance of many modern production processes is substantially influenced by process yields. Their first effect is on product cost – in some cases low yields can cause costs to double or worse. Yet measuring costs alone can substantially underestimate the importance of yield improvement. We show that yields are especially important in periods of constrained capacity, such as new product ramp-up. Our analysis is illustrated with numerical examples taken from hard disk-drive manufacturing. A one percentage point increase in yields can be worth about 6 percent of gross revenue and 17 percent of contribution. In fact an eight percentage point improvement in process yields can outweigh a $20 per hour increase in direct labor wages. We interpret this to mean that yields, in addition to or instead of labor costs, should be a focus of attention when making decisions such as new factory siting and type of automation. The paper also works out relative effects of first pass yield and rework yield.

PRODUCTION YIELDS, COST OF QUALITY, VALUE OF YIELDS, REWORK, LOCATION DECISIONS, INTERNATIONAL OPERATIONS

*Corresponding author: Roger Bohn, Information Storage Industry Center, IR/PS 0519, UCSD, La Jolla CA 92093-0519. Rbohn@ucsd.edu.

1. Introduction[1] Sept. 8 version

Many modern production processes and services are driven by process yields. Not every unit of material that starts into the production process makes it to the end as a salable, high quality product. Some "fall out" along the way due to problems of various kinds. Often some of the fallout can be reworked, but always a fraction of it must be scrapped. This means that materials and effort go into making something that ultimately cannot be not sold.

The effect of yield losses on the economics of the product, factory, and business can be dramatic. The comprehensive Berkeley project on semiconductors has documented many examples of integrated circuit factories with yields below 50% for years (Leachman 1996). The impact of this is, crudely, that costs per good unit are multiplied by two compared with what they would be at 100% yield. The impact on profit is much greater.

The main purpose of this paper is to analyze the economics of yield-driven production processes. Despite the widespread and important role of yields, their impact on economic performance is treated casually in management accounting systems, and has received little attention by operations management researchers. The result, we observe, is that some decisions are driven by analysis and intuition developed from inadequate models.

A secondary purpose of this paper is to compare the importance of yields with that of labor costs. Specifically, we show that under common conditions in "high-tech" industries, the impact of direct labor wage rates can be overshadowed by the effect of yields. Even eliminating direct labor entirely can have less effect on profit than modest changes in yield levels. Thus yields matter when asking questions such as "Where to site the next factory?" and "Should we automate a process?"

Our analysis is illustrated with examples from a high-tech industry, hard disk drives. Disk drive production starts with the fabrication of key components (heads, media disks, and semiconductors). All of these fabrication processes are strongly yield driven, i.e. much less than 100 percent of what goes "in" to the process comes "out" as good components. The components are then assembled in multi-step, labor- and testing- intensive processes. These assembly steps are also yield driven. The industry is sensitive to yield issues, as illustrated by the following quotation. Nonetheless it has not had good tools for quantifying their effects.

[1]Funding for this research was provided by Alfred P. Sloan Foundation through grants to the UCSD Information Storage Industry Center. We thank the following for useful comments: Neil Brumberger, Scot Burton, Scott Hampton, Michael Lapré, and two anonymous referees. Rick Dehmel provided invaluable mentoring on yields in the semiconductor industry.

> *It is how you can improve your yield that will get your productivity up. We are not in a business where you have a 99% yield. In many cases there are initial yields on high-end products that are in the 50% range. So a 5% or 10% improvement in these yields is significant.* (Richard Downing, a senior VP of manufacturing at Seagate, quoted in *Electronic Business Asia* Feb. 1997 p 35.)

Section 2 of this paper starts with some general discussion and examples of yield-driven processes. Section 3 analyzes yields in multi-stage production process. Section 4 examines the economics of rework and scrap in detail for a simple process. It concentrates on variable cost and output as the main effects of yield. In Section 5, we apply this model to hard disk drives, with numerical examples. Section 6 gives our conclusions and points out some limitations of the analysis.

2. Yields in high-tech manufacturing and other situations

Our primary domain is manufacturing in typical "high technology" industries. We define high tech as meaning that the company is on the cutting edge, with rapid product innovation. In hard disk drives, for example, product generations frequently last less than one year. Furthermore, because competitive pressure forces products to be brought to market before they or their manufacturing processes are fully understood, production techniques are at low stages of knowledge. A low stage production process is one that is not well understood and may behave unpredictably (Bohn 1994).

This situation usually has two key consequences. First, production yields are well below the ideal. Yield is the ratio between good output and gross output, so in the best case it is 100%. Because the production process is poorly understood, inevitably much of what is made does not work properly. Over time as more is learned and process problems are identified and solved, yields increase, but they never reach 100% and often never get close to it. In the Berkeley semiconductor data, for example, few CMOS factories ever reach 80% yields.

The second consequence of being on the cutting edge is that the product is in short supply. Initial production volumes are usually low, because of a variety of problems at the manufacturer or its key suppliers. If the product is successful, demand exceeds supply. Over a period of months, the manufacturing plant strives to increase output through a process known as "ramp-up," the gradual acceleration of manufacturing output from zero to full capacity. Although other forces also come into play, again the key driving force

behind ramp-up is usually learning of various kinds. Machine downtime decreases as causes are identified and fixed. Bottlenecks are detected and circumvented. More workers are trained for the labor-intensive production steps.

Notice that low yields exacerbate the problem of short supply. After all the other production problems are dealt with and units are produced, not all of them work properly. Thus one way in which output increases is by increasing yields. A key economic implication is that during ramp-up, it is generally not cost per unit from the product that most affects the company's bottom line, but its total contribution margin. Contribution, of course, is determined by production volume and selling price as well as cost. Premium pricing is the norm for new products when they are in short supply, and customers are willing to pay because of the product's special and previously unobtainable characteristics. Later, prices will fall rapidly as competitors enter and ramp up their own production. In hard disk drives, for example, selling price per megabyte has fallen at a compound rate of about 50 percent per year over the last decade. Semiconductor prices follow a similar price decline curve, albeit with bumps in response to market conditions.

The impact of low yields, therefore, is not just that they increase average unit manufacturing cost. They also decrease revenues, and therefore have a second impact on contribution and profits. In comparison, variable labor cost often has only a minor effect on contribution. Later in the life-cycle of the same product, the situation may be very different. If ramp-up was done well, yields will ultimately be high and stable, and production capacity will exceed demand for the product. At this time, cost is a key driver of profit, and labor cost often is an important expense.

2.1. Prior research on yields

The subject of process yields has received considerable attention in various disciplines, which is not surprising considering how many industries it affects (see Section 2.2). We can group this research into four streams. First, engineering reports describe yield problems in specific industrial processes and provide technical solutions. Second, operations management and operations research models support production management of yield-driven processes. Typical concerns are inventories, inspection plans, order releases, scheduling and sequencing, and other issues related to production planning. Third, there is an organizational learning literature on how to improve yields and reduce "waste". Much of it is empirical or case based. Fourth, quality management research outlines a number of principles to reduce the "cost of quality". Yield losses correspond to internal quality problems, i.e. problems caught before goods leave the factory.

There are a number of engineering articles and technical reports describing methods of dealing with yield driven production processes, especially in the semiconductor industry. For example, *IEEE Transactions on Semiconductor Manufacturing* has several articles per issue related to yields or "defects". The emphasis is on methods, concepts, and tools that will improve yields by detecting diagnosing and solving specific problems. Examples include methods of defect classification (Breaux and Kolar 1996), yield-loss modeling (Stamenkovic *et al.* 1996), in-line product inspection (Wang *et al.* 1996), statistical software to analyze process control data (Burggraaf 1996), and expert systems to provide estimates on quality of certain batches (Khera *et al.* 1994). This literature is vital to continued technological progress in these industries. As new products and processes push the state of the art, yields fall, and new cycles of yield improvement are needed.

The random nature of yield driven processes and the resulting challenges for managing production have attracted a number of operations management researchers. Most of this literature takes the production technology, and thus the yield problems, as given and provides models supporting standard production decisions such as how to manage work in process and congestion (e.g. Chen *et al.* 1988), inspection plans and quality improvement (e.g. Barad and Bennett 1996), scheduling and sequencing (e.g. Ou and Wein 1995), and other issues related to production planning (e.g. Denardo and Tang 1997).

A smaller group within the operations management literature argue that the overall yields of a production process can be improved by effective management of the process. Proposed methods for yield improvement include inspection policies for quick feed-back on the quality of the process (e.g. Tang 1991), keeping the work in progress level low (e.g. Wein 1992), and effectively combining items from different batches (e.g. Seshadri and Shanthikumar 1997). In contrast to the engineering literature, these papers focus on improving various performance measures, including yields, without really changing the underlying production technology. This makes them more general across processes, at least potentially, but limits their potency.

The third stream of yield research is at the intersection between production management and organizational research, especially organizational learning, and has contributed some in-depth empirical studies on yield improvement. Mukherjee *et al.* (1995) categorized various quality projects undertaken at a major manufacturer of wire cord depending on the type of learning approach taken in the projects. A follow-up study (Lapré *et al.* 1996) links these quality projects to waste reduction (yield improvement). Bohn (1995a,b) looks at factors which influence the speed of yield improvement in semiconductor manufacturing. Kantor and Zangwill (1991) give a theoretical model of waste reduction learning.

Like the engineering literature, the organizational literature has little to say on the economic value of yield improvement. For the most part yield improvement is implicitly treated as a way to reduce costs, without looking at other effects.

Finally, under the "cost of quality" paradigm as outlined in Juran and Gryna (1993), yield losses are viewed as part of internal failure costs and thus as one of the main drivers of the costs of quality. Juran and Gryna emphasize the need to assign economic values to these quality costs, to make them easier to understand for top management decision makers. The cost of quality approach is valuable in its recognition of hidden effects from quality problems, and its emphasis on quantifying them. For example this approach would show that when first-pass yields get high enough, in-process inspection can be eliminated, which has various desirable effects. However, one of the main benefits of yield improvement is ignored, namely the improvement in effective capacity and output.

In the quality literature, yield loss is the extreme form of a defect - the product is unsalable. Therefore, much of the quality improvement literature is applicable to yield improvement. Probably most important are the tools and concepts of statistical process control to yield monitoring and improvement. Again this is most active for semiconductors; see the survey/tutorial by Spanos (1992). Typical issues include how to detect a defective machine quickly, what inspection policies to set, and how to modify SPC tools such as control charts to cope with the huge amount of data produced by automated semiconductor manufacturing lines.

Although the literature reviewed above has substantially improved our understanding of yield issues in production processes, none of it has provided the basic economic analysis of how yields matter. We attempt to extend the literature in three directions:

- we assign concrete economic values to yield issues (Juran and Gryna 1993)
- we do not take yields as given, rather we concentrate on the value of improvement
- we look beyond the cost impacts of yield improvement.

This article can be viewed as an effort to estimate the value of yield improvement.

2.2. Examples of yield-driven processes

Yield issues affect a variety of processes. In assembly processes, parts and subassemblies are integrated together over the course of a line. Subassemblies are built on feeder lines, from their own components. Various tests are performed throughout the line or at the

end of each subassembly, and overall functional performance is tested at the end of the line. At any test, if the device fails it is set aside for diagnosis, and rework or scrapping. Examples include electronic circuit boards, hard disk drives, and personal computers. Across assembly processes, a general rule is that the lower the yields and the more complex the device, the more extensive the testing. Diagnosis and rework can be very complex. For example Hampton (1996) documents a rework process, for one subassembly of a hard disk drive, with about 38 separate operations, including tests.

Are there any processes which are *not* yield-driven, considering that all industries have defects? We find it useful to classify a process as yield-driven when it meets three criteria: final inspection of all units produced, defective units cannot be sold, and first-pass yields are below 95 percent. This definition, for example, excludes automobile assembly even though new cars average greater than one defect per car. While the modern auto industry can be considered "quality-driven", we do not find it useful to consider the assembly portion of it as "yield-driven". Cars with defects can still be sold, so that the level of attention to defects is a management choice, not a technological necessity.

Integrated circuits are among the industries which clearly are yield-driven. They are fabricated in very complex processes that have over 100 stages, any of which can go wrong. The basic processing unit is the wafer, typically 8 inches in diameter and containing hundreds of individual devices called dice. After fabrication, every device on each wafer is functionally tested. The testing machines physically mark bad devices with ink. Yields at this stage (fraction of dice on the wafer that are good) are anywhere from 90 percent down to zero on some wafers. Rework of bad dice is not possible; they must be scrapped, or in a few cases, downrated. After fabrication and testing comes the "back end", usually in a different factory. Here the wafers are sawn into dice, which are assembled into individual plastic packages. The packages are heavily tested, and again there is yield fallout. Rework of assembly errors is sometimes possible, although the yield from rework can itself be low due to damage to the small, delicate devices.

In the hard disk drive industry, read-write heads are fabricated in processes similar to simple integrated circuits, although they go through a rather different and labor intensive assembly process. Media (disks) are fabricated using sputtering. Rework of bad disks is often not economical.

Many selling and other people-based processes are yield-driven. Direct mail selling starts with "raw material" of a list of names and other information. The list is screened according to some criteria (with yield fallout). Candidates who pass this screen are sent a letter. Not all the letters arrive at their destination (fallout). Each letter contains a phone number

or postcard to be sent back by the potential customer. There is usually massive yield loss at this point, as only a few percent of the people receiving the letters respond. Those who respond receive further information and go through additional stages; eventually a fraction of them buy something.

Continuous process industries such as steel, paper, and glass move bulk materials through a series of large, expensive, and highly automated machines. When the outputs (billets, sheets, bottles, etc.) fail final inspection, they cannot be reworked in the usual sense, although they can be recycled.

3. Multi-stage yield-driven production processes

In this section we discuss production processes consisting of a sequence of sub-processes, of which at least one has yield below 100%. Figure 1 provides an example. Production stages are interspersed with test/inspection points.

INSERT FIGURE 1 ABOUT HERE

Although defects can occur anywhere, they are detected mainly at the test points. An important question in designing processes with yield losses is the positioning of tests or inspections. Tests are costly, and can sometimes reduce yields themselves. There are various formulations of where to put them. Common rules are to position them before extensive or irreversible operations, at the end of modules in modular subassembly, after low yield operations (to avoid adding more value to bad units), or immediately after operations targeted for process improvement (to provide fast feedback).

At each inspection point, items are classified into "good items" and various categories of "defective items". Whereas good items can continue processing (at the next operation e.g. by moving from Test 2 to Stage 3), defective items are removed from the line. They can then be either *reworked* or *scrapped*. These items are yield losses, which reduce output in various ways. The effects differ depending on whether defective items are scrapped or reworked.

Rework means that some operations prior to the defect detection point must be redone, or defects must be otherwise repaired. For example an item would move from Test 2 back to Stage 1 or 2. Depending on the diagnosis and the defect, rework could be a single repair operation, or a whole network of operations involving disassembly, replacement, and reassembly, taking more time than the original processing. As rework typically is not perfect, it may be necessary to rework the same item multiple times. This is especially likely when diagnosis of the initial problems is difficult, as in complex electronic assembly.

Scrap occurs when bad items are discarded and final output is correspondingly reduced. Rework is generally preferable, but sometimes it is technically infeasible or uneconomic. It is uneconomic if the expected cost of rework, divided by the probability of success, is greater than the value of a good unit at that point in the process. This calculation can be tricky since both cost and value numbers should include the opportunity cost of capacity, as we will show.

3.1. Yields, capacity, and variability

When yields are below 1 at a stage, additional material must be processed at that stage either to compensate for items scrapped later, or as rework of defects where that is possible. The magnitude of this effect is inversely proportional to the yield. In many situations the increased need for capacity extends back before the stage where the yield fallout occurs. This capacity effect would come up even if yields were deterministic, e.g. for continuous processes where a fixed percentage of the material is lost.

However, in reality yields are almost never deterministic. A yield of 90% means not that every tenth item is bad, but that there is a 10% chance that a given item is bad. Thus yield losses increase variability, which is the enemy of capacity. The best stochastic case is that yields are Bernoulli, i.e. that the process has no memory. Suppose that bad items at Test 2 are immediately reworked by repeating Stage 2. Even if the actual processing time of Stage 2 is itself deterministic, the yield losses force items into multiple passes through Stage 2, and thus make the effective processing time for a *good* item a random variable. Hopp and Spearman (1996, section 12.3) show for this case that the variability (measured by the squared coefficient of variation) in the effective processing time of Stage 2 increases linearly with $(1 - y)$ where y is the yield of Stage 2.

Figure 2 illustrates the relationship between work-in-process inventory and throughput (capacity of the process) for our sequential production line in Figure 1. Assume that the processing times of Stages 1, 2, and 3 are deterministic, with production rates of 19, 20, and 15 units per hour, respectively. The operation at Stage 2 is subject to yield losses with probability $1 - y$.

For the case of no yield losses ($y = 1$), three units of WIP suffice to keep all stages busy. Adding more WIP does not result in larger throughput. The capacity of the process is solely determined by the slowest operation (the bottleneck, in this case Stage 3), and the process produces 15 parts per hour.[2]

[2]We assume that Stage 1 has sufficient raw material and the line is operate with finite buffers such

INSERT FIGURE 2 ABOUT HERE

However, for any yield less than 1, the effective processing time at Stage 2 (and thus the work flowing to Stage 3) is stochastic, and a buffer after Test 2 is needed to reduce random starvation of Stage 3. This is true even though Stage 3 remains the bottleneck for any yield at Test 2 greater than 15/20. Capacity reduction occurs, even when yield problems are at a non-bottleneck station. But by adding more and more WIP, asymptotically throughput approaches the ideal rate of 15 parts per hour. This is illustrated by the middle curve in Figure 2.

If yield at Test 2 falls below 15/20, the bottleneck shifts from Stage 3 to Stage 2. Now, capacity is reduced because of yield loss in ways which WIP cannot help. The problem is that even if it is never starved or blocked, Stage 2 can only produce 20y good units per hour on average, which is less than Stage 3. Again there is an additional stochastic variability effect due to the interaction between Stages 2 and 3. This can be compensated by allowing more WIP, so that the capacity of the process approaches 20y. This is shown as the lower line in Figure 2. With a more complex production system with yield losses at multiple points, the stochastic effects create production time variability in many places, compounding the throughput loss even if, in a deterministic process, there were only a single bottleneck.[3]

If the yield losses are only detected at Test 2, does it matter where the failures occur? It does not, but it does matter whether the rework must go through Stages 1 and 2 or only Stage 2. Stage 1 is slightly slower than Stage 2, and if it is needed for rework, then it becomes the bottleneck at y = 15/19 = .79. The uncertain processing load on Stage 1 also calls for a buffer between Stages 1 and 2. No matter how much WIP is allowed in the system and where it is, throughput can never exceed $19y$.

The amount of WIP actually needed to buffer yield-induced variation depends on the underlying distribution of uncertainty. A favorable case is a memoryless process, such as a Bernoulli process. Unfortunately, real yield losses often have persistence, due, for example, to setup errors, environmental disturbances, and bad lots of raw material. In this case the optimal buffer size after Test 2 can go up dramatically. Ideally the buffer should be large enough to cover multiple occurrences of bad events, e.g. several bad vendor lots.

as by a CONWIP approach, i.e. it only produces an additional item, if an item is released out of the system after Stage 3. If buffer size were not limited in some way, WIP would grow infinitely even in the deterministic case.

[3]We follow convention in speaking as if the process has a single bottleneck, but in a stochastic world it is also reasonable to think in terms of multiple bottlenecks, i.e. multiple stages whose capacity affects overall process capacity. What we refer to as "the" bottleneck is the most important among several.

This is rarely practical, and it is often more effective to have a system for fast detection and correction of persistent problems. These goals call for WIP to be reduced rather than increased. Fortunately, the size of allowed WIP buffers can be adjusted dynamically in most processes, in response to current needs and conditions.

Summary (Capacity effects in lines with rework)[4]

> Capacity: If yield problems are severe enough to make a machine a bottleneck (or, even worse, the yield losses are on the bottleneck machine), they substantially reduce the capacity of the process. The nominal capacity of production stage i is reduced by the yield losses at all tests from i onwards which must go through i during rework. This effect is not ameliorated by allowing buffers.
>
> WIP and variability: Lower yields cause higher variability in all stations which handle rework, further reducing effective capacity, even at non-bottleneck machines. This can be partially compensated by allowing WIP after each affected process. However this increases costs, lead times, and throughput times, and can hurt problem detection and solution, thereby reducing yields.

3.2. Further complications: scrap, batching, and capacity planning

If yield fallout is taken as scrap rather than being reworked, the effects on system capacity are even stronger, since once a unit is lost it cannot be made up. In addition, the stochastic variation in load is felt at all stages downstream of the yield loss, not just at the stages involved in the rework loop. In order to get 100 good parts at the end of the process, more than $100/y$ must be started at the beginning, where y is the cumulative yield all the way through the process.

This points to the importance of capacity planning in yield-driven processes. If yields and resulting rework requirements are known at the time a line was laid out, and remain roughly constant, then capacity planning and line balancing is done by increasing the capacity at each station enough to handle its anticipated yield-caused extra load. With scrap, it takes the form of increasing the capacity enough at all upstream stations that they can keep up with demand at the end of the process. With rework, the effects are more localized; extra capacity is needed only for the rework loop sections of the process.

Commonly, however, yields are neither known accurately in advance nor are they constant over time. Instead the aggregate yield shows both a positive trend (learning) and week by

[4]These results as well as the example in Figure 2 are adapted from Hopp and Spearman (1996).

week variation which cannot be buffered out economically, even by finished goods inventory. There is still more uncertainty and variability for individual stages and individual failure causes. Therefore once a process starts up, the actual capacity at each stage usually will be "suboptimal" by static criteria. There will be shifting bottlenecks and other capacity problems. In manual assembly situations, they can be ameliorated by temporarily shifting workers around, but in automated processes this is generally not useful, and the only solutions are either to buy additional capacity at the beginning, or live with reduced and fluctuating output during ramp-up. Again these problems tend to be more extensive with scrap than with rework, due to the localized nature of rework loops.

Note that in a system with scrap, the ratio of optimal capacity at different process stages changes systematically over time. At the beginning, by far the most capacity is needed at the beginning of the process, but as learning takes place the optimal capacity distribution shifts to a more gradual taper. In the example of Figure 1, suppose the yields at all three stages start at 80% and gradually improve to 95%. To produce 100 units of output at the beginning, stages 1 and 2 need 195 units of capacity each while stage 3 needs $195 \times .8^2$ or 125. After learning, stages 1 and 2 need 117 units of capacity while stages 3 needs 105. In the common case where capacity must be sized and purchased all at once, before process start-up, this means that at all times some parts of the process will have excess capacity relative to other parts. The line can never be balanced, even with perfect foresight!

Scrap and rework cause further reductions in effective capacity when lot sizes are fixed for equipment reasons, such as in semiconductor fabrication where the lot size is exactly one "cassette" of wafers. Once a unit falls out of a station, the remaining lot has a hole. If this hole is not filled, capacity at all downstream operations is wasted due to the smaller lot size. One approach in rework situations would be to hold the incomplete lot until the defective unit was reworked and catches back up. This, however, would mean very small rework lots, and long delays and WIP as the incomplete lot waited. Jaikumar (1988) solved this problem through a mechanism called *e-lots*, which are special lots that serve both as buffers for stochastic yields, and as replacement inventories to fill holes left by yield loss. By carefully choosing the size of the e-lot, it can serve as an SPC signal as well, differentiating between normal statistical variation in yield, and special problems. When yield loss is in the form of scrap, however, there is basically no good way to fill the holes.

A related complication arises in make-to-order situations with scrap. The customer wants a fixed number N of good items at the end of the process. In order to compensate for the expected yield losses we must start N/y at the beginning. This approach would work fine, if yields were deterministic. However since they are not, the production scheduler has to

trade off the costs of making too much against the cost of making too little. (Mathematically this is a newsboy-type problem.) Sometimes the only way to deal with undersupply, caused by unexpectedly low yields, is to run a second, small, lot all the way through the process, causing long delays to the order.

Summary (Capacity when rework not possible)

> Capacity is always reduced by scrap, even when the bottleneck operation does not shift. To compensate, additional capacity must be added at all stations upstream of yield test points, with the most capacity needed at the start of the process.[5]
>
> Uncertainty and variability in yields make perfect line balancing almost impossible. WIP buffers can help with short term variation, but they cannot help with long term shifts due to learning.
>
> Fixed lot sizes suffer additional capacity penalties due to yields.
>
> Make to order situations with scrap suffer additional penalties due to yield variability.

Our discussion translates into the following costs of yield losses. The capacity effects in Table 1 show up in two ways: increased fixed costs due to more capacity, and reduced output due to not having enough capacity. The value of lost output can be quite large. Section 4 discusses this in detail.

INSERT TABLE 1 ABOUT HERE

3.3. Cost and value at different stages of the process

Not only is the final cost of items important, but managers also need to know the way the costs change from stage to stage. More precisely, they need the value of a good item at each point in the process. One place where this is key is deciding where to concentrate process improvement efforts. A two point yield improvement has different value at different places in the process. Another is for the scrap versus rework decision. For example, suppose that after any test a defective item can be reworked for a labor cost of \$10, with a 90% chance of success and a 10% chance that the item must be scrapped. Is it better to pay for rework, or to scrap the item? Clearly if x is the value of a good item at that point,

[5]It does not matter where the defective unit is actually created, only where it is detected.

the decision rule is to rework if $10 < .9x$. Early in the process where x is small it may be better to scrap, while near the end it is better to rework.

Some general insights into these values are given by the following:

- At the beginning of the process, the value of a good item equals the cost of raw materials (which may themselves be the output of another yield-driven process).
- At the end of the process, the value is given by the marginal revenue from a good item that can be sold.
- The value of a good item increases as it moves through the process, even if no additional material is being added. Let y_n be the yield at the n'th stage. If there are no binding capacity constraints, the value leaving stage n is approximately $1/y_n$ times the sum of the value entering stage n plus variable costs at stage n.
- This gives two different ways to calculate value: cost-based working forward, and price-based working backwards. The two will be equivalent if there is no binding capacity constraint, and differ if there is one. This is a basic intuition behind the results of Section 4.
- The discontinuity in value comes at the bottleneck operation(s). After the bottleneck, value is based on selling price; before the bottleneck, it is based on cost.

These results are exact for a process with no rework loops. With rework, the calculations can get quite complex. Hampton (1996b) has examples for hard disk drives.

4. A Two-stage model

We now turn from a general discussion of yields in a multi-stage process, to a quantitative analysis of yield effects. Consider the simplified production process depicted in Figure 3. The first stage corresponds to the "normal" production process while the second stage is a special rework process where defective items are repaired to eventually meet the quality specifications. We will index our variables with in for initial production, and rew for rework process. The amount of direct labor required in stage in is denoted by L_{in} [hours/unit], with a wage rate w [$/hour]. M_{in} denotes the material costs per incoming item at stage in [$/unit]. Initial production ends with a quality inspection of every single item. We define the first-pass yield (y_{in}) as the proportion of items that pass this test and

can thus be put on the market. For the moment, we will assume that all good items are sold, at a price p.

Insert Figure 3 about here

If the item does not pass the test at the end of the initial stage, it enters a rework process where some of its components are replaced, adjusted, or repaired. Rework can be a rather complicated production process, including multiple workstations, a substantial amount of testing and diagnosing devices, and multiple passes. For our economic analysis, we do not need the microstructure of this rework process. All we need are data describing the aggregated behavior of the rework system, including:

y_{rew}: rework yield (proportion of items that are successfully reworked so that they can pass the quality test)

L_{rew}, M_{rew}: average direct labor requirements / average material costs, per part entering rework

If the rework is successful, the item can be sold at price p; thus the full functionality of the product can be reached in the rework. Otherwise the item is scrapped.

Let K be the number of items started at stage in in a given time period, e.g. a month. Of these K items, Ky_{in} can be put on the market without any rework. $K(1-y_{in})$ enter the rework process, of which $K(1-y_{in})y_{rew}$ can be reworked to become sellable output. This creates an overall (composite) yield of:

$$y_c = y_{in} + (1-y_{in})y_{rew} \tag{4.1}$$

Thus, rework raises effective yields from y_{in} to y_c, an improvement of $(1-y_{in})y_{rew}$.

In high-tech industries, rework is typically more difficult than initial production, and therefore y_{rew} is often less than y_{in}. Reasons rework is harder include the need to disassemble the defective items, which may cause damage; faulty initial diagnosis so that the real problem is not what gets repaired; and some problems simply cannot be fixed but rather the whole item must be discarded. On the other hand, the rework process can be repeated several times to improve its yield ("rework of the rework"). As a result y_{rew} can be higher than y_{in}. Typical values in assembly of a high end disk drive are y_{in}=60%, y_{rew}=70% and therefore y_c=88%.

4.1. Capacity constraints

As we discussed in Section 3, capacity issues are very important in yield-based processes. How much good output will come out of the process in Figure 3, if K units are started? There are two cases, capacity constrained and capacity unconstrained. The easier case is the one where production has sufficient capacity to keep up with demand. The factory can make as much as it can sell.

Let D denote the demand per period. To have final output D, the factory must start $K = D/y_c$ items into the initial stage. We assume that D is known and that demand is fulfilled at a market price p, which is – as a result of many other suppliers on the market – out of control of the individual company. Similarly, we do not consider any price discounts or other aspects of marketing or competitive interaction.

The capacity constraint is related to the production process, and happens especially during ramp-up periods. The factory cannot produce as much volume as it would like to. Such constraints can be a result of component shortages, limited production equipment, limited trained workers, or other factors.

Let K_{max} describe the number of units available of a scarce resource. In disk drives, K_{max} could be the number of heads available from a supplier or the overall testing capacity in the period. The K_{max} units of the scarce resource are consumed at a rate of k_{in} units per incoming item at initial production and at a rate k_{rew} per reworked item. Then for K items started, Kk_{in} units of resources will be used at stage in, and on average $K(1 - y_{in})$ items will need rework, consuming a further $K(1 - y_{in})k_{rew}$ units of scarce capacity. This must be kept at or below available capacity K_{max}, giving us:

$$K \leq \frac{K_{\max}}{k_{in} + (1 - y_{in})k_{rew}} \tag{4.2}$$

Combining the two cases, production unconstrained and production constrained, we have

$$K = Min\left\{\frac{D}{y_{in} + (1 - y_{in})y_{rew}}, \frac{K_{\max}}{k_{in} + (1 - y_{in})k_{rew}}\right\} \tag{4.3}$$

This is the number of starts into the factory. The effective output level is given by Ky_c.

4.2. The effect of yield on contribution

Our analysis focuses primarily on one performance measure: contribution per period, which we define as revenues minus variable costs (materials and labor). Contribution roughly equals profit before tax plus fixed costs, so that a one dollar change in contribution

gives a one dollar change in profits. The advantage of working with contribution per period is that it includes both cost and revenue aspects. Traditional analysis using Cost of Goods Sold (COGS) or other cost-based measures neglects the positive effects of yields on sales and revenues. We will discuss the relationship between these measures further below.

Define π as the per period contribution. Then

$$\pi = \underbrace{Ky_c p}_{revenues} - \underbrace{K\,[\overbrace{wL_{in} + M_{in}}^{initial} + \overbrace{(1-y_{in})(wL_{rew} + M_{rew})}^{rework}]}_{\text{costs}} \tag{4.4}$$

The objective of our analysis is to develop a better understanding of how changes in the wage rate w, or changes in process performance such as yields y_{in} and y_{rew}, affect the overall economics. We will do this in two steps. First, by deriving the partial derivatives of π with respect to yield rates and wage rates, we develop a number of qualitative insights about how these variables influence economic performance. Second, by providing actual case data from the hard disk industry, we estimate concrete economic values for changes in yields and wage rates.

Two cases have to be considered, depending whether the capacity constraint is binding. If it is not, we assume that input and production capacity is unlimited and the only constraint is provided by the market. In this case the number of starts required to meet demand can be computed as $K = \frac{D}{y_c}$. The second case is the opposite. The factory can sell whatever it can make, but there is a limited supply of components and capacity. The two cases are now analyzed in greater detail.

4.2.1. Analysis of Case 1: Market Limit Only

If the factory is not capacity constrained, it will make and sell D units. The number of starts is provided by $K = \frac{D}{y_c}$, which, substituted into (4.4), gives an overall per period contribution which we denote by π_{market}:

$$\pi_{market} = \underbrace{Dp}_{revenues} - \overbrace{\frac{D}{y_{in} + (1-y_{in})y_{rew}}}^{starts-to-meet-demand} \overbrace{[wL_{in} + M_{in} + (1-y_{in})(wL_{rew} + M_{rew})]}^{\text{cost}-per-start} \tag{4.5}$$

The first term in (4.5) describes the revenues and the second term the direct costs. Note that the second term in (4.5) can be interpreted as the cost per good unit, which is cost-per-start divided by composite yields.

There are several interesting questions about the production process, which this model allows us to answer. Specifically, what is the effect of increases in the yield at each stage?

What effect do wages have on the contribution? Which of the two effects dominates under what conditions? We can approach these questions by looking at the derivatives of contribution with respect to each of these variables. First, consider wages.

The partial derivative of contribution with respect to wage rate can be computed as:

$$\frac{\partial \pi_{market}}{\partial w} = -D\frac{L_{in} + (1 - y_{in})L_{rew}}{y_{in} + (1 - y_{in})y_{rew}} \tag{4.6}$$

From (4.6) we see that an increase in wage rate reduces the contribution proportionally to $(L_{in} + (1 - y_{in})L_{rew})$, a factor that describes the expected amount of labor that is spent per start.

Second, consider the partial derivative with respect to first-pass yield:

$$\frac{\partial \pi_{market}}{\partial y_{in}} = \frac{D}{\left[y_{in} + (1 - y_{in})y_{rew}\right]^2}\left[wL_{in}(1 - y_{rew}) + M_{in}(1 - y_{rew}) + wL_{rew} + M_{rew}\right] \tag{4.7}$$

As yield enters the contribution expression only through costs, the partial derivative does not include any changes in revenues. Thus, in the market limited case the yield increase pays off purely in the form of cost reduction. This cost reduction effect can be decomposed into

- savings in the rework process at rate $M_{rew}+wL_{rew}$ as fewer items have to be reworked
- savings in the production process, as, with increased yields, less items have to be started in order to fulfill the same amount of demand D.

Finally, the partial derivative with respect to rework yields is similar to (4.7) and can shown to be

$$\frac{\partial \pi_{market}}{\partial y_{rew}} = \frac{D(1 - y_{in})}{\left[y_{in} + (1 - y_{in})y_{rew}\right]^2}\left[wL_{in} + M_{in} + wL_{rew}(1 - y_{in}) + M_{rew}(1 - y_{in})\right] \tag{4.8}$$

The impact of a yield increase on the production cost $(wL_{in} + M_{in})$ has to be scaled with a factor $(1 - y_{in})$: as now more demand is filled with reworked parts, less of the normal production is needed. Rework costs decrease also, because with less starts required, fewer items are likely to enter the rework process.

4.2.2. Analysis of Case 2: Capacity Constraint

Binding constraints on both raw material and equipment are typically observed during the ramp-up of a new process. As an example, consider the case of disk-drives. For a

new generation disk, the key components are redesigned, pushing the component supplier to the frontier of what currently is producible. This translates into low yields and low production volumes at the supplier, and ultimately to a shortage of components for the disk manufacturer. An example of this is the shortages of platters (media). If a drive contains four platters, and per reworked drive on average one platter must be replaced, we have k_{in} =4 and k_{rew}=1.

Another scarce resource during the ramp-up is the capacity of testing equipment.Given the substantial cost of testing equipment, testing is frequently the bottleneck activity in the overall process. Rework is especially testing intense, and one drive can need four hours of testing (k_{rew}=4). For the normal production process, the numbers are lower, but still reach one hour per start (k_{in} =1).[6]

If the factory is constrained to K_{max} units of the scarce resource per period(e.g. 100,000 available heads per month, 10,000h available testing time), the number of starts is restricted to $K = \frac{K_{max}}{k_{in}+(1-y_{in})k_{rew}}$, as stated in (4.2). Substituting this into (4.4) gives an overall per period contribution of

$$\begin{aligned} \pi_{capacity} &= \overbrace{\frac{K_{\max}}{k_{in}+(1-y_{in})k_{rew}}\left[y_{in}+(1-y_{in})y_{rew}\right]p}^{\text{revenues}} \\ &\quad \underbrace{- \frac{K_{\max}}{k_{in}+(1-y_{in})k_{rew}}\left[wL_{in}+M_{in}+(1-y_{in})(wL_{rew}+M_{rew})\right]}_{\text{production costs}} \end{aligned} \tag{4.9}$$

where $\pi_{capacity}$is defined as the per period contribution in the capacity constrained case. As in (4.9), the first term is revenue, and the second is production cost.

As before, the effect of a change in wage rate is straightforward:

$$\frac{\partial \pi_{capacity}}{\partial w} = -\frac{K_{\max}}{k_{in}+(1-y_{in})k_{rew}}\left[L_{in}+(1-y_{in})L_{rew}\right] \tag{4.10}$$

which again is linear in the expected amount of labor per start.

The effect of first pass yield is more complicated, with a partial derivative of:

$$\begin{aligned} \frac{\partial \pi_{capacity}}{\partial y_{in}} &= \frac{K_{\max}p}{\left[k_{in}+(1-y_{in})k_{rew}\right]^2}\left[k_{rew}+(1-y_{rew})k_{in}\right] \\ &\quad + \frac{K_{\max}}{\left[k_{in}+(1-y_{in})k_{rew}\right]^2}\left[k_{in}(wL_{rew}+M_{rew})-k_{rew}(wL_{in}+M_{in})\right] \end{aligned} \tag{4.11}$$

[6]In fact, k_{in} and k_{rew} tend to fall during ramp-up. Initially, drives must be "extra-tested" to ensure all problems are caught. Later, engineers learn how to test more narrowly for specific problems, allowing faster testing.

Note that unlike the market constrained case, a change in yield now affects the revenue as well as the costs. Thus during the ramp-up, an improved yield not only reduces unit costs, but also creates an increase in net capacity. This is captured in (4.11) where the first expression describes the increased revenue resulting from a better usage of the K_{max} units of the scarce resource. The revenue effect comes from two sources. One is the direct effect of yield improvement on output. The second is an indirect effect of reducing capacity consumption during rework, which allows an increase in starts, and therefore in output and revenues.

> **Result (Capacity effect of yields)** When production capacity is constrained, higher first pass yields have extra leverage because they increase gross revenues, as well as decreasing the cost per unit. The increased revenue comes from more output per start (direct effect) and from less capacity consumed per start (indirect effect).

Finally, consider the partial derivative of contribution with respect to y_{rew}:

$$\frac{\partial \pi_{capacity}}{\partial y_{rew}} = \frac{K_{\max} p}{k_{in} + (1 - y_{in}) k_{rew}} (1 - y_{in}) \tag{4.12}$$

Similar to (4.11) we see that for the capacity constrained case, a yield improvement creates an increase in net capacity (and thus in revenues). There is no impact on total costs, as once an item has entered the rework process, the investment for labor and material is determined. The same holds for capacity consumption, as once the item has entered rework, it uses up k_{rew} units of capacity. Thus, the only effect of y_{rew} is on output. Of course, the average cost per good unit falls as a result of the increased output.

Table 2 summarizes the effects of wage rate and yield changes on contribution per period. First, we see that the effect of wage rate is the same for both cases, and is linear in the expected amount of labor time per item started. Second, depending which of the two constraints on starts is binding, an improvement in first-pass yield creates cost reduction and an increase in revenue (capacity is binding) or a cost reduction effect only (demand is binding).

`Insert Table 2 about here`

Typically, in industries with long product lifecycles and few introductions of new products, the cost benefit is dominant. This explains why previous OM literature has largely emphasized cost aspects of yield management. In industries with short lifecycles and a high rate of product replacements, however, the importance of the revenue effect is higher and

one major effect of yield improvement is an increase in net capacity. As Table 2 shows, this effect is even stronger than linear. The relative importance of these two effects will be examined for the disk-drive industry in Section 5.

The ambiguous effect of yield improvement on production costs can lead to incentive problems. The second expression in equation (4.11) defines the change in the overall labor and material costs. Interestingly, total variable costs can rise or fall. Suppose

$$\frac{(wL_{rew} + M_{rew})}{k_{rew}} < \frac{(wL_{in} + M_{in})}{k_{in}}$$

That is, rework cost per unit of capacity consumed in rework is larger than the production cost per unit of capacity consumed in production. Then the overall costs actually increase rather than decrease if yields improve. An example of such situation would be the case of testing devices being the capacity constraint. An improved first-pass yield frees up testing capacity in the rework process. This capacity can be reallocated to the normal production process, enabling the factory to increase its starts, which ultimately leads to an increase in total costs. This effect, of course, is more than compensated by the increase of revenue, so that in any case there is an economic gain from the yield improvement.

This situation where both costs and profits go up can have perverse incentive effects on production managers. In one company we studied, overhead costs of the factory got allocated to production lines based on their direct production costs. The manager of a production line who improved her yield had higher total cost, and got punished through a larger allocation of overhead costs. The revenue and profit enhancing benefits of the yield improvement were not included in the accounting system. Note that such perverse incentive effects are created by cost-driven measures. Thus, any measure that does not take into account the revenue aspects of changes in the production process is misleading.[7]

4.3. When to rework?

So far, we have assumed that rework is always desirable. Sometimes, rework is technically infeasible - rework yield is zero. All our equations hold for this situation by setting:

$$0 = y_{rew} = k_{rew} = L_{rew} = M_{rew} \tag{4.13}$$

Not surprisingly, the effects of first-pass yield become larger when rework is infeasible.

[7] A further complication is that initial production and rework costs may be charged to different cost centers. If yields go up, rework costs come down but initial production costs always rise, possibly further embarassing the manager whose yields improved.

More interesting is the case where rework is technically feasible, but uneconomic. It turns out that the criteria on whether to rework depend on all the cost and profit parameters, including whether the factory is market or capacity limited. We look at the market limit case first. In this situation, the decision rule on when to rework is to compare contributions with and without rework. More precisely, choose the larger of π_{market} in (4.5) compared with π_{market} when (4.13) is substituted into (4.5). After manipulation, this leads to the rule:

Do rework if

$$\frac{wL_{in} + M_{in}}{y_{in}} > \frac{wL_{rew} + M_{rew}}{y_{rew}} \tag{4.14}$$

In other words, compare the expected *cost* per good unit from new production with the expected *cost* per good unit from rework. Since, in most situations, $M_{rew} \ll M_{in}$ and labor costs are small compared with material costs, this means that rework is almost always a good idea, unless it has very low yield compared with initial production.

A generalization of (4.14) covers the case of multiple defect types or symptoms. Rework each defect type j for which (4.14) holds with appropriate values of y_j, L_j, M_j on the right hand side. Sometimes, a few defect types are so unlikely or expensive to repair, that units with that defect type should be scrapped.

The logic for analyzing the capacity constrained case is similar, with (4.9) the relevant contribution equation, but the result is very different. The decision rule becomes:

Do rework if

$$\frac{py_{in} - (wL_{in} + M_{in})}{k_{in}} < \frac{py_{rew} - (wL_{rew} + M_{rew})}{k_{rew}} \tag{4.15}$$

The numerators of (4.15) are the expected contribution per unit started into new production (if no rework is done) and into rework. So (4.15) amounts to comparing the *contribution per unit of scarce capacity*. Thus the decision of when to rework is based on very different criteria in the market limited and capacity constrained cases.

It is possible that the decision of what to rework will change during ramp-up. For example, at the start components may be the scarcest resource, with rework needing many fewer components than new build ($k_{rew} \ll k_{in}$) so that rework is good. Later, as the vendor ramps up component production, test capacity may be the scarcest resource, with $k_{rew}/k_{in} = 4$, and selling prices high so that the expected contribution is about the same for rework as for new build. In this situation, rework reduces profit. Finally, when ramp-up ends, capacity is no longer a constraint, expected cost becomes the deciding criterion,

and rework is again desirable.

More realistically, instead of on/off all-or-nothing policy shifts for rework, the decisions about which kinds of defects and how to rework them will change in increments over time. In the beginning, rework almost everything, at least enough to salvage all scarce components. In the middle, rework only a few defect types. At the end, rework most defects, but perhaps still scrap units with symptoms that are hard to diagnose, so that probability of success is low.[8]

> **Result (Decision to rework)** The choice of whether to rework or scrap a defective unit is an economic one, and may change over time. Even the correct criteria to use depends on whether production is capacity constrained.

4.4. When are yields most potent?

An interesting question is "When should managers pay the most attention to yields? What conditions lead to a high payoff from yield improvement?" The first answer is that yield improvement has the biggest impact when production capacity is a binding constraint. In this situation, yield improvement raises sales as well as reducing cost per unit. It is even more valuable when there is a high gross margin for the product, since additional output from better yield is multiplied by the gross margin to get the contribution impact.

This capacity result has to be qualified in two ways. First, the transition from the capacity constrained to the capacity-unconstrained case is actually more gradual than in our model. In most industrial settings, extra capacity has a positive value even when capacity is higher than demand. It can be used to reduce lot-sizes, reduce shop-floor congestion, respond faster to orders, or run controlled trials to further reduce yields. Hence, a more accurate statement is "The higher demand is relative to capacity, the higher the value of yield improvement."

The second qualification is a technical one. When we say the value of yield improvement is higher when production capacity is scarce, this is true if other conditions are the same. In particular compare two situations, one where capacity is greater than demand D and a second one where demand is still D but capacity is only a small fraction of D. Then, it is

[8] A further complication is that static profit maximization is not the only reason to rework. Ramp-ups should be managed for rapid learning and yield improvement. (Jaikumar and Bohn 1992) Under this condition, rework can also give useful information about how to change the product or process to further improve yields. This means working on failure types which, at present, are not profitable to rework according to the above formulas.

possible that the cost savings of the first situation outweigh the revenue enhancement of the second situation. This is a result of cost reductions being "leveraged" by the amount of production K in equation (4.7). However, as long as production capacity is roughly the same in the two cases, the capacity constrained case has a larger effect from yield changes.

Another situation that makes yield improvement more valuable is where rework is difficult, meaning it is low yielding, is expensive, or consumes a lot of capacity relative to initial production. In the extreme version of this situation, it is better to scrap defective units rather than attempt to rework them, as discussed. To see why difficult rework makes it more valuable to improve yields, consider the opposite situation. If rework is almost free, always works, and consumes no scarce capacity, then bad first pass yield can be made up with very little penalty.

5. The disk-drive case

We now present a case study using our model to gain detailed insights about the economics of yields for the disk-drive industry. Table 3 summarizes the key model parameters. The numbers are appropriate for a high-end drive. With these parameters, the average cost per good drive is $175.50.

Insert Table 3 about here

To explore the effects of yield and wage changes on contribution, we now reproduce Table 2 with actual cost data (Table 4). For the case where the factory can produce all the demand (first column), we assume that 150,000 units can be sold per month. For the next column we are working with 150,000 read-write heads available and for the last column, the constraint is given by 200,000 hours of available testing time. Table 4 shows the impacts of different process changes. As a result of the different constraints, the output varies across columns: 150,000 in the capacity-unconstrained case, 120,000 for the read-write heads, and 98,000 for the case when the testing equipment provides the binding constraint. All numbers are thousands of dollars per month of contribution, unless stated otherwise.

Insert Table 4 about here

The effect of a reduction in wage rate is straightforward. As seen in (4.6) and (4.10) a change in wage rate has a linear effect on contribution. The contribution gains are, depending on the availability of capacity, between 160,000 and 245,000 dollars per month.

The effect of a 5% improvement (5 percentage points, from 60% to 65%) in first pass yields is more complicated. Consider the market constrained case first. The yield improvement

results in a cost improvement at both regular production and rework. This translates into a reduction of 4.90 dollars per good unit, or, in other words, a $735,000 /month improvement in contribution and a 2.7% cost reduction.

For the read-write head constrained case, there is no improvement in production costs. Some savings can be achieved in the rework process, but the costs in the regular production go up rather than down. This is a result of having 1.15 percent more starts, enabled by a better usage of the read-write heads. More important, the increase in starts provides a 2.87 percent increase in output (and thus in revenue), corresponding to a one million dollar increase in monthly contribution. These effects are even stronger in the testing constrained case. Although production costs increase by over .8 million dollars per month, this is more than off-set by a two million dollar increase in revenues, making up for a net-benefit of 1.4 million dollars per month.

Note that depending on what measure is used (costs vs. overall), the economic evaluation of changes differs substantially.

- Looking at production cost alone is obviously misleading, since in the last column (the biggest improvement in contribution) production costs actually increase.
- Working with unit costs, however, is also misleading. In all three cases, unit costs went down by 4.90 dollars per unit. Looking at the economic performance, however, we can see that the contribution effect in column three is actually twice as strong as in the first column: 1451 versus 735.

To capture the overall effect, any metric used for evaluating changes in the production must include revenue aspects.

The effect of an improvement in rework yield is similar. Note again that the improvements are stronger for the capacity constrained cases than for the market limited case. In addition to the first three rows, which illustrate the partial derivatives presented above and thus mirror Table 2, we also show the consequences of three other process improvements.

- First, in most cases yield figures change in concert across stages, as a result of a better understanding of the underlying process. To illustrate this, consider a rise in first-pass yield and in rework yield of 5%. (The composite yield, however, only goes up by about 3%.) Such an improvement creates an economic benefit of 1,226,000 to 2,069,000 dollars per month depending on the capacity constraint. This is a rise of between 6.5 percent and 17 percent of contribution.

- Second, what would it pay to design the product or automate production, so that its labor content is reduced by 20 minutes per unit? This has an effect similar to wage reduction, creating cost savings from $320,000 per month to $490,000 per month.
- Finally, consider the effect of having 10% more units of the scarce resource (e.g. testing capacity). In the market limited case, added resources do not change anything. For the other two cases, all variables in the model, including the overall contribution, are scaled by a factor 1.1, giving an increased contribution of $1,494,000 per month and $1,217,000 per month.

Result (Wage versus yield improvement) Wage reduction (from $6/hour to $5/hour) has comparatively little impact on contribution. Yield improvement effects, especially in first pass yields, are stronger, especially in capacity constrained (ramp-up) situations.

5.1. Making location decisions

When companies make decisions about new plants and processes, they often have to choose among a range of geographic locations, technologies, and workforces. These choices often affect yields, as well as the more visible fixed and variable costs. We now illustrate the resulting tradeoffs, using the example of disk drive assembly. The disk-drive industry is characterized by a strong separation into two geographic clusters: most product development is done in the US, whereas 65% of the assembly is done in Southeast Asia, especially Singapore. Further, there is a trend towards moving some manufacturing to countries with even cheaper labor, such as the Philippines and mainland China. For a detailed description of the global patterns of this industry, see Gourevitch *et al.* (1997). In many cases, moving into a new country has the potential to affect yields, particularly during ramp-up of advanced products. Workers and engineers in the new factory will not be as fast at debuging problems, or as able to communicate with developers for joint problem solving. In addition, infrastructure differences among countries may also affect ramp-up and yields. The effects of moving manufacturing to benefit from cheap labor is connected to the questions we raised during our presentation of our model. What effect do wages have on the contribution? What is the effect of yields? Which of the two effects dominates under what conditions?

Mathematically speaking, the first two questions have been addressed above. The third one requires a relative comparison between partial derivatives. The following equations show

the ratio between the change in contribution from yield improvement and the change in contribution from wage rate reduction. These apply to the disk-drive example introduced in Table 3. As before, we have to distinguish between the market-limited case and the capacity constrained case:

$$\begin{aligned} \frac{\partial \pi_{market}}{\partial y_1} / \frac{\partial \pi_{market}}{\partial w} &= \frac{149}{-245} = -0.6 \\ \frac{\partial \pi_{capacity}}{\partial y_1} / \frac{\partial \pi_{capacity}}{\partial w} &= \frac{274}{-160} = -1.71 \end{aligned}$$

In other words a 1% improvement in initial yield has the same value as a \$.60/hour reduction in wage rate, in the market limited case. From the numbers, we see that the yield effect is relatively dominating, especially in the capacity constrained case.

To further explore wages versus yields, we plot the monthly contribution over a range of possible yields for different wage levels in Figure 4. We assume for this graph that rework and first pass yields move in concert. The lower curve corresponds to a wage rate of \$20 per hour, a level representative of the US or Europe. The middle curve is based on a wage rate of \$6 per hour, which is about a typical number in Singapore. The upper curve is the extreme case where labor is free, thus the wage rate is equals to zero. Note the distinct change in slope at about 85% yield. This is where production capacity catches up to market demand.

`Insert Figure 4 about here`

There are three directions to look at the graphs in Figure 4. Direction (a) is similar to Table 4. It shows how a yield improvement influences contribution. The left of the two (a)-arrows is at a low yield. This typically corresponds to a new product. With an increase in yields, the contribution goes up quickly, until at some point market demand is satisfied. Hereafter, any further improvement in yield has a much lower effect because it does not increase revenue.

Direction (b) corresponds to decreasing the wage rate from \$20 per hour to \$6 per hour and ultimately down to \$0 per hour. The effect is constant in the sense that regardless of the yield level (or the stage in the lifecycle if we take a more dynamic perspective), it gives the same improvement in contribution.

Finally, direction (c) shows the relative comparison between yields and wage rates. Consider the left arrow labeled c first. By moving from the beginning to the end of the arrow, we see that an 8% improvement in yields corresponds to moving from a high wage country into a country with zero (!) wages, and still getting the same contribution. Improve the process by 8 percentage points and get all labor for free! If we look at the right arrow

labeled (c), this picture changes dramatically. At high yields a much larger yield improvement is needed to compensate for any wage hike. The reason is that capacity is no longer constrained, so that yield affects costs but not revenue. This confirms our earlier analysis that yield effects are large relative to wage effects, especially when capacity is constrained[9].

To what extent is there actually a tradeoff between wage rates and yields in hard disk drives? Evidence on this is sketchy and anecdotal, in part because of the general confidentiality of yield information, and in part because it is a lot easier to measure wage effects of a workforce than to measure yield effects. One disk media company, HMT, says publicly that it manufactures in California because it is easier to ramp up new products to high yield quickly there. However, many of HMT's competitors are building their capacity additions near their customers' assembly plants in SE Asia. In assembly, there is general agreement that Singapore has assembly capability and yield as good as anywhere in the world. But whether Singapore is significantly better than Thailand or China is disputed. Even plant by plant comparisons of average yields within a company are misleading, because the higher wage country's plants generally are given more technically demanding and shorter-lived products. Therefore, we can only say that the effects of yields should be evaluated at the same time as other consequences of factory location, and are likely to have a similar magnitude of impact on the bottom line.

5.2. Making automation decisions

Figure 4 was generated by assuming that the labor hours per drive remained constant, and the wages per hour changed. Since only the product of these two factors determines labor cost, it can be reinterpreted as showing what happens when the process is partially or fully automated. The wage = 0 line corresponds to a fully automated process.[10]

Automation generally improves yields, especially as components get smaller and smaller. For automation, we interpret Figure 4 as showing not how automation should be *traded-off* against yields, but as how automation should be *evaluated* with respect to both yields and labor costs. For example, in a $20 per hour wage factory where automation will reduce labor requirements by half (i.e. halfway betwen the $20 and $0 lines in Figure 4), a

[9]The reader may notice that Figure 4 has steeper slopes than some of the earlier calculations. The reason is that the Figure is based on both initial and rework yields improving simultaneously. This is commonly what happens - new knowledge benefits both.

[10]One partnership in the industry, MKE/Quantum, is noted for running "lights-out" factories. Seagate tends to use highly labor instensive methods, while others are in-between. A more sophisticated model of automation would look at differential automation levels for production and rework. Our equations can be used in this way, but Figure 4 implicitly assumes the same degree of labor displacement for both stages.

yield improvement of four percentage points will double these benefits during the ramp-up period. Once capacity catches up with demand, the labor-saving benefits continue while the yield-improving benefits get smaller.

Again, we interpret this as saying not that yields outweigh other factors, but that they are of roughly the same magnitude, and should be analyzed just as carefully as labor costs and capital costs when automating.

6. Discussion and Conclusion

6.1. Concluding Implications

This paper has provided a formal model that shows how yields drive manufacturing economics in some industries. By applying the model to the specific example of one process, hard disk drive assembly, we compared the economic values of wage reduction, yield improvement, and automation. Our findings have managerial and even political implications. We find that the effect of yield improvement in increasing contribution and profit can be very strong. Especially during ramp-up periods of scarce resources or capacity, it is critical to focus attention on yields.

Accounting systems are quite poor at dealing with yield issues, both prospectively and retroactively. Scrap costs are often treated as a separate cost pool, which is not carefully allocated back to individual points in the process. Even more basic, accounting systems only look at the cost-based numbers, not the price-based values. Sensitivity analysis on the effects of alternative production methods with different yields is very difficult with conventional cost accounting. Because of these problems with accounting numbers, experienced managers in yield-driven industries often rely on intuition for relevant decisions, while inexperienced managers make mistakes. Even the decision on what to rework and what to scrap, seemingly a technical decision, turns out to be an economic choice, and one not capturable in a cost-based accounting system.

We have applied the analysis to the location decision of a disk-drive manufacturer. At least during the initial phases of the product lifecycle, there is no wage rate low enough to compensate for even modest yield losses. A yield drop of 8% has a bigger effect on contribution than going from twenty dollar per hour wages down to free labor. Thus the quality of work done in a specific location by a specific labor force is more important than their wage. Once the product is mature, wages become more important relative to yields, and in some situations a cheaper labor force could be justified even if it reduced yields.

However, the calculations have to be done explicitly. Even a "productivity adjusted wage rate" will not properly adjust for the effect of a cheaper labor force on yields, since yields affect material costs and revenues, not just labor costs.

Of course, wages and yields are not the only things affected by siting decisions. We have investigated overseas factory siting by the hard disk drive industry, and find that a number of other cost and non-cost factors, such as tax incentives, appear to be important (Gourevitch *et al.* 1997). Thus when a country and a firm consider tax incentives, they should investigate yield issues in as much detail as labor cost issues, and weigh their effects against other criteria.

An analogous situation exists for choice of technology, such as the extent and nature of automation. Many automated technologies affect yields, often for the better. The yield effects of a technology can easily be more important than its effects on labor costs.

6.2. Further research

Our model is static in that our variables were treated as fixed, rather than changing over time. Yet change is a key element of short life-cycle production processes. Further research is developing a dynamic model of the phenomena including learning, price reduction, and changes in market demand. As a process successfully ramps up yields and capacity, value of yield improvement falls, either gradually or abruptly. Over the life cycle of a product, how much net present value is gained by higher initial yields, faster yield learning, or faster capacity ramp-up? What does this say about where managerial attention should be directed?

Next, we need to explore other industries and processes to see the relative importance of capacity, rework, initial yields, and the other factors in our model. For example, component fabrication segments of the hard disk drive will have rather different economics than assembly.

Finally, our focus in this article has been on production, simplifying the competitive environment of the producing firm. If firms improve yields more rapidly, they have some choice about whether to exploit this as higher margins or for lower prices.

In addition to the managerial lessons discussed above, we believe this paper has political implications. Our findings suggest that U.S. wage rates are not very relevant to "bringing manufacturing jobs back to America". The key in the high-tech manufacturing game can be found in yields and speed of bringing products into volume production. These are results of the organization's understanding of the production process. Therefore training

and education of all levels of the current and future workforce, as well as direct development of new technological capabilities at the organization and national levels, are crucial.

7. References

Barad, M. and G. Bennett, 1996, "Optimal yield improvement in multi-stage manufacturing systems", European Journal of Operational Research, Vol. 95, No. 3, 549-655.

Bohn, R. E., 1994, "Measuring and Managing Technological Knowledge," Sloan Management Review, Vol. 36, No. 1, 61-73.

Bohn, R. E., 1995a, "The Impact of Process Noise on VLSI Process Improvement," IEEE Transactions on Semiconductor Manufacturing, Vol. 8, No. 3, 228-238.

Bohn, R. E., 1995b, "Noise and Learning in Semiconductor Manufacturing," Management Science, Vol. 41, No. 1, 31-42.

Breaux, L. and D. Kolar, 1996, "Automatic defect classification for effective yield management", Solid State Technology, Vol. 39, No. 12, 89-96.

Burggraaf, P., 1996, "Yield analysis software solutions [IC manufacture]", Semiconductor International, Vol.19, No.1, 79-85.

Chea, K. S., 1997, "Pilot Production, Transition and Ramp-up of a New Product across an International Boundary: Hard Disk Drives in Singapore", Master thesis in the Program of Advanced Manufacturing, University of California at San Diego, CA 92093.

Chen, J. M. H., A. Mandelbaum, A. Van Ackere and L. H. Wein, 1988, "Empirical Evaluation of A Queueing Network Model for Semiconductor Wafer Fabrication," Operations Research, Vol. 36, No. 2, 202-215.

Denardo, E. V. and C. S. Tang, 1997, "Control of a Stochastic Production System with Estimated Parameters", Management Science, Vol. 43, No.9, 1296-1307.

Gourevitch, P., R. E. Bohn and D. McKendrick, 1997, "Who is US? The Nationality of Production in the Hard Disk Drive Industry", The Data Storage Industry Globalization Project, Report 97-01, University of California at San Diego, CA 92093.

Hampton, S, 1996a, "Engineering a Cost Analysis System for Hard Disk Manufacturing" Master of Science, University of California, San Diego.

Hampton, S., 1996b, "Process Cost Analysis for Hard Disk Manufacturing", The Data Storage Industry Globalization Project, Report 96-02, University of California at San Diego, CA 92093.

Hopp, W. J. and M. L. Spearman, 1996, "Factory Physics: Foundations of Manufacturing Management", Irwin.

Jaikumar, Ramchandran, 1988, "Contingent Control of Synchronous Lines: A Theory of JIT" Harvard Business School working paper 88-061.

Jaikumar, R. and R. Bohn, 1992, "A Dynamic Approach to Operations Management: An Alternative to Static Optimization," International Journal of Production Economics.

Juran, J. M. and F. M. Gryna, 1993, "Quality Planning and Analysis", 4th edition, McGraw-Hill.

Kantor, P. B. and W. I. Zangwill, 1991, "Theoretical Foundation for a Learning Rate Budget", Management Science, Vol. 37, No. 3, 315-330.

Khera, D., M. W. Cresswell, L. W. Linholm, G. Ramanathan, J. Buzzeo and A. Nagarajan, 1994, "Increasing profitability and improving semiconductor manufacturing throughput using expert systems", IEEE Transactions on Engineering Management, Vol. 41, No. 2, 143-151.

Lapré, M. A., A. S. Mukherjee and L. N. Van Wassenhove, 1996, "Behind the Learning Curve: Linking Learning Activities to Waste Reduction," Working Paper 96/24/TM, INSEAD, 77305 Fontainebleau, France.

Leachman, R. C., 1996, "Competitive semiconductor manufacturing survey: Third report on the results of the main phase," Berkeley report CSM-31, University of California at Berkley, CA 94720.

Mukherjee, A. S., M. A. Lapré and L. N. Van Wassenhove, 1995, "Knowledge Driven Quality Improvement," Working Paper 95/48/TM, INSEAD, 77305 Fontainebleau, France.

Ou, J. and L. M. Wein, 1995, "Dynamic scheduling of a production/inventory system with by-products and random yield", Management Science, Vol. 41, No. 6, 1000-1017.

Seshadri, S. and J. G. Shanthikumar, 1997, "Allocation of chips to wafers in a production problem of semiconductor kits", Operations Research, Vol. 45, No. 2, 315-321.

Stamenkovic, Z., N. Stojadinovic and S. Dimitrijev, 1996, "Modeling of integrated circuit yield loss mechanisms", IEEE Transactions on Semiconductor Manufacturing, Vol. 9, No. 2, 270-272.

Tang, C. S., 1991, "Designing an Optimal Production System with Inspection", European Journal of Operational Research, Vol. 52, No. 1, 45-54.

Wang, P., F. Lee, K. M. Chan, R. Goodner and R. Ceton, 1996, "Yield enhancement in a high-volume 8-inch wafer fab. II. Yield enhancement programs", Semiconductor International, Vol. 19, No. 8, 217-222.

Wein, L.M., 1992, "Random Yield, Rework and Scrap in a Multistage Batch Manufacturing Environment," Operations Research, Vol. 40, No. 3, 551-563.

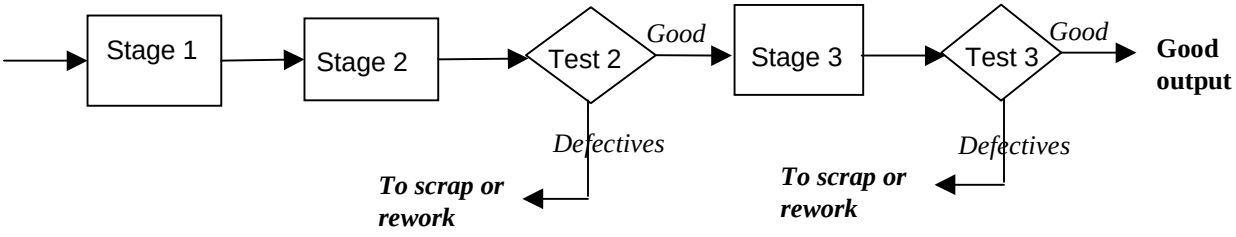

Figure 1: Example of a sequential production process

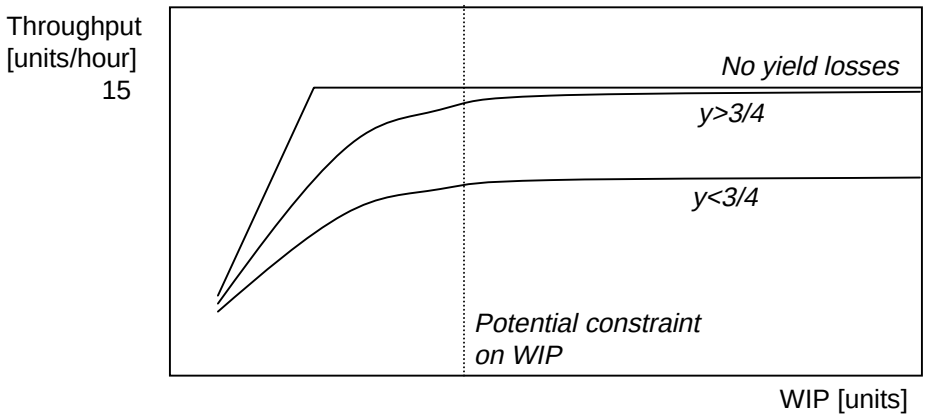

Figure 2: Throughput versus WIP for different yield levels

	Rework is done	**Scrap**
Material related costs	Incremental material to replace bad components	All material up to failed test is lost
Labor related costs	Rework labor	All labor up to failed test is lost
Capacity related costs	More capacity needed in the rework loops of process	More capacity needed at all stages upstream of failed tests
Variability related costs	WIP cost to buffer variability	WIP still needed but less effective; more capacity needed to counteract
	Lead time variability in make to order systems	Extra large lots needed in make to order systems
		Line never perfectly balanced; more capacity needed to counteract

Table 1: Summary of yield effects on cost

	Limited Market	**Capacity constrained**
Effect of wage rate reduction	+ Linear increase of contribution per period (proportional to the expected labor time per started item)	
Effect of increase in first-pass yield	+ reduction in stage 1 costs + reduction in stage 2 costs => *lower total and unit costs*	+ more output (direct effect) + better use of scarce resource, this allows more starts (indirect effect) => *overproportionally more revenue* + reduction in stage 2 costs - increase in stage 1 costs => *lower unit costs, total costs unclear*
Effect of increase in rework yield	+ reduction in stage 1 costs + reduction in stage 2 costs => *lower total and unit costs*	+ more output (direct effect) => *linearly more revenue,* *lower unit costs*

Table 2: Effects of wage rate and yield changes on contribution per period

	Initial Production	**Rework**
Material Cost	135 [$/drive]	27 [$/drive]
Direct Labor	.9 [h/drive]	1.35 [h/drive]
Yield Rate	60 [%]	70 [%]
Testing Time	1 [h/drive]	2 [h/drive]
Set of Heads	1 [unit/drive]	0.25 [unit/drive]
Selling Price	300 [$/drive]	
Demand	150,000 [drives/month]	
Wage rate	$6 per hour	

Table 3: Typical hard disk-drive data

	Limited Market	**Capacity constrained (read-write heads)**	**Capacity constrained (testing equipment)**
Output per month	*150,000 units*	*120,000 units*	*98,000 units*
Revenue per month	*45,000*	*36,000*	*29,333*
Contribution per month	*18,675*	*14,940*	*12,173*
Effect of $1/h wage reduction	*+ 245*	*+ 196*	*+ 160*
Effect of 5% increase in the first pass yield (60% to 65%)	*+ 401 (at stage 1)* *+ 334 (at rework)* ⇨ *735 (overall)* *(unit cost reduced by $4.90/disk)*	*2.87% more output* *1.15% more starts* ⇨ *1034 (in revenues)* *- 220 (at stage 1)* *+ 220 (at rework)* ⇨ *0 (in costs)* ⇨ *1034 (overall)*	*7.69% more output* *5.88% more starts* ⇨ *2254 (in revenues)* *- 917 (at stage 1)* *+ 114 (at rework)* ⇨ *-803 (in costs)* ⇨ *1451 (overall)*
Effect of a 5% increase in the rework yield	*+ 532 (at stage 1)* *+ 53 (at rework)* ⇨ *585 (overall)*	*818 (overall)*	*666 (overall)*
Effect of a 5% improvement in y_{in} and y_{rew}	*1226*	*1758*	*2069*
Effect of 20min. lower labor content L_{in}	*490*	*392*	*320*
Effect of 10% more units for the resource K_{max}	*0*	*1494*	*1217*

Table 4: Effects of wage rate and yield changes in the disk-drive case (All numbers are thousands of dollars of contribution per month, unless otherwise noted)

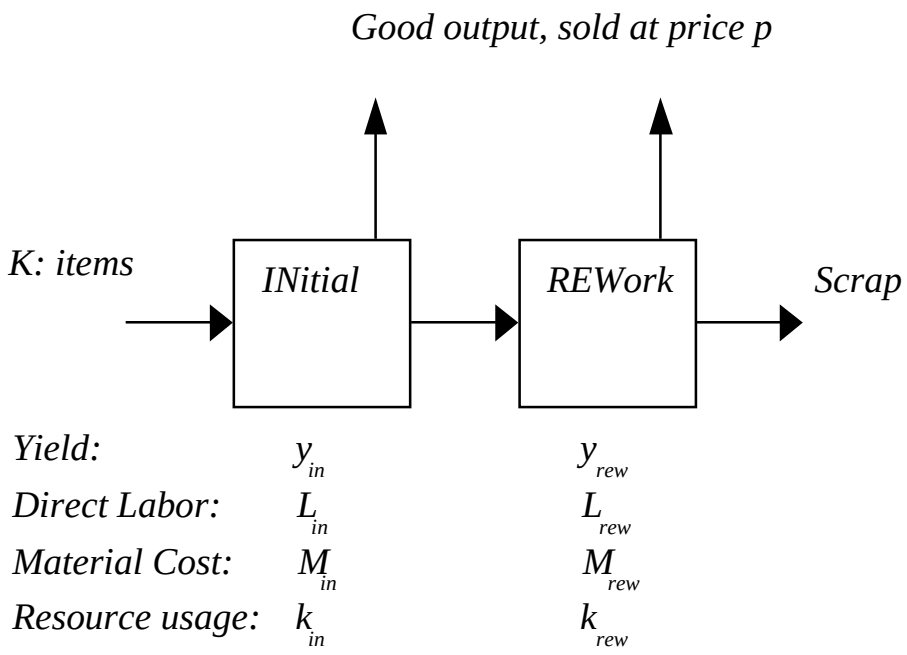

Figure 3: Production process with rework

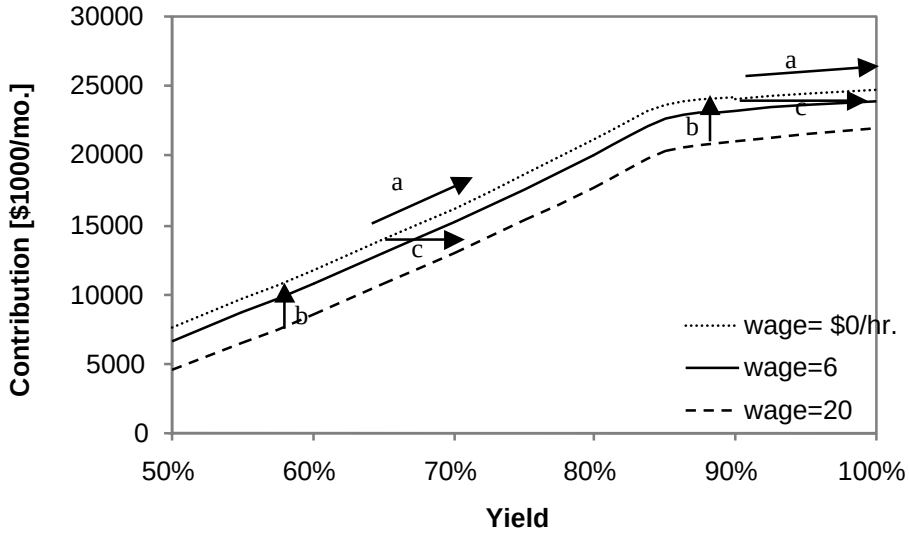

Figure 4: Impact of yield changes (both types) and wages on contribution