Insights from Modeling the Dynamics of Process Improvement

J. Bradley Morrison
Brandeis University
International Business School
bmorriso@brandeis.edu

July 2007

Insights from Modeling the Dynamics of Process Improvement

ABSTRACT

Process improvement takes place in the context of ongoing work, so people usually face the dual pressure to produce output and to build capability. Repenning and Sterman's (2002) study of two process improvement initiatives developed a causal loop diagram characterizing first-order improvements which boost output from existing processes and second-order improvements which enhance the capability of underlying processes, but the study stopped short of building a simulating model. This paper starts from the feedback structure they present and constructs a system dynamics model that formalizes the critical interaction between first- and second-order improvements as options for governing production. Analytical results characterize the optimal tradeoff between working harder and working smarter. However, practitioners generally must manage this tradeoff lacking adequate knowledge of the parameter space to find the optimum. Results demonstrate tipping points in the dynamics of process improvement, identify perverse behaviors that are likely to thwart the good intentions of practitioners, and show how the feedback structure of process improvement presents challenges to agents facing the dual pressures to produce and improve. By moving from causal loops to a simulating model, the paper also provides an example of how formal modeling yields more nuanced understanding.

Insights from Modeling the Dynamics of Process Improvement

The complex and problematic nature of process improvement has attracted the attention of both scholars and practitioners. Organizations often strive to improve the execution of their core tasks and processes, employing a variety of intentional approaches to enhancing organizational performance that have at their core a notion of process improvement. Examples include business process reengineering, total quality management, six sigma, the Toyota Production System, and many techniques based on ideas of lean manufacturing (Monden 1983; Womack, Jones et al. 1990; Hammer and Champy 1993; Cole and Scott 2000; Rigby 2001). The essential challenge to both practitioners and scholars is that the track record of process improvement initiatives is an inconsistent one. On the one hand, there is ample evidence that these initiatives are sometimes successful in yielding improvements in organizational performance. But, on the other hand, many efforts fail to yield the desired benefit, often exhibiting a pattern of short-lived improvement followed by a decline in performance to levels at or below those before the improvement initiative began. These initiatives are examples of implementation failure (Klein and Sorra 1996). The reasons that many organizations face difficulties in implementing what they know to be good ideas remain at best poorly understood.

An emerging stream of literature examining the phenomenon of problematic process improvement has explicitly considered feedback explanations. In this literature, one class of explanations points to factors in the broader organizational context that undermine the sustainability of the improvement activity. Sterman, Repenning and Kofman (1997) (Sterman, Repenning et al. 1997) highlight the impending fear of losing jobs as improvements yielding greater productivity imply a need for fewer employees. Keating and Oliva (2000) (Keating and Oliva 2000) point to the challenges of simultaneously undertaking multiple improvement projects. Repenning (2002) (Repenning 2002) shows the dynamic effects of waning employee commitment to process improvement. A second class of feedback explanations regarding problematic process development takes a more micro view and identifies critical interactions in the work of process improvement itself. Repenning and Sterman (2002) develop a causal loop model of the dynamics of process improvement that distinguishes first-order improvements (working harder) and second-order improvements (working smarter). The explanation for problematic behavior is rooted in understanding the links between these two options for responding to pressures to improve performance.

The purpose of this paper is to examine the dynamics of process improvement. Specifically, the paper constructs a dynamic mathematical model building on the feedback structure presented in Repenning and Sterman's (2002) study. The model formalizes the critical interaction between first- and second-order improvements as options for governing production. Moving from Repenning and Sterman's causal loop model to a fully formulated system dynamics model enables a more rigorous examination into how the feedback structure of process improvement presents challenges to people in a system facing the dual pressure to produce output and to build capability.

The paper is organized as follows. The next section presents a brief summary of Repenning and Sterman's causal model to provide a starting point for the modeling in this paper. The next section describes the model, describing the formulations and discussing some modeling choices made in the model formulation step. The next section presents analytical results of equilibrium analysis, identifying optimal sustainable performance. The next section uses simulation analysis to explore several polices for managing process improvement. Finally, the concluding section discusses the findings and some implications for theory and practice.

A MODEL OF PROCESS IMPROVEMENT

Repenning and Sterman (2002) studied two process improvement initiatives aimed at reducing cycle times in the electronics components division of a major U.S. automaker. Motivated by the observation that one initiative succeeded while the other languished even though both were launched and managed by the same manager, they developed a causal loop model to explain the evolution of the two initiatives. The model is grounded in their field data and consistent with the principles of operations and quality management, organizational theory, and the literature on human decision making. Figure 1 is a replication of their causal loop model.

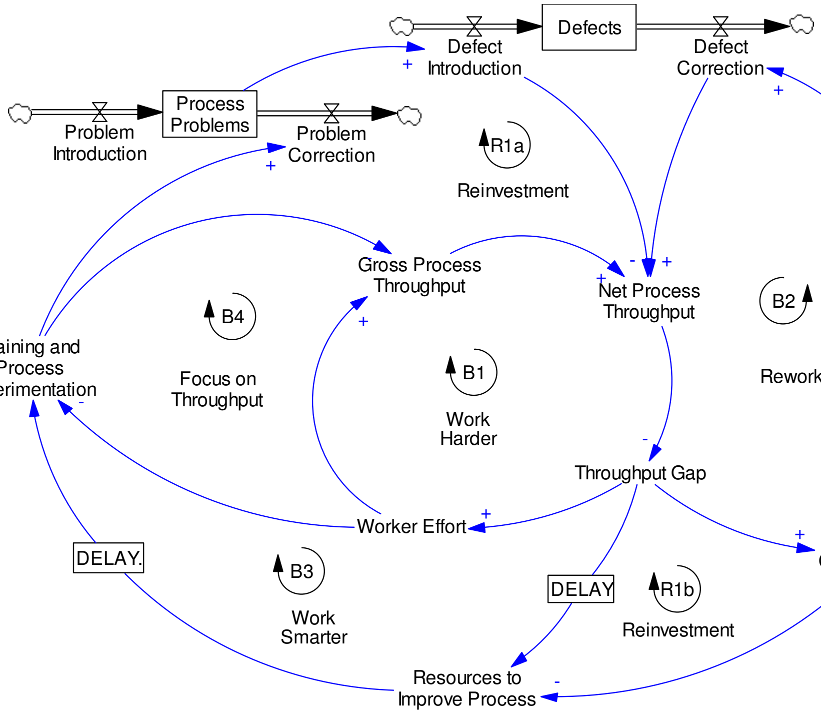


Figure 1: A Model of Process Improvement, taken from Repenning and Sterman (2002)

A central construct in their model is Net Process Throughput, the basic metric for organizational performance. Net Process Throughput results from the combination of Gross Process Throughput (the rate at which new work gets done), defect introduction (the portion of new work that is defective), and defect correction (the rate at which previously defective work is made usable, for example by fixing errors). The model describes several options available to managers and workers to regulate output in order to meet the exogenous goal for Net Process Throughput. Each option is a balancing feedback loop that seeks to eliminate the Throughput Gap. Repenning and Sterman divide these options into first-order improvement and second-order improvement.

First-order improvement achieves greater usable output from the existing process. Figure 1 includes two examples of first-order improvement. One is shown as balancing loop B1, the Work Harder Loop. The response to throughput pressure is to increase worker effort, thus boosting gross throughput. The second example of first-order improvement is balancing loop B2, the Rework Loop. In this loop, the response to throughput pressure is to allocate more resources to correcting defects in previous work done incorrectly, thus increasing the rate of defect correction as a contribution to Net Process Throughput.

Second-order improvement achieves higher rates of output by enhancing the capability of the underlying process. Figure 1 shows second-order improvement in balancing loop B3, the Work Smarter Loop. In response to throughput pressure, managers and workers allocate more time to process improvement activities that lead to problem correction. The result is a decrease in the stock of process problems, so rate of defect introduction declines, and net process throughput is increased as a greater proportion of new work is done correctly. Second-order improvements bolster the process capability and thus contribute a more enduring benefit compared to first-order improvements, which contribute to net throughput only at a recurring cost. First-order improvements are analogous to expenses, such a direct labor, whereas second-order improvements are analogous to investments, such as the purchase of more efficient equipment.

Repenning and Sterman use their causal loop model to explain the evolution of the two initiatives and then develop some important insights regarding how the difficulties of process improvement “are rooted in the ongoing interactions among the physical, economic, social, and psychological structures in which implementation takes place” (Repenning and Sterman 2002 p. 275). They note that four individual level psychological biases serve to favor first-order improvements over second-order improvements. Compared to those of second-order improvements, the outcomes of first-order improvements are more salient, more certain, and achieved with less delay. Moreover, because a stock of defective products eligible for repair presents an opportunity to recoup past expenditures, the sunk cost fallacy contributes to an even greater bias towards first-order improvement in the form of correction. These biases set in motion the reinforcing loops in Figure 1 that can work as vicious cycles and lead to a capability trap. First-order improvements generate some immediate gains, but eventually process capability erodes, increasing throughput pressure, shifting time away from process improvement, and leading to even further reliance on first-order improvement. The reinforcing loops trap the organization in a state of low process capability. Since managers are subject to the fundamental attribution error, they are likely to attribute performance problems to problems with the workforce. Based on this attribution, efforts to get the workforce to work harder (first-order improvement) are favored. As workers respond to pressure to work harder, shifting time away from process improvement to direct production activity, the immediate consequence will be an improvement in throughput. The perverse short-term effectiveness of first-order improvement means that managers observe improvement, providing evidence to support their initial incorrect attribution that inadequate worker effort is the cause of low throughput. The self-confirming attribution error can become institutionalized and leave the organization paralyzed in a state of conflict and mistrust, incapable of any useful change.

FORMULATING THE MATHEMATICAL MODEL

This section presents the system dynamics model developed from the casual loop diagram in Figure 1. The modeling work begins with a choice of model boundary that excludes some of the causal loop structure presented by Repenning and Sterman. Their causal loop structure represents two types of first-order improvement as seen in balancing loops B1, the Work Harder Loop, and B2, the Rework Loop in Figure 1. In the interest of parsimony, the model developed here considers only one method of first order improvement – the Work Harder Loop. This simpler model retains the ability to examine the key interactions between first- and second-order improvement, and an extension to include another method of second-order improvement such as defect correction would be straightforward. Figure 2 shows a stock and flow diagram of the model developed here. The remainder of this section describes the model formulations.

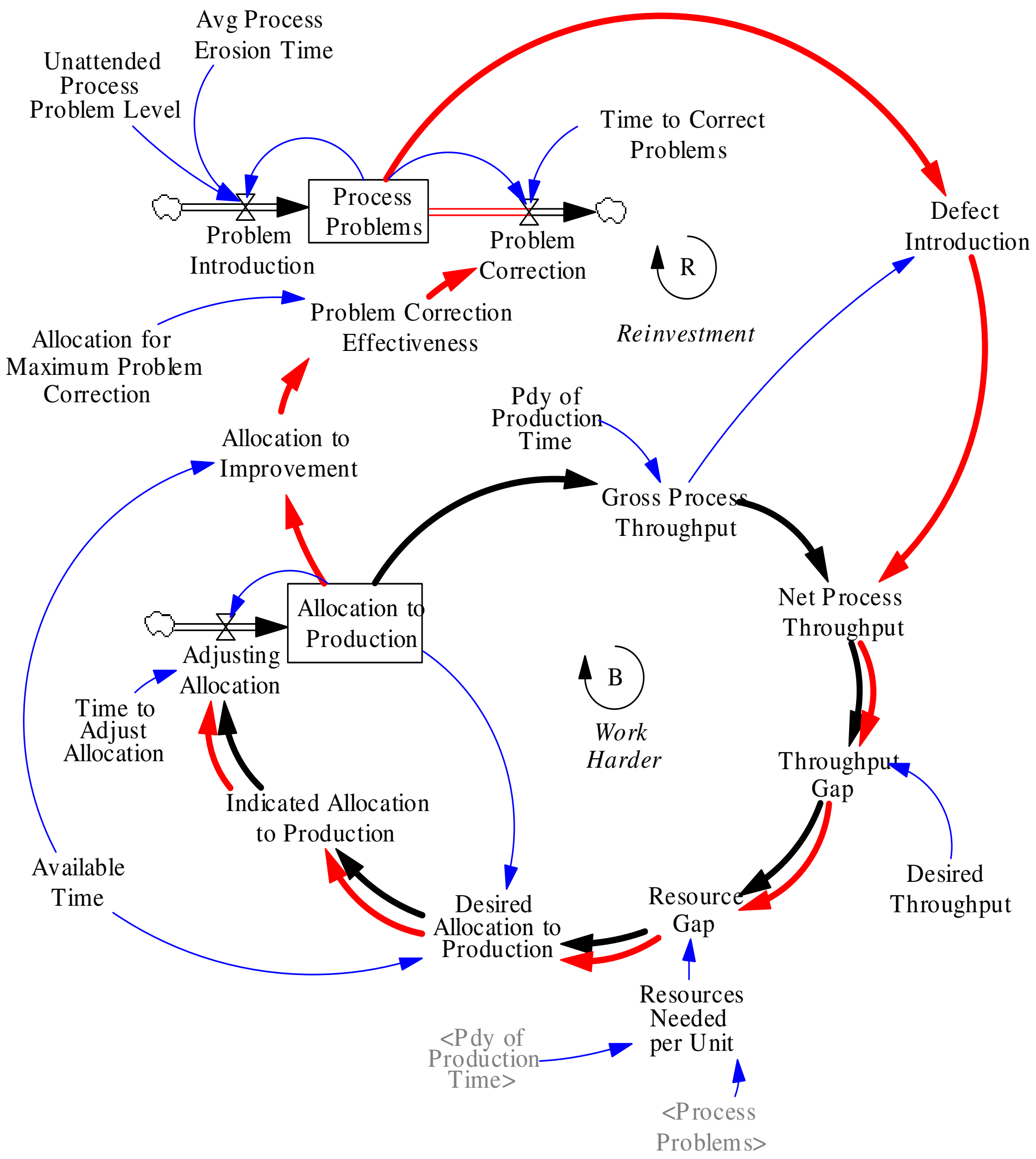


Figure 2: Modeling the Interaction of First- and Second-order Improvement

The measure of organizational performance for the stylized production organization in the model is Net Process Throughput. Net Process Throughput is the rate of Gross Process Throughput less the rate of Defect Introduction. Gross Process Throughput is the product of the amount of time workers spend on production activities, the Allocation to Production, times the Productivity of Production Time.

Net Process Throughput = Gross Process Throughput - Defect Introduction

Gross Process Throughput
= Allocation to Production * Productivity of Production Time

The Allocation to Production is a stock that is increased or decreased by Adjusting Allocation. The Adjusting Allocation flow is some fraction of the gap between the Indicated Allocation to Production and the Allocation to Production. The fraction is given by 1/Time to Adjust Allocation.

Allocation to Production
= INTEGRAL(Adjusting Allocation, Indicated Allocation to Production)

Adjusting Allocation
= (Indicated Allocation to Production – Allocation to Production)/Time to Adjust Allocation

The key policy rule represented in the feedback structure of this model is the allocation of the workers' time among two activities: production, which is a type of first-order improvement, and problem correction, which is a type of second-order improvement. The model assumes that all of the workers' time is allocated to these two activities. To model the Work Harder Loop as described by Repenning and Sterman, the workers are assumed to respond to throughput pressure created by a Throughput Gap equivalent to the shortfall of Net Process Throughput relative to Desired Throughput. From the standpoint of these workers, Desired Throughput is an exogenous goal. The model also assumes, contrary to fact, that the allocation decision is made with full knowledge of the state of the system, including instantaneous and completely accurate knowledge of the throughput rate, the defect introduction rate, the productivity of production time, and the current allocation to production. The reason for this assumption is to eliminate any flaws in perception, information processing, or allocation decision making as possible causes of the pathologies that will be observed in model behavior. There are no “mistakes” in decision making, although the policies that govern the ongoing allocation decisions may be flawed.

The Indicated Allocation to Production is the current Allocation to Production adjusted to respond to the Resource Gap. The Indicated Allocation to Production is constrained to be nonnegative and to not exceed the Available Time. The Resource Gap is determined by the Throughput Gap and the Resources Needed per Unit. The Throughput Gap is the difference between the Desired Throughput and the Net Process Throughput. The Resources Needed per Unit depends on the Productivity of Production Time and the fraction of Process Problems that generate defects.

Indicated Allocation to Production
= max[0, min(Available Time, Allocation to Production + Resource Gap)]

Resource Gap = Throughput Gap*Resources Needed per Unit

Throughput Gap = Desired Throughput – Net Process Throughput

Resources Needed per Unit = Productivity of Production Time/(1 - Process Problems)

Defect Introduction arises as some of the production output accomplished according to the Gross Process Throughput rate is done incorrectly. The fraction of the Gross Throughput that is done incorrectly depends on the process capability as defined by Process Problems. Process Problems is a stock that is increased by Problem Introduction and decreased by Problem Correction. Process Problems are measured as a dimensionless index ranging from 0 to 1, so the variable is a direct indicator of the fraction of Gross Process Throughput that is done incorrectly.

Defect Introduction = Gross Process Throughput * Process Problems

Process Problems

= INTEGRAL(Problem Introduction – Problem Correction, Initial Process Problems)

Problem Introduction is an inflow to the stock of Process Problems. Problem Introduction is modeled as a process of natural entropy. If the process is left unattended by any improvement activity, over time the process will deteriorate to a state of high process problems as given by the Unattended Process Problem Level. The Problem Introduction flow closes a fraction of the gap between the current Process Problem level and the Unattended Process Problem Level at rate given by the Average Process Erosion.

Problem Introduction

= (Unattended Process Problem Level – Process Problems)/Average Process Erosion Time

Problem Correction takes place when workers spend time conducting improvement activities such as investigating problems, conducting experiments, and implementing process changes. Empirical analyses of rates of process improvement over time show that they exhibit characteristic half-lives, depending on such factors as the technical and organizational complexity (Schneiderman 1988). Absolute rates of improvement are relatively high when processes are in states of low capability, but these absolute improvement rates decline as the process capability increases. The formulation used here thus models the potential improvement rate as a constant fractional decrease in process problems. The potential improvement rate is adjusted to account for the Problem Correction Effectiveness, which is a function of how much time workers spend on problem correction (Allocation to Problem Correction), relative to the Allocation for Maximum Problem Correction. The Allocation to Problem Correction is the Available Time less the Allocation to Production.

Problem Correction =
Problem Correction Effectiveness* (Process Problems/Time to Correct Problems)

Problem Correction Effectiveness
= Allocation to Problem Correction/Allocation for Maximum Problem Correction

Allocation to Problem Correction = Available Time – Allocation to Production

Notice that the Allocation to Problem Correction is the amount of time "left over" after the desired Allocation to Production is made. The decision rule implied here is that the production activities take a higher priority than the improvement activities, consistent with the field study data in Repenning and Sterman. For example, a respondent describing a pilot improvement project said, "People had to do their normal work *(production activity)* as well as keep track of the work plan *(improvement activity).* There just weren't enough hours in the day, and the work *(production activity)* wasn't going to wait." (Repenning and Sterman 2002 p 273. Comments in italics added.) The strict priority of first-order improvement also implies that second-order improvement takes place not as a direct response to a Throughput Gap but as an investment when resources are available. The model in Figure 2 does not explicitly represent the Work Smarter Loop, loop B3 from Figure 1, following this assumption.

OPTIMALITY ANALYSIS

This section presents analytical results. The analytical strategy is to identify and characterize the conditions for long-term optimal throughput for a given quantity of resources. While there may be opportunities for temporary increases above the long-term optimal level, the analysis in this section will solve for the optimal allocation of workers' time between production and problem correction in order to achieve the maximum Net Process Throughput in steady state. Consider first the intuition to suggest such a maximum exists. For a very low allocation to production, the organization will be using resources for improvement that could be more productively employed in production activity that would boost output by increasing Gross Process Throughput. Process Problems would be at a relatively low level, so additional time spent producing would yield much usable output. That is, the marginal benefit of an additional hour of production exceeds the marginal opportunity cost. The marginal opportunity cost is a consequence of the increase in Process Problems that would result from allocating less time to problem correction. Conversely, for a very high allocation to production, the organization will be using resources for direct production that could be more productively employed in Problem Correction activities. Process Problems will be at a relatively high level, so additional time spent correcting problems would boost Net Process Throughput by improving the proportion of usable work (i.e., reducing the rate of Defect Introduction). That is, the marginal benefit of an additional hour of problem correction exceeds the marginal opportunity cost. Thus, somewhere between these two extremes there lies at least one local peak at which Net Process Throughput will be (locally) maximal.

There are three conditions for the system to be in steady-state equilibrium. First, the stock of Process Problems must be in equilibrium, implying that the inflow Problem Introduction must equal the outflow Problem Correction. Second, the stock Allocation to Production must be in equilibrium, implying that Adjusting Allocation must equal zero, which occurs when the Allocation to Production is at its desired level, the Indicated Allocation to Production. Third, the Work Harder balancing loop must be in equilibrium, implying that the Throughput Gap is zero which occurs when the Net Process Throughput is equal to Desired Process Throughput. It can be seen by inspection, that setting the Desired Process Throughput to the optimal level and the Allocation to Production to the level required to achieve this optimal throughput will satisfy the second on third conditions. Thus, the analysis turns to the first condition to find the optimal allocation.

The optimal solution is found starting from two equations:

(1) Net Process Throughput =

Allocation to Production * Productivity of Production Time * (1- Process Problems)

(2) Problem Introduction = Problem Correction.

Note that both Problem Introduction and Problem Correction are functions of Process Problems. The optimal Net Process Throughput is found by first substituting into equation (2) and rearranging to give a univariate expression for Process Problems as a function of model parameters and the variable Allocation to Production. Substituting this expression into equation (1) yields an equation for Net Process Throughput as a function of Process Problems. Differentiation of the resulting equation with respect to the Allocation to Production and setting the derivative equal to zero gives a quadratic equation for the first order conditions for optimality. Solving with the quadratic formula finds the two roots of the equation, which can be translated into values of Allocation to Production. One of the roots is infeasible, as it implies an allocation to production in excess of the Available Time. The other root is the optimal allocation, and the optimal throughput can be found from this root and the expression for Net Process Throughput derived above. The full text of these calculations is available from the authors on request. Figure 3 plots the equilibrium value of Net Process Throughput as a function of the Allocation to Production for the set of parameters shown in Table 1.

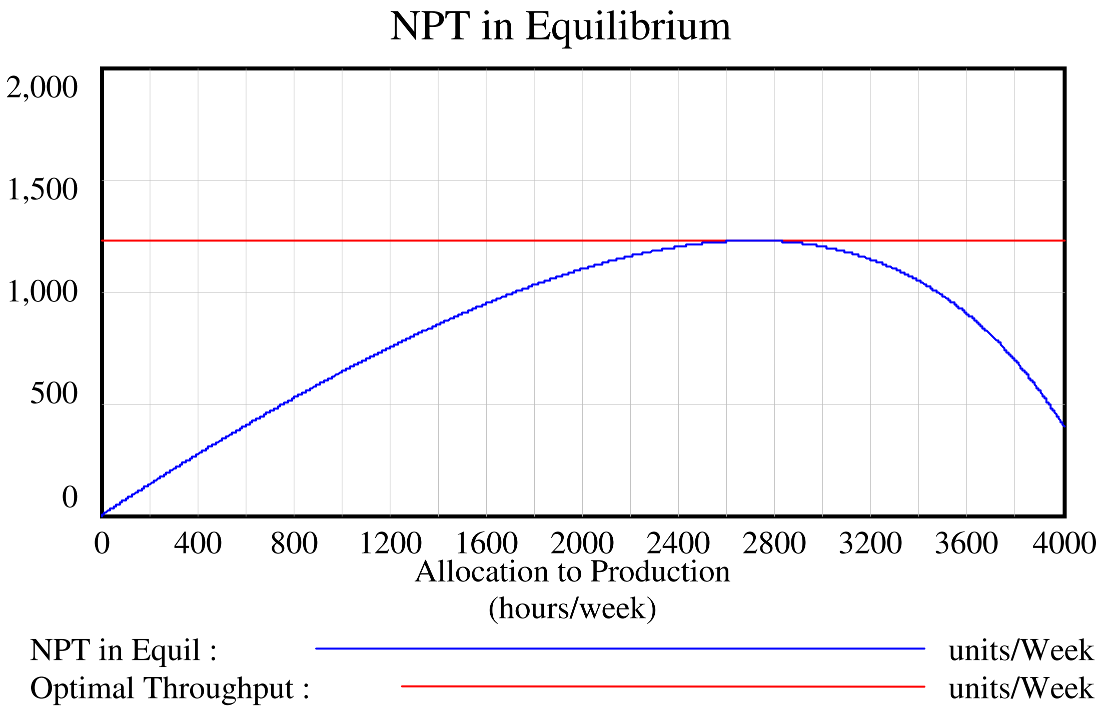


Figure 3: Relationship of Steady State Equilibrium Rates of Net Process Throughput to Allocation to Production

Table 1: Baseline Parameter Values for Simulation Analysis

Parameter	Value	Units
Unattended Process Problem Level	0.9	Dimensionless
Avg Process Erosion Time	36	Weeks
Time to Correct Problems	16	Weeks
Productivity of Production Time	1	Unit/hour
Allocation for Maximum Problem Correction	4000	Hours/week
Available Time	4000	Hours/week
Time to Adjust Allocation	1	Week
Initial Process Problems	0.4	dimensionless

SIMULATING THE DYNAMICS OF PROCESS IMPROVEMENT

This section presents the results of simulation analysis to investigate the dynamic behavior of the stylized production system under various improvement scenarios. The baseline simulations use the parameter settings shown in Table 1. The Desired Throughput and the Initial Allocation to Production are set so the simulations begin in equilibrium conditions. The first tests are to establish basic behavior patterns of the system. The simulation in Figure 4 introduces a pulse increase in the Desired Throughput in week 10. The pulse input causes an increase in the Throughput Gap that stimulates greater Allocation to Production, resulting in more Gross Process Throughput and therefore more Net Process Throughput. Because the increase in Desired Throughput is only temporary, following the initial pulse, the Net Process Throughput smoothly returns to its original equilibrium rate as the Allocation to Production adjusts to its initial level. This test demonstrates that, in response to moderate disturbances, the allocation rule achieves its intended goal, bringing the Net Throughput to equal Desired Throughput.

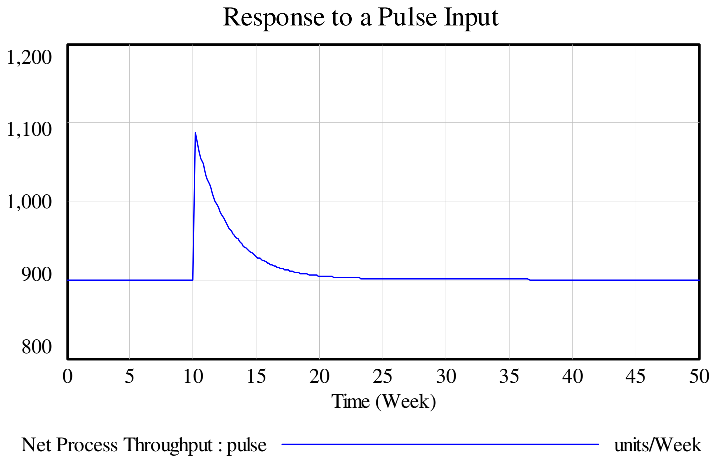


Figure 4: Response to a Pulse Input

Next, consider how to achieve higher levels of performance n this system. The “lever” available for a manager of this system is to set the Desired Process Throughput. As Figure 4 shows, the system will respond, at least under some conditions, by adjusting the Allocation to Production in order to achieve the target throughput. Figure 5 shows the result of introducing a one-time permanent step increase in the Desired Throughput in week 10. The Allocation to Production increases in response to higher Desired Throughput, so Net Process Throughput increases and remains at a higher level. The graph on the right of Figure 5 displays the behavior of the stock of Process Problems. With a higher Allocation to Production, there is a lower Allocation to Problem Correction, so stock of Process Problems grows and then stabilizes at a higher level as the throughput target is achieved. This simulation shows an example of conditions under which first-order improvement is effective. The organization is able to achieve a higher level of performance, although the capability of the process is compromised, as seen by the increase in Process Problems. Because Process Problems were low to begin, there was sufficient organizational slack to absorb the stress from the increased target output.

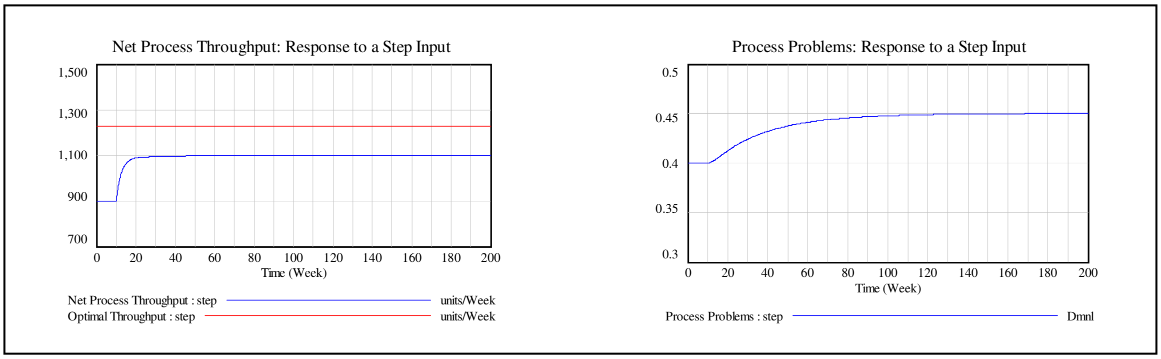


Figure 5: Response to a Step Input

It is instructive to examine the scenario in Figure 5 a bit more closely, so Figure 6 plots four variables in the simulation on the same set of axes. The red line shows the Desired Throughput, which increases in a step fashion at time 10 weeks. The blue line shows the response of Net Process Throughput. The workers adjust to the new goal and successfully achieve the higher level of throughput. The blue line smoothly approaches the red line. Together, the red and blue lines can be interpreted as the manager's view of what has happened. All appears well, the story seems to be over by about week 24, and the manager might even consider increasing the goal yet again once the new target has been reached. But let us look a bit more carefully at what else is changing. The green line shows the workers Allocation to Production. Note that they allocate more time to production and continue to do so even after the manager believes the change has been completed. Moreover, as the grey line shows, the consequences of reducing the time spent on problem correction (because more time is spent on production) manifest as a slow increase in process problems. The workers must thus continue to allocate more time to production since the defect rate is increasing, and they are caught in a treadmill created by the reinforcing Reinvestment loop working as vicious cycle. Long after what the manager would observe as process throughput reaching its goal, the workers experience the ongoing deterioration of the process capability and are forced to work more and more on first-order production activity. The dysfunctional attributions that are central to the story in Repenning and Sterman are easy to imagine based on the results of this simulation.

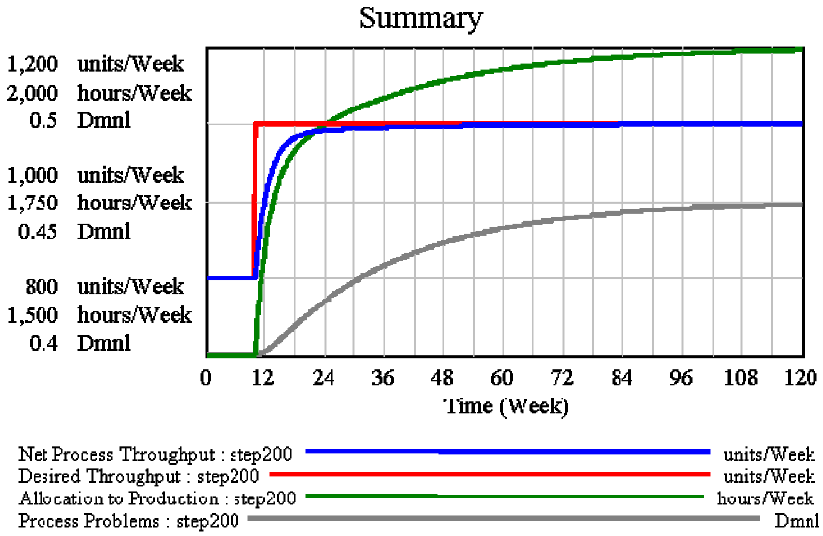


Figure 6: Summary of Dynamics of Response to a Step Input

The graph on the left of Figure 5 also includes a red line showing the optimal steady-state Net Process Throughput. Note that the step increase in Desired Throughput stimulates an increase that is sustained but that leaves the organization performing below the optimal level. The expression for the optimal Allocation to Production derived in the previous section depends on the following parameters: Unattended Process Problem Level, Average Process Erosion Time, Time to Correct Problems, Allocation for Maximum Problem Correction, and Available Time. Of these, only the Available Time is directly observable, so in practice managers are highly unlikely to know the parameters necessary to choose their Desired Throughput targets in accordance with the optimality conditions. Instead, they must discover other policies to manage this system. The next simulations explore the response to other attempts to achieve higher performance.

Consider next the effect of setting Desired Throughput even higher. Figure 7 shows the results of a larger step increase, one that brings the Desired Throughput above the optimal level. The system response is to increase the Allocation to Production. The immediate effect is to increase the Net Process Throughput, as more time doing productive activities boosts Gross Process Throughput. However, the increase in Allocation to Production is accomplished at the expense of a decrease in the Allocation to Problem Correction. With less effective problem correction, capability begins to deteriorate as new problems creep in. The stock of Process Problems grows, increasing the rate of Defect Introduction and reducing the fraction of output that is usable. An even greater Allocation to Production is needed to boost Gross Throughput to the even higher levels needed to achieve the target net throughput in a state of higher Process Problems. The Reinvestment Loop R1 works as a vicious cycle, and the organization gets locked into a downward spiral. Eventually, Available Time is allocated entirely to production, and Process Problems increase until they reach their natural limit. The system reaches a steady state characterized by low levels of capability and performance, despite the full allocation of available time to productive activity. The left graph in Figure 7 includes a green line that shows this equilibrium level as the "No Maintenance Throughput," the rate achieved by setting Process Problems to the maximum given by Unattended Process Problem Level and setting Allocation to Production to its maximum given by Available Time.

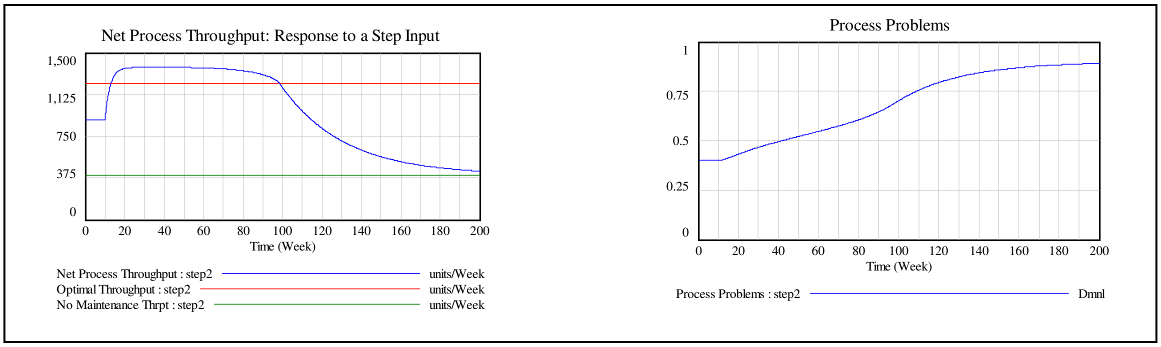
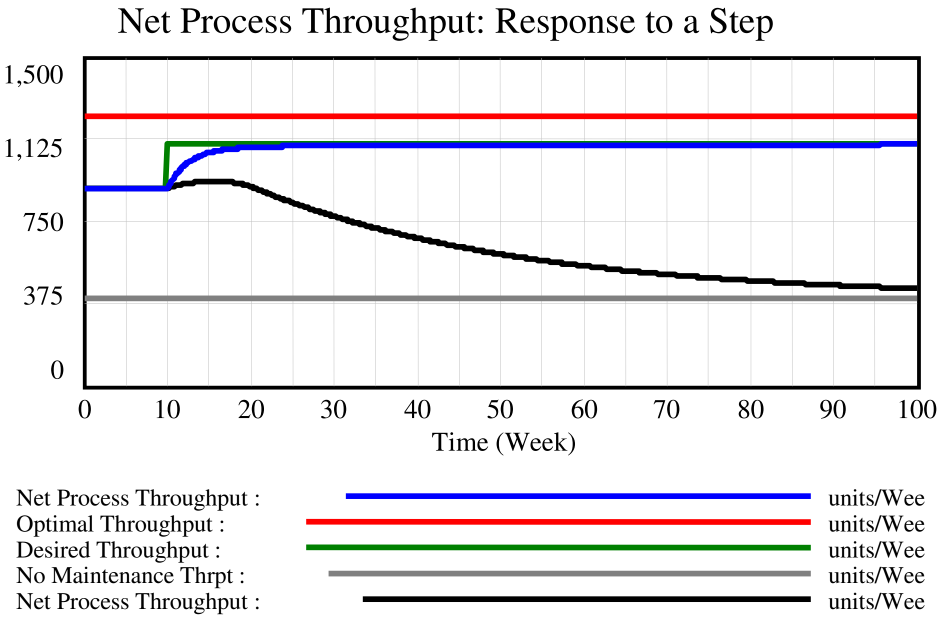


Figure 7: Response to a Step Increase above the Optimal Steady-State Net Process Throughput

Taken together, the simulations shown in Figures 5 and 7 demonstrate that the system has a tipping point. Attempts to increase performance by increasing the Desired Throughput by moderate amounts can result in sustained improvement as shown in Figure 5. However, if the targeted increase crosses a critical threshold, the mode of behavior changes to one that displays a better before worse pattern. Throughput improves at first, but only at the expense of deteriorating capability that sends the system towards a steady state performance at a level worse than where it began, even though resource levels are the same. Tipping points have been described in a variety of dynamical systems, such as fads, disease epidemics, product development processes, and disasters (Gladwell 2000; Sterman 2000; Repenning, Goncalves et al. 2001; Rudolph and Repenning 2002). Tipping points are unstable equilibrium points such that a perturbation in one direction will send the system towards one steady-state behavior while a perturbation in the opposite direction will send the system towards a meaningfully different steady-state behavior. The tipping point in the process improvement system here is the same as the point of optimal allocation for maximum net process throughput. Allocating more than this amount of time to production causes a temporary increase in net process throughput but will result in long-term deterioration towards the lowest possible performance.

The results so far have demonstrated that the feedback structure of process improvement, characterizing the critical interaction between first- and second-order improvement, defines performance (Net Process Throughput) as a curvilinear function of the key decision variable (Allocation to Production). The curvilinearity implies that the system has a tipping point. Moreover, since the location of the tipping point is unknowable in practice, stretching throughput goals risks pushing the system beyond sustainable levels and locking in to a cycle of deterioration. Let us return to Figure 3 to demonstrate another important implication of the curvilinear production function. Imagine a horizontal line crossing the production function somewhere below the optimum. There are in general two values of the Allocation to Production that will achieve any given rate of Net Process Throughput. The value to the left of the optimum characterizes an equilibrium in which there is an opportunity to improve output by shifting time away from second-order improvement to make more time for first-order production activities. The stock of process problems is relatively low, indeed lower than optimal, so the system can accommodate increased throughput pressure by allowing some deterioration of this stock. The lower than optimal stock of process problems (i.e., higher than optimal stock of organizational capability) is a form of organizational slack. Under these conditions, an increase in the throughput goal leads to improved performance, as we saw in the simulation in Figures 5 and 6, and which is shown again as the blue line in Figure 8. But what if the system begins at the Allocation to Production to the right of the optimum? The black line in Figure 8 shows the results of another simulation that begins with the organization achieving the same Net Process Throughput and is then subjected to the same increase in Desired Throughput as the blue line, but in this test the starting Allocation to Production corresponds to the value to the right of the optimum. This point characterizes a system in which workers have diligently overcome a somewhat weaker process by working harder (with a higher Allocation to Production) in order to achieve the desired throughput. The stock of Process Problems is relatively high, in particular higher than optimal. Increasing the throughput goal in this case sets in motion the vicious cycle of deteriorating capability and pushes the system to low levels of output. The same change has resulted in a dramatically different outcome.



DISCUSSION

The paper developed a system dynamics model starting from the causal loop model Repenning and Sterman (2002) present based on their field study of two improvement initiatives. By moving beyond a causal loop diagram to a simulating system dynamics model, this paper is able to more rigorously examine the dynamics of process improvement. First- and second-order improvement options are effective in boosting short-term organizational performance, but only second-order improvement builds sustainable capability. Analytical results characterize the optimal choice for the allocation of the constrained resource of worker time among these two types of activities. While the optimum can be found analytically in the stylized system, where all parameter values are known, managers in real systems are unlikely to have the information to do so in practice. Consequently, policies for managing must be discovered by experimentation.

Simulation analysis shows that increases in the goal for performance can lead to improved performance. For moderate increases in the goal, and with some organizational slack available, the organization can achieve an enduring improvement. However, for increases in the goal that are more severe, the organization's attempt to meet the goal results in a temporary improvement followed by decline that sends the organization towards a steady state of performance inferior to the starting point. The system has a tipping point. Below the tipping point, stretching the goal to reach higher levels of output is a successful strategy. But stretching to goal above the tipping point triggers a vicious cycle of over-reliance on first-order improvement that sends the organization into a downward spiral of performance. The tipping point in this system is at the optimal level for process throughput. An especially perverse characteristic of the behavior in these simulations is the similarity in the short-term responses to changes that are either below or above the tipping point. The response to an “appropriate” increase in a goal to a level in the safe region below the tipping is an improvement in throughput. Likewise, the response to an “inappropriate” increase in a goal to a level in the dangerous region above the tipping point is an improvement in throughput. The delayed effects of deteriorating process capability reverse the initial increase in throughput, but a manager focused on throughput may not observe the difference until it is too late to recover from the underinvestment in building or maintaining organizational capability. Managers need to find more robust strategies for decision making in such systems. Future work will explore other types of goal setting and monitoring strategies.

REFERENCES

Cole, R. E. and W. R. Scott (2000). The Quality Movement & Organization Theory. Thousand Oaks, CA, Sage Publications.

Gladwell, M. (2000). The Tipping Point: How Little Things Can Make a Big Difference. Boston, Little, Brown and Company.

Hammer, M. and J. Champy (1993). Reengineering the Corporation: A Manifesto for Business Revolution. New York, Harper Collins.

Keating, E. and R. Oliva (2000). "A Dynamic Theory for Sustaining Process Improvement Teams in Product Development." Advances in Interdisciplinary Studies of Work Teams **5**: 245-281.

Klein, K. J. and J. S. Sorra (1996). "The Challenge of Innovation Implementation." Acadmemy of Management Journal **21**(4): 1055-1080.

Monden, Y. (1983). Toyota Production System. Atlanta, GA, Institute of Industrial Engineers.

Repenning, N. P. (2002). "A Simulation-based Approach to Understanding the Dynamics of Innovation Implementation." Organization Science **13**(2): 109-127.

Repenning, N. P., P. Goncalves, et al. (2001). "Past the Tipping Point: The Persistence of Firefighting in Product Development." Califronia Management Review **43**(4): 44-63.

Repenning, N. P. and J. D. Sterman (2002). "Capability Traps and Self-Confirming Attribution Errors in the Dynamics of Process Improvement." Administrative Science Quarterly **47**: 265-295.

Rigby, D. (2001). "Management Tools and Techniques: A Survey." California Management Review **43**(2): 139-160.

Rudolph, J. W. and N. R. Repenning (2002). "Disaster Dynamics: Understanding the Role of Stress and Interruptions in Organizational Collapse." Administrative Science Quarterly **47**: 1-30.

Schneiderman, A. (1988). "Seting Quality Goals." Quality Progress(April): 55-57.

Sterman, J. D. (2000). Business Dynamics: Systems Thinking and Modeling for a Complex World. Chicago, Irwin-McGraw Hill.

Sterman, J. D., N. P. Repenning, et al. (1997). "Unanticipated Side Effects of Successful Quality Programs: Exploring a Paradox of Organizational Improvement." Management Science **43**(4): 503-521.

Womack, J. P., D. T. Jones, et al. (1990). The Machine that Changed the World. New York, Harper Collins.