Process Improvement Dynamics Under Constrained Resources:
Managing the Work Harder versus Work Smarter Balance

J. Bradley Morrison
Brandeis International Business School
Mailstop 032
P.O. Box 549110
Waltham, MA 02454
USA
Tel: 1 (781) 736-2246
Fax: 1 (781) 736-2269
bmorriso@brandeis.edu

Brad Morrison is Assistant Professor of Management at the Brandeis International Business School. His research focuses on the paradoxes of implementation and building organizational capability. He holds a Ph.D. in Management (System Dynamics and Organization Studies) from the Sloan School of Management at MIT.
Process Improvement Dynamics Under Constrained Resources:
Managing the Work Harder versus Work Smarter Balance

ABSTRACT

This paper considers the problem of managing process improvement when resources are constrained. The paper develops a system dynamics model that formalizes the critical interaction between first- and second-order improvements as options for governing production. Both first-and second-order improvement are effective in boosting short-term organizational performance, whereas only second-order improvement builds sustainable capability. Analytical results characterize the optimal tradeoff between these options (working harder and working smarter). Simulation results show how pressuring the workforce to work harder can set off a vicious cycle of deteriorating process capability that squeezes out improvement activity. Results also highlight the key role of process capability as a form of organizational slack and demonstrate tipping points in the dynamics of process improvement. By moving from causal loops to a simulating model, this paper provides an example of how formal quantitative modeling can lead to important insights not discovered with causal loop diagramming alone.

To prosper in the face of competitive markets and increasingly demanding customers, organizations must strive not only to produce goods or services to meet the market's requirements but also to improve the performance of their fundamental business processes. Whether they originate in internal aspirations or external market forces, pressures to improve feature prominently in classic theories of organizational performance (Cyert & March, 1963; Nelson & Winter, 1982). Organizations often employ intentional approaches to enhancing organizational performance centered on the notion of process improvement, systematically seeking better ways to execute their core tasks and processes (Van de Ven & Poole, 1995). The simultaneous pursuit of a goal to deliver for the market and a goal to improve core business processes creates a fundamental resource allocation problem that requires striking an appropriate balance between these two demands for resources. This paper considers the resource allocation
problem for managers facing the dual pressures to produce and improve, examining the
dynamics of process improvement under constrained resources.

Examples of process improvement are numerous, including business process reengineering, total
quality management, statistical quality control, theory of constraints, six sigma, the Toyota
Production System, and many techniques based on ideas of lean manufacturing (Cole & Scott,
2000; Cox & Goldratt, 1986; Hammer & Champy, 1993; Monden, 1983; Rigby, 2001; Womack,
Jones, & Roos, 1990). Process improvement initiatives may be focused on reducing cost,
reducing cycle times, improving quality, building flexibility, or boosting throughput. The
widespread availability notwithstanding, there is considerable disagreement as to whether these
programs are helpful in improving organizational performance (Byrne, 1997; Easton & Jarell,
The complex and problematic nature of process improvement has attracted the attention of both
scholars and practitioners (Dean & Bowen, 1994). 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 benefits, 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.
Because these improvement approaches have been successful in some instances, we must look
beyond an explanation for failure that rests on claims of any inherent lack of efficacy of the
improvement approach itself (Klein & Sorra, 1996; Pfeffer & Sutton, 2000; Repenning, 2002).
Instead, scholars see these initiatives as examples of implementation failure (Klein & Sorra,
1996). While scholars have begun to recognize that the structure, resources, practices and
systems of implementation play an important role in influencing implementation effectiveness (Choi & Chang, 2009; Purvis, Sambamurthy, & Zmud, 2001), the reasons that many organizations face difficulties in implementing what they know to be good ideas remain at best poorly understood (Cole & Scott, 2000; Klein, Conn, & Sorra, 2001; Pfeffer & Sutton, 2000).

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. For example, the impending fear of losing jobs as improvements yield greater productivity, implying a need for fewer employees, hinders progress in some initiatives. (Sterman, Repenning, & Kofman, 1997). Simultaneously undertaking multiple improvement projects can overly tax the organization and lead to stagnation of quality improvement (Keating & Oliva, 2000). The insidious dynamic effects of waning employee commitment to process improvement, especially when managerial commitment is not apparent, can explain some instances of failure of otherwise successful programs (Repenning, 2002). 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 first- and second-order improvement as options for responding to pressures to improve performance. The motivation for the current study is the premise that research that clarifies the connections between such front-line behaviors and the macro level
outcomes they generate will shed light on the paradox of problematic process improvement. An important step in this direction is to frame the problem in a manner consistent with the challenges faced, that is, the choices that need to be made, by front-line workers and managers.

In practice, managers facing the need to improve (or perhaps even maintain) their firm’s productive capability must make decisions about the allocation of resources to these activities. This resource allocation decision has attracted a considerable amount of scholarly attention, especially in the stream of literature on quality improvement. For example, one view holds that organizations should continually aspire to achieve zero defects, implying that more improvement is always better (Crosby, 1979; Deming, 1982). A different view holds that quality improvement practices have positive returns, but only up to a point, so managers should choose quality levels based on economic trade-offs (Juran, 1979). Researchers in this stream have studied and modeled costs and tradeoffs in order to determine the optimal level of, or optimal policies for, investment in process (or quality) improvement (Carrillo & Gaimon, 2000; Fine, 1986; Li & Rajagopalan, 1998). In either case, framing the question as the choice of the level of investment implicitly assumes that resources are available to scale up to optimal levels and perhaps that resources can be reduced when excessive. This firm-level view of the question is helpful to set long-run strategies, but it does not address the fundamental challenge of day to day management on the front lines. The front-line challenge is to make do with the available resources and allocate them to achieve the best outcomes.

This paper considers the problem of managing process improvement when resources are constrained. Specifically, we consider the case in which total resources in the form of employee
time available for use in both production and improvement are held constant. Resources may be constrained for a variety of reasons, such as shortages in local labor markets, budget restrictions imposed by enterprise management, differences in realized versus forecast market demand, and even more simply the time needed to adjust overall resources through right-sizing the workforce or reassigning people to or from other areas. Although these practical constraints can often be overcome in the long-run, we believe such resource constraints accurately characterize the problem from the perspective of a mid-level supervisor or a front-line worker. Despite the widespread occurrence of this resource-constrained problem in practice, the problem has received limited attention from scholars of process improvement.

The purpose of this paper is to examine the dynamics of process improvement when resources are constrained. Specifically, the paper constructs a system dynamics model that formalizes the critical interaction between using resources to produce primary output and investing resources in process improvement as means to increase throughput. The paper builds on the conceptualization of feedback structure presented in Repenning and Sterman’s (2002) study, which developed causal loop diagrams that mapped the critical interaction between first- and second-order improvements as options for governing production. This paper extends that work by constructing and analyzing a system dynamics model, enabling a rigorous examination of 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.

This paper contributes to our understanding of the problematic nature of process improvement by formalizing the interaction between first- and second-order process improvement. The formalized model enables the characterization of a tipping point in the dynamics of process improvement, demonstrating the influence of the resource allocation decision. The simulations
and analysis highlight why finding the right balance between working harder and working smarter will be difficult under the familiar conditions of constrained resources and pressures to enhance performance. The paper also contributes to a topic of longstanding interest among system dynamics practitioners: when, or indeed whether, formal quantitative modeling is warranted (Richardson, 1996; Richardson, 1999; Sterman, 1994; Wolstenholme, 1999). This paper provides an example of how quantitative modeling, moving beyond the qualitative mapping of feedback structure, can lead to important insights not discovered with causal loop diagramming alone, even by using a small model for simulation experiments (Ghaffarzadegan, Lyneis, & Richardson, 2011; Repenning, 2003). The paper also contributes as a scholarly work that replicates and extends other published work in the field of system dynamics, contributing to a culture of accumulation, as called for by prominent scholars (Forrester, 2007; Richardson, 1996).

The paper is organized as follows. The next section is a brief summary of the qualitative model that provides a starting point for the quantitative modeling in this paper (Repenning & Sterman, 2002). The following section describes the system dynamics model, detailing 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.

MODEL CONCEPTUALIZATION
Consider a firm that aims to maximize the rate of net process throughput given a production process with a certain capability, an option to undertake process improvement to increase the
capability of the process, and a fixed quantity of labor available to allocate between two activities: producing output and conducting process improvement work. We will model net process throughput as the gross process throughput (the rate at which new work gets done) less the rate of defect introduction (the portion of new work that is defective), which is a function of the capability of the process. The problem is to choose the allocation of resources between these two activities that will achieve the optimal net throughput.

Repenning and Sterman (2002) report a qualitative study of a similar problem in 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 qualitative model to map the feedback structure of the problem. Figure 1 is a replication of their causal loop model.
The structure in Figure 1 describes several options available to managers and workers to regulate output in order to meet an exogenous goal for Net Process Throughput. Each option is a balancing feedback loop that seeks to eliminate the Throughput Gap. Repenning and Sterman (2002) divide these options into first-order improvement and second-order improvement. First-order improvement achieves greater usable output from the existing process. An example of first-order improvement in Figure 1 is shown as balancing loop B1, the Work Harder Loop. The response to throughput pressure is to increase worker effort, thus boosting gross 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 the 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 (2002) use their causal loop diagram to explain the evolution of the two initiatives, focusing on important insights about how the difficulties of process improvement, “rooted in the ongoing interactions among the physical, economic, social, and psychological
structures in which implementation takes place,“ (Repenning & Sterman, 2002 p. 275), give rise to a capability trap. Psychological biases that favor first- over second-order improvement set in motion the reinforcing loops in Figure 1 that can work as vicious cycles. 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. The authors develop the important capability trap insights by discussing how managers, subject to the fundamental attribution error, are likely to attribute performance problems to problems with the workforce and thus favor efforts to get the workforce to work harder (first-order improvement). As workers respond to the managerial pressure, shifting time away from process improvement to direct production activity, throughput increases. This perverse short-term effectiveness 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.

The causal loop diagram of Figure 1 was the basis from which Repenning and Sterman (2002) developed their key insights, but they did not develop a mathematical model. The sections that follow develop and analyze a simulating model, using the dynamic hypothesis represented in this causal loop diagram as a starting point (Randers, 1980).

FORMULATING THE MATHEMATICAL MODEL
This section presents the system dynamics model developed to explore the dynamic behavior of first- and second-order improvement. The modeling work begins by selecting the feedback structure that is essential for the dynamic hypothesis. The causal loop structure in Figure 1 represents two types of first-order improvement as seen in balancing loops B1 and B2, the Work Harder Loop and the Defect Correction Loop. 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. Where possible, the model uses the same variable names as in Figure 1, and other variables are added to complete the formulation. The remainder of this section describes the model formulations.
Figure 2: Modeling the Interaction of First- and Second-order Improvement
Loop names are taken from the corresponding loops in Figure 1. Bold causal links identify the causal paths in the two main feedback loops.

The measure of organizational performance for the stylized production organization 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 Resources to Production, times the Productivity of Production Time.

\[
\text{Net Process Throughput} = \text{Gross Process Throughput} - \text{Defect Introduction}
\]

\[
\text{Gross Process Throughput} = \text{Resources to Production} \times \text{Productivity of Production Time}
\]

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

\[
\text{Resources to Production} = \text{INTEGRAL(Adjusted Allocation, Indicated Allocation to Production)}
\]

\[
\text{Adjusting Allocation} = \frac{(\text{Indicated Allocation to Production} - \text{Resources to Production})}{\text{Time to Adjust Allocation}}
\]

The key choice represented in the feedback structure of this model is the allocation of the workers’ time among two activities: production, which is a form of first-order improvement, and problem correction, which is a form 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, 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 execution, although the policies that govern the ongoing allocation decisions may be flawed.

The Indicated Allocation to Production is the current Resources 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.

\[
\text{Indicated Allocation to Production} = \\
\max[0, \min(\text{Available Time}, \text{Resources to Production} + \text{Resource Gap})] \\
\text{Resource Gap} = \text{Throughput Gap} \times \text{Resources Needed per Unit} \\
\text{Throughput Gap} = \text{Desired Throughput} - \text{Net Process Throughput} \\
\text{Resources Needed per Unit} = \text{Productivity of Production Time} \times (1 - \text{Process Problems})
\]

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.

\[
\text{Defect Introduction} = \text{Gross Process Throughput} \times \text{Process Problems}
\]

\[
\text{Process Problems} = \text{INTEGRAL(Problem Introduction} - \text{Problem Correction, Initial Process Problems)}
\]

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, where the
fraction is given by \(1/\text{Average Process Erosion Time}\).

\[
\text{Problem Introduction} = (\text{Unattended Process Problem Level} - \text{Process Problems}) / \text{Average Process Erosion Time}
\]

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 (Resources to Improve Process), relative to the
Resources for Maximum Problem Correction. The Resources to Problem Correction is the
Available Time less the Resources to Production.
Problem Correction = Problem Correction Effectiveness* (Process Problems/Time to Correct Problems)

Problem Correction Effectiveness
= Resources to Improve Process / Resources for Maximum Problem Correction

Resources to Improve Process = Available Time – Resources to Production

Notice that the Resources to Improve Process is the amount of time “left over” after the desired allocation of Resources 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 & 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. Following this prioritization, the model in Figure 2 does not explicitly represent the Work Smarter Loop, loop B3 from Figure 1, but it does assume that after allocating sufficient time to production so as to achieve the desired throughput, workers will use all available time for improvement. That is, there is no slacking off, and the workers are indeed doing as much improvement work as they can consistent with the goal of meeting their immediate throughput objective.

OPTIMALITY ANALYSIS

To begin exploring the relationship between the feedback structure and the dynamic behavior it generates, this section presents analytical results. The following section extends the exploration using simulations and sensitivity analysis. The analytical strategy will be 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 an organization that does little or no second-order improvement - a state characterized by a very high allocation of resources 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 (from lost production). Conversely, 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. Thus, somewhere between these two extremes there lies at least one local peak at which Net Process Throughput will be (locally) maximal.

To find the allocation to production that will yield the maximum steady-state Net Process Throughput, we first recognize that 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 Resources to Production must be in equilibrium, implying that Adjusting Allocation must equal zero, which occurs when the Resources 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 equals Desired Process Throughput.

\[ (1) \quad \text{Problem Introduction} = \text{Problem Correction} \]

\[ (2) \quad \text{Resources to Production} = \text{Indicated Allocation to Production} \]

\[ (3) \quad \text{Net Process Throughput} = \text{Desired Process Throughput} \]

It can be seen by inspection that setting the Desired Process Throughput to the optimal level and the Resources to Production to the level required to achieve this optimal throughput will satisfy the second and third conditions. We want to find an expression for Net Process Throughput as a function of the Resources to Production so we can solve this expression for optimality. Substituting the equations for the variables on the right hand side of the equation for Net Process Throughput yields an expression as a function of the two stocks:

\[ (4) \quad \text{Net Process Throughput} = \text{Resources to Production} \times \text{Productivity of Production Time} \times (1- \text{Process Problems}) \]

We now turn to condition (1) to find an expression for Process Problems that we can substitute into Equation 4. Note in Equation 1 that both Problem Introduction and Problem Correction are functions of Process Problems. Substituting the equations for the variables in Equation 1 gives:

\[ (5) \quad \frac{(\text{Unattended Process Problem Level} - \text{Process Problems})}{\text{Average Process Erosion Time}} = \text{Problem Correction Effectiveness} \times \frac{(\text{Process Problems/Time to Correct Problems})}{(\text{Process Problems/Time to Correct Problems})} \]
Rearranging yields:

\[ (6) \quad \text{Process Problems} = \frac{(\text{Time to Correct Problems} \times \text{Unattended Process Problem Level})}{(\text{Problem Correction Effectiveness} \times \text{Average Process Erosion Time} + \text{Time to Correct Problems})} \]

Equation 6 is a univariate expression for Process Problems as a function of model parameters and the variable Problem Correction Effectiveness. Substituting for Problem Correction Effectiveness yields:

\[ (7) \quad \text{Process Problems} = \frac{(\text{Time to Correct Problems} \times \text{Unattended Process Problem Level})}{((\text{Available Time} - \text{Resources to Production})/\text{Allocation for Maximum Problem Correction}) \times \text{Average Process Erosion Time} + \text{Time to Correct Problems})} \]

Substituting this expression into equation (4) yields an equation for Net Process Throughput as a function of the Resources to Production:

\[ (8) \quad \text{Net Process Throughput} = \text{Resources to Production} \times \text{Productivity of Production Time} \times (1 - \frac{(\text{Time to Correct Problems} \times \text{Unattended Process Problem Level})}{((\text{Available Time} - \text{Resources to Production})/\text{Allocation for Maximum Problem Correction}) \times \text{Average Process Erosion Time} + \text{Time to Correct Problems})}) \]

Differentiating equation (8) with respect to the Resources to Production and setting the derivative equal to zero gives a quadratic equation for the first order conditions for optimality:

\[ (9) \quad \text{Productivity of Production Time} - \left[ \frac{\text{Allocation for Maximum Problem Correction} \times \text{Time to Correct Problems} \times \text{Productivity of Production Time}}{(\text{Allocation for Maximum Problem Correction} \times \text{Time to Correct Problems} + \text{Average Process Erosion Time} \times \text{Available Time}) - \text{Average Process Erosion Time} \times \text{Resources to Production}} \right] = 0 \]
Solving with the quadratic formula finds the two roots of the equation, the optimal values of Resources to Production:

\[
\text{Resources to Production}^* = \frac{(\text{Allocation for Maximum Problem Correction} \times \text{Time to Correct Problems} \times \text{Average Process Erosion Time} \times \text{Available Time} \pm \sqrt{\text{Allocation for Maximum Problem Correction} \times \text{Time to Correct Problems} \times \text{Unattended Process Problem Level} \times (\text{Allocation for Maximum Problem Correction} \times \text{Time to Correct Problems} + \text{Average Process Erosion Time} \times \text{Available Time})})}{\text{Average Process Erosion Time}}
\]

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.

It is useful to examine a numerical example of how the steady-state Net Process Throughput varies as a function of Resources to Production. To do so, we specify a set of model parameters, shown in Table 1, that we will use for the numerical analyses presented throughout this paper. These parameters describe a stylized production setting that is reasonable and consistent with managerial experience. Figure 3 plots the equilibrium value of Net Process Throughput as a function of the Resources to Production (Equation 8) using these parameters. As the intuition described above suggested, the curve follows an inverted U-shape. Figure 3 also identifies the optimal steady-state Net Process Throughput, the maximum of this curve, with the horizontal dotted line. The omniscient manager would achieve the optimal long-term production outcome by setting the Resources to Production to the level that maximizes Net Process Throughput. But, the challenge the practicing manager faces is not so simple, for two reasons. First, notice, from Equation 10, that the value of the optimal allocation depends on the values of Allocation for Maximum Problem Correction, Unattended Process Problem Level, Time to Correct Problems, Average Process Erosion Time, and Available Time. In practice, managers are likely to have a
reasonably good estimate of the Available Time. The other four parameters all relate to the
dynamics of the stock of process problems, and managers are unlikely to have more than a
cursory understanding of them at best.

![Net Process Throughput in Equilibrium](image)

**Figure 3:** Steady-state equilibrium rates of Net Process Throughput for the range of Resources to Production

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unattended Process Problem Level</td>
<td>0.9</td>
<td>Dimensionless</td>
</tr>
<tr>
<td>Average Process Erosion Time</td>
<td>36</td>
<td>Weeks</td>
</tr>
<tr>
<td>Time to Correct Problems</td>
<td>16</td>
<td>Weeks</td>
</tr>
<tr>
<td>Productivity of Production Time</td>
<td>1</td>
<td>Unit/hour</td>
</tr>
<tr>
<td>Allocation for Maximum Problem Correction</td>
<td>4000</td>
<td>Hours/week</td>
</tr>
<tr>
<td>Available Time</td>
<td>4000</td>
<td>Hours/week</td>
</tr>
<tr>
<td>Time to Adjust Allocation</td>
<td>1</td>
<td>Week</td>
</tr>
<tr>
<td>Initial Process Problems</td>
<td>0.4</td>
<td>dimensionless</td>
</tr>
</tbody>
</table>
A second reason that the managerial challenge is far more complicated than simply choosing the optimal parameter value is that, given that the optimal choices are not a priori knowable, the manager who attempts improvement will set in motion the dynamics of this system. The manager who wishes to overcome a shortfall in process throughput (whether due to recent declines in output or recent increases in desired production) must somehow search for the choices that will achieve an improved performance level. But, as we shall see in the following section, moving along the transition path from current performance towards improvement is fraught with unexpected challenges. To examine more closely the dynamics of the underlying feedback structure constituted by the interaction of first- and second-order improvement and develop the key insights, we turn to simulation experiments.

SIMULATING THE DYNAMICS OF PROCESS IMPROVEMENT

This section presents the results of simulation experiments to investigate the dynamic behavior of the stylized production system under various improvement scenarios. The simulations begin in equilibrium conditions and use the parameter settings shown in Table 1. To parallel a real-world setting with managers and front-line workers, we assume that the manager sets a Desired Throughput target (an exogenous input in the model) and that the workers respond over time to try to achieve this target (an endogenous response generated by the model structure). This conceptualization describes a setting in which the workers take the production target as given and do their best under the available resources, using intentionally rational choices, to achieve the target. The “lever” available for a manager of this system is to set the Desired Throughput. The first experiments helps to establish basic behavior patterns of the system.
We begin with an experiment that mimics how a manager might act to achieve higher levels of performance in this system, that is by setting an explicit goal for more Desired Throughput. To model this scenario, we begin with the system in equilibrium and then introduce a one-time permanent step increase in the Desired Throughput in week 10. Figure 4 shows the response of the system to a 20% increase in the target production. The system's response to the higher throughput target begins with workers recognizing the throughput gap, updating their sense of the indicated time to spend producing in order to reduce the gap, and consequently increasing the Resources to Production. With more time spent producing, gross throughput increases and so does the Net Process Throughput, enough to achieve the new Desired Throughput by approximately week 20. Notice, however, that with higher Resources to Production, there is a lower allocation of Resources to Improvement, so Problem Correction declines, as shown in the lower left panel of Figure 4, and with a lower outflow from Problem Correction the 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 system responds and returns rather smoothly to a new equilibrium where the workers are enduringly producing a higher Net Process Throughput. Throughput has indeed increased in response to the manager's action to increase the throughput goal. But, notice two more subtle features of these results. First, although the organization is able to achieve and sustain a higher level of performance, 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. The manager's actions have resulted in trading off some process capability (i.e., the inverse of Process Problems) for a higher level of throughput - but the net effect has been beneficial. Second, the
new level of performance is still far short of the optimal steady-state net throughput, marked by the dotted line in the top panel of Figure 4.

It is instructive to examine the dynamics of the scenario in Figure 4 a bit more closely, so Figure 5 plots four variables in the simulation on the same set of axes. The Desired Throughput increases in a step fashion at time 10 weeks. As the response of Net Process Throughput shows, the workers adjust to the new goal and successfully achieve the higher level of throughput. Together, the Desired Throughput and Net Process Throughput lines can be interpreted as the manager’s view of what has happened. All appears well, and the story seems to be over by about week 24. The manager's action to increase the goal has indeed resulted in the outcome the manager was seeking, 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.

![Figure 5: Response to a Step Input Showing Four Variables](image)

The dash-dot line in Figure 5 shows the workers' allocation of Resources 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. Even as the manager sees that the new throughput target is being reached, the workers are shifting more and more of their time away from improvement in
order to keep up with the production goal. Consequently, as the dashed line shows, reducing the
time spent on problem correction (because more time is spent on production) manifests as a slow
increase in process problems. More process problems would reduce net throughput, all else
equal, but net throughput stays high because workers continue to allocate even more time to
production. They are caught in a treadmill created by the reinforcing Reinvestment loop of
Figure 2 working as a vicious cycle. Although the system does adjust to achieve the new
equilibrium, the workers' very success in doing so masks the problematic nature of what is
happening. Workers are stuck on this treadmill chasing more gross throughput in order to meet
the objective while the system capability, as indicated by Process Problems, actually deteriorates.
Most insidiously, the salient signal to the manager - the Net Process Throughput - indicates not
only that things are doing fine but also that the managers' action to push harder on the goal has
been successful. The results of this simulation provide a vivid foundation for the dysfunctional
attributions that Repenning and Sterman suggest managers might make - that is, that inadequate
worker effort is the cause of the shortfall and that the rational response is to pressure them to
work harder.

As we saw in Figure 4, the step increase in Desired Throughput stimulates an increase that is
sustained but that still leaves the organization performing below the optimal level. It is tempting
to suggest that managers should set the Resources to Production to the optimal level as given by
Equation 10 (or set the corresponding optimal Desired Throughput). However, as discussed
above, this is not a practical option given the difficulties of knowing the values of the necessary
parameters. Instead, managers must discover other policies to manage this system. The next
simulation experiments explore how the system responds to other attempts to achieve higher
performance.
In the end state of the previous scenario, the system is producing more than before the managers' intervention but still less than the optimal throughput, so next we consider next the effect of setting Desired Throughput even higher. Figure 6 shows the results of a larger step increase, one that (unbeknownst to the manager) brings the Desired Throughput above the optimal level. As before, the workers' response is to increase the Resources to Production, which again has the immediate effect of increasing the Net Process Throughput, as more time doing production activities boosts Gross Process Throughput. However, the increase in Resources to Production is accomplished at the expense of a decrease in the Resources to Improvement. 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. With this higher Process Problem level, even more Gross Throughput is needed to achieve the target net throughput, and workers allocate even more Resources to Production. The Reinvestment Loop R1 works as a vicious cycle, and the organization gets locked into a downward spiral, shifting more and more time away from improvement activities even as the Process Problems continue to grow. But, the workers continue to achieve the target until approximately week 90, accomplishing this throughput objective at the expense of a deteriorating process. By then, all Available Time has been allocated to production, so as Process Problems increase still further, the workers are unable to boost Gross Throughput high enough to meet the goal and Net Process Throughput begins to fall. Worse, as the throughput pressure has squeezed out all improvement activity, Process Problems increase still further until they reach their natural limit, dragging down the Net Process Throughput as they grow. 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 top graph in Figure 6 includes a
heavy dotted line that identifies this equilibrium level as the “No Maintenance Throughput,” the rate achieved when Process Problems are at the maximum given by Unattended Process Problem Level and the Resources to Production are at the maximum given by Available Time. The overly aggressive target for throughput has led to a period of temporary success but has also triggered  a downward spiral that drives the system to this underperforming extreme.

The set of simulation experiments shown in Figures 4, 5 and 6 highlights three key features of the system's dynamics. First, 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 4. However, if the targeted increase crosses a critical threshold, the mode of

Figure 6: Response to a Step Increase above the Optimal Steady-state Net Process Throughput
behavior changes to one that displays a better before worse pattern, as in Figures 5 and 6. 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. Coupled with this potential for tipping, the oft-observed managerial logic to "push just a little harder" may unwittingly trigger a vicious cycle that is difficult if not impossible to recover from. Second, once the deterioration of system capability begins, it is self-sustaining. Once past the tipping point, the increasing stock of Process Problems causes more defects and therefore less Net Throughput, forcing the workers to shift even more time away from improvement, and the vicious cycle gains even more momentum. The exponential decline in the Resources to Production shown in Figure 6 is a tell-tale sign of the dominance of this vicious reinforcing loop. Third, system performance continues to deteriorate even after the workers have completely shifted their time away from improvement. Because of the stock of Process Problems increases only somewhat slowly relative to the other aspects of the system, the absence of problems endows the process with a certain capability to continue to produce, even when this stock is neglected. We see here that this process capability is a form of organizational slack that is eventually consumed.

Our final set of simulation experiments explores the role of slack in the interaction of first- and second-order process improvement. Recall the curvilinear shape of the steady-state net throughput as a function of the Resources to Production shown in Figure 3. Our previous simulations all began with an initial allocation of Resources to Production that is less than the optimum, that is, to the left of the optimum as marked in Figure 3. The horizontal line in Figure 3 that passes through this initial allocation also intersects the Net Process Throughput curve at another point, highlighted in the figure to the right of the optimum, marking another allocation of
resources that will achieve the same steady-state Net Process Throughput. The second intersection describes a system equilibrium with a higher allocation of Resources to Production, therefore fewer Resources to Improvement, and larger stock of Process Problems. Figure 7 shows the system's response to the same 20% step increase in Desired Throughput as before when starting from the alternative initial conditions in Figure 3 that achieve the same initial Net Process Throughput (which using the current parameter values occurs when Initial Process Problems are 0.75 and Initial Resources to Production are 3600 hrs/week).

![Graph](image.png)

**Figure 6: Response to a Step Increase above the Optimal Steady-state Net Process Throughput**

Starting from this state of high process problems, the system in unable to adjust enough to achieve the new Desired Throughput. There is a very short period of modestly improved performance, as the workers discontinue the little time they are spending on improvement and shift the time to production, but doing so reduces Problem Correction and Process Problems begin to grow. Just as we saw in the previous scenario, the system deteriorates until it reaches the No Maintenance Throughput level, characteristic of the system when all improvement activity has been squeezed out. Moreover, just as in the previous scenario, the action taken in
the hopes of improving performance has triggered a dynamic response that leads to a significant worsening of performance, an outcome exactly opposite of what was intended.

The simulations in Figure 6 highlight two key features of the systems' dynamics. First, despite the identical net process throughput at the start and the identical stimulus, the system produces qualitatively different patterns of behavior. In systems prone to these tipping dynamics, managerial lessons learned in one setting not only do not transfer well to another, but may actually translate to the wrong actions in a different setting. Second, the distinction between the outcome of improvement and the outcome of demise hinges on an attribute of the process capability - the stock of Process Problems - that may have a rather low degree of salience. Given the elusive nature of this critical distinction, even the manager who has some rudimentary understanding of the potential dynamics of this system may have difficulty recognizing where the organization's current situation lies on the spectrum of improvement potential.

DISCUSSION

This paper develops a system dynamics model to examine the dynamics of first- and second-order process improvement, using the conceptual causal loop diagram from Repenning and Sterman (2002) as a starting point. By moving beyond a causal loop diagram to a simulating system dynamics model, we are able to more rigorously examine the dynamics and discover several important insights into the managerial challenges these systems present, especially when resources are constrained. Both first- and second-order improvement options are effective in boosting short-term organizational performance, whereas only second-order improvement builds sustainable capability. But as our simulation experiments show, the dynamic interactions that unfold when attempting to implement improvements in organizational performance are far from
straightforward trade-offs. Instead, this work highlights three particularly problematic implications of the feedback structure of process improvement.

First, while first-order improvements (e.g., allocating more worker time to production) can boost output, they may do so at the expense of compromising process capability (e.g., capability to produce with low defect rates). Shifting worker time to first-order improvement rather quickly increases the net process throughput, but the consequences of shifting time away from maintaining and improving process capability manifest only later, as process capability decays over time due to lack of attention. While such inter-temporal tradeoffs are widely documented in the system dynamics literature, the dynamics here pose a particularly vexing managerial challenge. Because the success of first-order improvement in boosting output provides a salient signal that things are going well, it often masks the underlying, less salient degradation of process capability. More insidiously, continued reliance on first-order improvement, as might be encouraged by these very signals that things are going well, compensates for and continues to mask the degradation of process capability. Although the workers who industriously work harder to overcome the deficits in the process even as the process is deteriorating may be well aware that the neglect of second-order improvement is causing other problems, managers more distant from these front line activities may easily be misled. Observing the increased output, the manager might be fooled into believing not only that stretching the goal was a successful intervention but even that stretching still further would be warranted.

Second, our simulation analyses underscore the role of an organization's built-up stock of process capability as a form of organizational slack. As our final set of simulations
demonstrated, when process capability is high (i.e., the stock of process problems is low) the organization is positioned to "trade-in" some of this process capability in exchange for higher rates of performance. Process capability is a stock that must be built-up (and maintained) over time, but once accumulated it deteriorates only slowly. Following the language of strategy scholars, process capability is an example of a strategic asset stock that accumulates time paths of flows over a period of time and cannot be adjusted instantaneously (Dierickx & Cool, 1989). This "stickiness" of the accumulation is the essence of the dynamic behavior of any stock variable, but the important implication here is that the economic value of this stock of process capability can be realized in one of two ways: as a capital-like asset that can be operated to yield production or as a resource pool that can be drawn down while workers do production activities that generate production. The stock of process capability is the "enabler" of production activity of the workforce, but it is also the "accumulator" of process improvement activity by the production same workforce. Because the resources that could be used to conduct process improvements, maintaining an inflow to the stock, can be redeployed to instead conduct production activities that generate immediate value, the stock endows the organization with a form of organizational slack. Traditional conceptualizations of organizational slack focus on the notion of surplus resources relative to the output requirements for the organization (Cyert & March, 1963; Cyert & March, 1956). The analysis here suggests alternatively that organizations may be endowed with (or better, may have accumulated) surplus capability that, just like surplus resources, means the organization if it so chooses can achieve higher production with the current level of resources.
That organizational slack can exist in the form of surplus process capability has two important managerial implications. First, when surplus process capability exists, managers have the option of trading it in to boost production. Managers may benefit from making explicit the understanding that they are trading in surplus capability in order to boost output. In particular, due to the curvilinear nature of the relationship of the production rate to the resource allocation between production and capability building (See Figure 3), attempts to shift this allocation more towards production will yield a higher production rate sustainably only when surplus capability exists (i.e., to the left of optimal in Figure 3). In other cases (i.e., when right of optimal in Figure 3), a similar shift in allocation towards production will set the system on a path towards the vicious cycle of shifting more and more resources to production in response to immediate pressures, the pathologies of deteriorating capability masked by working harder, and the misattributions that such scenarios foster (Repenning & Sterman, 2002). Second, intentionally accumulating some slack process capability is both a potentially useful buffer for the future and a means of "storing" excess capacity. During periods when a firm has excess production resources (perhaps due to a temporary decline in demand), a typical response is to downsize the workforce. However, we speculate that an alternative is to deploy the surplus production resources (that are typically generating the flow of productive output) to build process capability, thus accumulating a stock of slack that can later be traded in. These two implications are complements of each other: the first says that when the capability is in surplus, it can be consumed, and the second says that when production worker time is in surplus, it can be used to accumulate capability. This policy may have significant disadvantages, especially if overreliance leads to chronic cycles of building and then eroding capability. But, if the common response is to use layoffs during market declines, managers may find the disadvantages of such capability cycles are less
problematic than those associated with chronic cycles of hiring and downsizing. More modestly, convincing managers that available resources can be usefully deployed to do second-order improvements, despite the challenge that they are less salient, more uncertain, and yield benefits only with a delay, is an important step forward.

A third implication of the feedback structure of process improvement is that production systems facing the interactions of first- and second-order improvement alternatives have tipping points, thresholds of stretch goals beyond which performance rapidly deteriorates. Simulation analysis shows that a managerial intervention to increase the performance goal can indeed 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 a 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 immediate response to an “inappropriate” increase in a goal to a level in the dangerous region above the tipping 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.

Analytically, we find the allocation of worker time between first-order (production) and second-order (problem correction/capability building) activities that optimizes the long-run steady-state net process throughput. 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. Examining Equation 10, we see that the optimal allocation of Resources to Production is a function of the Allocation for Maximum Problem Correction, the Time to Correct Problems, the Average Process Erosion Time, the Unattended Process Problem Level and the Available Time. All of these parameters represent recognizable real world concepts, but managers will often have reasonable knowledge about only the amount of Available Time. The other four parameters are factors governing the inflows to and outflows from the stock of Process Problems (or more generally its inverse, the stock of process capability). Despite the clear importance, as the analysis here demonstrates, of these parameters in understanding the dynamics of production systems undergoing improvement, knowledge about them remains rather elusive. Future research to carefully examine the dynamics of stocks of process capability is needed. In the absence of good knowledge of these parameter values, policies for managing must be discovered by experimentation - with the attendant risks of pushing past the tipping point as we saw in our simulation experiments. Even if good estimates of these parameters were available, policies based on point solutions to the allocation decision are not likely to be robust to changes in the marketplace, the production environment, the
available technology, and even the workforce itself. Research to develop managerial policies suitable for systems that are prone to tipping would also be useful. Tipping points have been described in a variety of dynamical systems, such as fads, the behavior of crowds, the ethnographic transformation of city neighborhoods, climate, disease epidemics, product development processes, and disasters (Alley et al., 2003; Gladwell, 2000; Granovetter, 1978; Lenton et al., 2008; Repenning, Goncalves, & Black, 2001; Rudolph & Repenning, 2002; Sterman, 2000). Developing theory for managing near the tipping point and using theory to generate a suite of managerial practices - in this system and more generally - holds great promise for enhancing our ability to cope with the vexing challenges posed by the interactions of first- and second-order improvement. Such theory and practice tools might include identifying signals that a system is approaching a tipping point as well as actions to take once those signals are noticed.
REFERENCES


