Towards Evaluation of Systems Thinking Interventions: A Case Study

Steven Cavaleri
Department of Management
Central Connecticut State University
New Britain, CT 06050

John D. Sterman
MIT Sloan School of Management
Cambridge MA 02142

Introduction

Recent innovations in systems thinking have fueled growing interest among managers in the practical application of the tools (Senge 1990, Morecroft and Sterman 1994, Senge et al. 1994). A number of technologies and protocols are useful for developing systems thinking capability in organizations and individuals, such as management flight simulators, experiential exercises, and causal loop diagramming. Although various intervention techniques that fall under the rubric 'systems thinking' have become quite popular, little is known about their effectiveness in enhancing organizational effectiveness or productivity. In general, the relationship between the use of systems thinking and organizational performance remains the province of anecdote rather than rigorous follow up research. In this paper we argue that such rigorous follow up research must be developed if we are to build a strong foundation for the effective use and refinement of the tools of system dynamics and systems thinking. At the same time, such evaluative research is extremely difficult. In this paper we evaluate a well-known and often-cited systems thinking intervention in an organization. The evaluation suggests the intervention did have positive effects on the organization, but because the original intervention was not designed with evaluation in mind, the study also illustrates many of the difficulties encountered in conducting such evaluations.

The history of the intervention is well told elsewhere (Senge 1990, Senge and Sterman 1992, Moissis 1989, Bergin and Prusko 1990); we summarize briefly here. In the late 1980s members of the MIT System Dynamics Group worked with a top management team of the Hanover Insurance Company, a mid-size property-casualty insurer, to develop a system dynamics model describing the interaction of claims management, quality, and costs. The model suggested important insights into the industry-wide problem of rising costs and falling quality, including some high-leverage policies to improve the situation. To diffuse the insights from the model more widely throughout the organization, the model was converted into an interactive management flight simulator, the "Claims Game". The flight simulator was incorporated into a 'learning laboratory' in which the participants were introduced to various systems thinking tools, play of the "Beer Game" (Sterman 1989), a workshop with the claims game, and a seminar which dealt with systems thinking skills, such as causal loop diagramming. The learning laboratories were initially run at corporate headquarters with participants from different regional offices and functions, but were eventually devolved to individual regional offices where intact management teams could participate. The program began in 1988 and continued through 1991.

Data Sources and Level of Analysis

We sought to evaluate the impact and effectiveness of the intervention. The focal point of the research is assessing changes in the attitudes, practices, and business results subsequent to the training program. Our research takes a 'formative perspective' (Gagne, 1985) in the sense that generally accepted tools and processes for evaluating the impact of systems thinking on organizational performance do not yet exist. The original intervention had not been designed with longitudinal evaluation in mind. Thus we were forced to conduct a retrospective study, greatly complicating the task of assessment. We return to this theme in the conclusion.

Our evaluative research was conducted in a regional office of the company. The primary

function of the office, which employs about seventy people, is claims management. The office had participated in the workshops as an intact team. The training was aimed at individual members of the organization and at the claims unit team. The claims unit is one of the primary determinants of the profitability of the office. In our evaluation we considered all employees of the regional office, but distinguish between managerial and non-managerial personnel because the managers received more extensive training than the non-managers. All employees attended the Beer Game module, but the Claims Game flight simulator session and systems thinking seminar were attended primarily by members of the claims unit.

Data collection commenced with a series of interviews with the key managers of the regional office. The interviews were used to convey the goals of the research and to identify indicators for assessing changes in performance. We also solicited the managers' views regarding the usefulness of the training program. A twenty-two item questionnaire was designed and distributed to all of the participants in the training who were still with the organization (full documentation is available from the authors). The questionnaire was designed to measure any changes in the systems thinking capabilities of the participants, including any differences in personal perceptions or behavior since the training and also to measure their ability to recognize specific systems principles demonstrated in the Beer Game and system thinking training sessions. Following each question opportunities were provided for respondents to offer clarifying comments.

Hypotheses

Three claims have been made in the literature about the effectiveness of systems thinking interventions of this sort: that they alter thinking, behavior, and results. In the context of this work, the first claim is that the insurance claims management learning laboratory should have altered people's *mental models* to be more systemic and more aware of the long-term dynamics of the business. The hypotheses below distinguish between the managers and the other workers in the regional office we examined. Thus,

H1-A: Managers attending the training will have an increased capability to think systemically.
H1-B: Participants other than managers will have an increased capability to think systemically.

Second, changes in *behavior* consistent with the long-term high-leverage policies identified in the simulation analysis should also be observed. Thus,

H2-A: Managers attending the training will experience changes in their patterns of behavior and organizational policies consistent with the long-term best interests of the system as a whole.
H2-B: Participants other than managers will experience changes in their patterns of behavior consistent with the long-term best interests of the system as a whole.

Finally, it is expected that changes in behavior will improve business results. Thus,

H3: The claims unit's operational performance will improve as a consequence of the training.

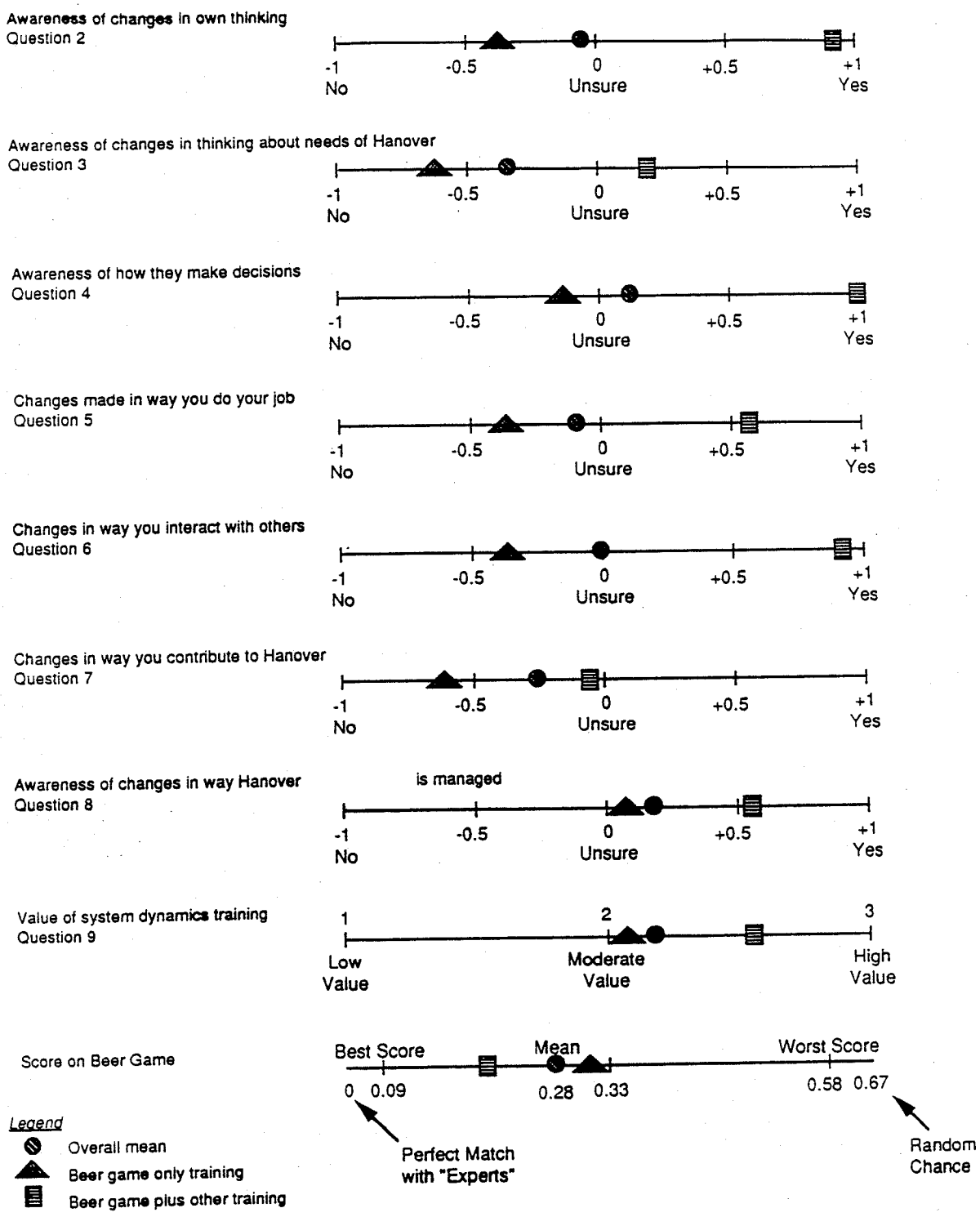
There are several common indicators of claims unit performance: (1) the claims pending ratio (the ratio of pending [unsettled] claims to the flow of new claims); (2) the production ratio (the ratio of claims settled to incoming claims); (3) average settlement size ($ per feature; a feature is a particular loss described in a claim [complex claims consist of many features]); and (4) average administrative cost per feature.

Results

Participant Attitudes and Beliefs (Survey Results)

The survey was given to all employees of the office who participated in the systems thinking intervention (59 out of a total of 70 employees). 36 usable responses were received (61%); 9 were managers and 27 were non-managers. The response rate is excellent; the main reason for nonresponse was that some personnel were on vacation at the time the survey was

Figure 1. Summary of Questionnaire Results.



administered. The survey results are organized in four areas: (1) self-reported cognitive changes, (2) self-reported behavioral changes, (3) perceived changes in the way the company is managed, (4) competency in understanding the principles demonstrated in the Beer Game (Figure 1).

1. Cognitive Change:

The overall effect of the training on this dimension is mixed. Non-managers, who received only the beer game component, report little awareness of changes in their thinking. Managers, who took the entire training, report much greater awareness of changes in their thinking.

2. Behavioral Change:

The overall effects of the training on behavior are unclear. There is no evidence of change in behavior among non-managers, but clear evidence managers believe their behavior changed as a result of the intervention. The interviews strongly suggest that the behavior change was a direct effect of the training.

3. Changes in Management Style

The survey data provide evidence of only marginal changes in the way the office is managed. Non-managers only noted a slight change in this area. However, the managers report that their own management style has changed and become more "systemic". Of course, the meaning of 'systemic' to these managers is somewhat ambiguous.

4. Understanding The Key Principles of the Beer Game

In the questionnaire, respondents were asked read a series of short case descriptions and rate the extent to which the lessons of the Beer Game were illustrated or relevant to the case. Responses were compared against the average rating of a panel of three experts. The scores ranged from .09 to .58 (where 0 would indicate perfect concordance with the experts), and .67 is the rating achieved if responses were random. The overall mean was .28. The mean for the managers was .19, while for non-managers it was .31. The participants appear to have learned some, but not all, the lessons of the beer game. Managers did better than nonmanagers.

Manager Comments

The interviews conducted with the management team and the written comments from the questionnaire were quite revealing. Respondents report that they developed new insight into the causal relations at work in their center by playing the Claims Game. In particular, the visualization of the interaction between the stocks and flows within the claims unit helped in the design of new policies. Several managers noted that playing the claims game helped to reinforce their prior understandings of systems thinking. It also helped them to understand the interconnectedness of elements in the system and to see the tradeoffs generated by alternative courses of action. Managers reported that the systems thinking training program helped shift many people's thinking from a reactive to more strategic mode, giving them an edge over competitors who rely on a traditional view of managing.

In sum, managers report that the training affected their mental models (attitudes and beliefs), and that they understood and were able to apply the principles of systems thinking to their jobs and in their interactions with colleagues. Other personnel do not report such changes. It is not possible to say whether the differences are due to the differences between the managers and other personnel or to differences in the amount of training in systems thinking they received, since the manager/worker distinction is confounded with the full training/partial training distinction.

In interpreting these data one must be careful of demand effects, that is, of being told what the informants believe you want to hear. Though we took pains to tell the managers that we were interested in an honest assessment, it is quite possible that the managers believed we were looking for a positive outcome, or that they themselves were invested in a positive outcome since they had committed substantial time to the training process. Such demand effects are a common problem in field work, especially when relying on retrospective reports.

Changes in Behavior

We next sought to document specific changes in behavior flowing from the intervention. These artifacts of change would show the tangible impact of the learning laboratory directly, rather than by unreliable retrospective self-reports. Our interviews and archival data revealed a range of changes in behavior, policy, and organizational structure (table 1). Many of the changes can be traced directly to the training. In interviews, managers were explicit in relating these new policies to their new mental models of the causal structure of the claims unit. They developed several new strategies to implement key recommendations emerging from the model, including recruiting more experienced and higher quality adjusters, attempting to retain experienced adjusters longer, focusing on the quality of the settlements rather than measures of throughput, and increasing total settlement capability by hiring new adjusters (Senge and Sterman 1994).

Table 1. Summary of Behavioral Changes and New Policies

I. Company policies:
1. New statement of performance expectations developed
2. Work quality redefined

Quality was emphasized more compared to throughput, and quality was redefined to include the long-term and system-wide effects of claimant contact, negotiation, documentation, and investigation

II. Structure:
1. supervisor converted into adjuster
2. adjuster jobs redefined
3. cases assigned to adjusters in new way

Claims were assigned to adjusters to better match expertise to the complexity of the claims. The result was faster case resolution with less stress.

III. Hiring:
1. new selection criteria implemented, stressing those with an aptitude for systemic thinking.
2. new recruitment methods implemented, in particular, pre-recruitment networking with experienced adjusters throughout the industry was instituted.
3. time to fill vacancies reduced
4. ideal candidate profile redefined
5. interviewing process changed (new questions added to select candidates with systemic thinking capability)
6. hiring experienced adjusters emphasized
7. addition to staff of adjusters requested

These measures resulted in a substantial increase in net adjuster headcount and effectiveness, even though the request to corporate headquarters to increase the authorized headcount (#7) was denied.

IV. Training:
1. Goal of training altered to stress adjuster empowerment; adjusters were given the ability to spend time to maintain quality in face of pressure to meet throughput goals.

V. Environment:
1. Outreach to attorneys, explaining systemic issues, to help legal staff understand the changes in policies at the claims office.

As examples of these changes, consider the hiring and staffing policies. The simulation model suggested that long-term, system-wide costs could be reduced by increasing the

organization's capacity to settle claims, and then using that extra capacity to increase the quality of settlements rather than increase the throughput. To increase capacity requires increasing the skill and experience of adjusters, as well as the total headcount. Headcount limits for the regional offices are set by corporate headquarters. The regional office did request an increase in their authorized headcount cap, but the request was denied. The office managers cleverly realized they could increase effective adjusting capacity without an increase in the authorized headcount of the office by filling the vacancies created by turnover faster. Before the learning laboratory management's goal was to minimize administrative expenses by delaying the replacement of departing employees. The average number of adjusters was well below the authorized headcount. After the intervention, the goal became rapid replacement of departed employees, effectively increasing headcount and adjuster capacity without requiring corporate approval (table 2).

Table 2. Average Time to Fill Vacancies in Claims Adjuster Positions

	1986-1988	1988-1992
Average time to fill vacancies (weeks)	16.6 weeks	4.6 weeks
Range	5 - 20 weeks	1 - 12 weeks

The organization also sought to reduce turnover, thus stanching the outflow of experienced adjusters. Turnover did decline after the intervention, falling from 27%/year in 1988 to 9% in 1991 (turnover rose substantially in 1992, however; see table 3). Two notes of caution: First, some of the turnover is desired, as it is necessary to weed out poor performers. We have no way to distinguish between 'desired' and 'undesired' turnover. Second, the decline in turnover, while consistent with the implementation of the new policies, can also be explained by exogenous events: the decline in turnover coincided with the national recession in 1990-1991. During recessions, voluntary quits drop as workers find it harder to land new jobs. Thus the decline in turnover could have been the result of changing macroeconomic conditions. It is simply not possible to rule this out with the small sample size and limited data available.

Table 3. Rate of Employee Turnover

Year	1988	1989	1990	1991	1992
Turnover (%/year)	27%	18%	18%	9%	27%*

* (9 month period, January-September 1992)

Changes in Business Performance

In the end, the effectiveness of any intervention rests on the changes in business outcomes (the so-called 'system improvement test'; see Forrester and Senge 1980). To be judged effective, an intervention must do more than affect attitudes and beliefs; it is not enough that participants like the intervention and rate the workshops highly. An intervention must also have positive effects on the states of the system. The performance of the claims unit was measured in terms of four standard measures used throughout the industry. These measures are tracked routinely within the company (table 4). Performance was examined over a six year period from 1986 through 1992. During this time period the number of incoming cases grew only slightly, by 4%, however, there was a qualitative increase in the complexity of the cases, as judged by the managers. In interpreting the data note that there is simply too little available from the organization to conduct statistical tests.

Table 4. Summary of Claims Unit Performance Indices. Numbers in **bold** are national averages.

Year	Pending Ratio	Production Ratio	Settlement Size ($)	Admin. Cost per Feature
1986	2.67	102.5	1432	115.25
	2.45	**98.4**	**1639**	**143.88**
1987	2.75	98.0	1895	166.93
	2.43	**98.3**	**1597**	**168.19**
1988	2.69	100.1	2102	195.54
	2.41	**98.8**	**1711**	**198.49**
1989	2.69	99.8	2443	243.42
	2.46	**101.0**	**2016**	**214.18**
1990	2.37	103.15	2898	233.45
	2.46	**99.92**	**2333**	**246.42**
1991	2.37	99.11	2719	277.75
	2.47	**100.84**	**2529**	**279.03**
1992	2.65	97.68	3222	274.24

The lack of large samples is partly the result of the fact that we were forced to conduct a retrospective evaluation. The pending ratio compares the number of claim features that remain unresolved against the number of features received per month. There is no definitive change discernible in this performance measure. The pending ratio improved 13% from 1989 to 1991. However, in 1992 it rose to 2.65 from 2.38 the previous year. The target pending ratio is between 2.00 and 2.25 months, depending on the mix of business. It becomes increasingly difficult to make further gains as the ideal ratio is approached.

The production ratio is the ratio of claims settled to claims received, and should not be less than 1.0 (in equilibrium incoming claims = settlement rate, so production ratio = 1). No clear pattern of improvement is evident. Over a six year period performance was both above and below the overall company average. By 1992 the ratio had fallen 1.4% below the prior year and 5.6% from the level of 1990. Comparisons to national averages are problematic as the incidence of disasters that can raise the incoming claim rate and thus reduce the production ratio are not distributed uniformly throughout the regions of the country in which the company writes business.

Average settlement size for a given time period is calculated by dividing the total payout for claims settlement by the total number of features settled during the same time period. Again, the data suggest no definite pattern of improvement. Pay-outs to customers in this regional office have typically been higher than the national average due to the higher cost of living and of auto repair in the states served by the office. Payouts declined in 1991 but increased substantially in 1992. Settlement size is influenced by regional factors such as cost of living, weather, and economic conditions. Substantial time delays may arise in reporting the settlement of cases due to accounting conventions. A case is not credited as settled until all features are resolved. The rise in costs in 1992 may also be explained by the change in the mix of claims for this office or by changes in underwriting standards. These exogenous variables confound the interpretation of the results.

Settlements involve a direct payout of funds to claimants but also incur administrative costs such as appraisal fees, adjuster salaries, legal fees, and overhead. Attorney's fees are the largest component of this index. High quality adjusting reduces the need for legal intervention; it was

believed that increasing the quality of claim adjusting would reduce the need for litigation and subrogation, reducing expenses (Senge 1990). Administrative costs show no improvement. The office's performance runs roughly parallel to the national company average; both increased steadily during the period considered.

In summary, there is no compelling evidence to suggest the overall performance of the claims unit has improved. Skepticism as to the business benefits of the intervention is warranted. However, the weak evidence does not mean the intervention was unsuccessful: the data cover a comparatively short period of time, while the model suggested many years are required for improvement to manifest. The financial results may reflect decisions taken many years before the intervention. Further, the model suggested a short-term/long-term tradeoff in which performance suffers between the time the organization invested in the quality of its adjusters and the time the effects of superior quality manifested in settlement costs, litigation success and volume, and reduced incidence of fraud. If so, then a deterioration in financial results after the intervention is consistent with its success (still, the lack of any period of 'better' results after the 'worse' results is not reassuring). The business results are simply too coarse and too recent a set of measures to resolve this issue definitively. The difficulty of relating business results to particular interventions in a complex dynamic system is both a thorny problem in evaluative research and a chief reason that organizational policies often produce dysfunctional results (Sterman 1994). We now provide more detail on the confounding issues

Externalities: Confounding Variables

One of the difficulties of longitudinal research is that the environment inevitably changes along with the changes introduced by an intervention or experimental treatment. Such is the case here. In particular, a number of substantial changes in the environment of the firm and insurance industry confound the interpretation of the data above.

1. Changing Mix of Business

The types of claims handled by the office changed in two significant ways since the intervention. One of the shifts has been evolutionary, the other is more sudden. The claims handled by the office became more complex. In general, they included more features and were more frequently the subject of litigation. Second, the branch office was given responsibility for processing claims on policies issued by a recently-acquired business unit. The new unit tended to write policies involving greater risk and requiring more time to settle. A higher portion of their claims involve personal injury claims which tend to be heavily litigated.

2. Change in Management

Subsequent to the start of this research the parent corporation experienced several large scale changes in ownership and management. Controlling interest in the firm was acquired by a former minority stockholder, and there was significant turnover in the top management team. The new team emphasized cost reduction as a primary focus, rather than quality improvement, as the systems thinking intervention stressed.

3. Change in Management Philosophy

The new top management team of the corporation adopted a different management philosophy. The impact of this decision on the branch office is hard to assess. A number of the new policies established following the training have since been 'reassessed' in light of the new management position.

Discussion

The hypotheses offered above suggest that the claims management learning laboratory should have altered the way people think about their business to be more systemic, altered their behavior to reflect high leverage policies for system-wide, long-term improvement, and, as a result, improved business results.

The results of our evaluation are mixed. The questionnaires, written comments, and

interviews do support hypothesis 1-A, that managers did experience a shift in their mental models towards a more systemic understanding of the claims system and its dynamics. The managers who participated in all of the training modules believe that their thinking shifted significantly towards a more systemic style. However, there is no evidence to support the idea that the intervention helped people other than managers to think more systemically (H1-B is not supported).

The evidence for behavioral change is again mixed. There is no doubt that behavior did change among managers (H2-A). New policies for recruiting, hiring, training, and retaining adjusters were put into place. Changes were made in the nature of the work, the assignment of work to adjusters, and in the emphasis of quality compared to throughput. However, the evidence for behavior changes among the non-managers is weaker (H2-B is not supported).

Finally, there is essentially no support for the claim that the intervention produced measurable improvements in business performance (H3). Neither the pending ratio, production ratio, settlement costs nor expense ratios show any consistent patterns of improvement.

The weak evidence of performance improvement seems at odds with the stronger evidence of change in mental models, behavior, and organizational structure and policy. The reasons for this are not fully known. There are several possibilities. One is the presence of confounding changes in the business environment. Since the intervention was not designed as a controlled experiment, it could always be argued by a supporter of the intervention that performance would have been even worse without the training, or that it is too early to see the beneficial effects, even though some four years have elapsed. We point out these possibilities not because we believe them; on the contrary, one should be highly skeptical of such arguments, especially when advanced by those with a vested interest in the success of the intervention. Rather, we point out these possibilities to highlight the difficulty of drawing strong inferences about the effects of an intervention when the intervention was not designed as an experiment or as a prospective study.

Finally, it may be that we are properly measuring the impact of the intervention, but that the theory developed in the original simulation study is wrong, so that the changes the intervention led to do not in fact lead to improvement. Despite the relatively long time that has transpired, the use of multiple data sources, and the extensive cooperation we received from the personnel at the regional office, it is simply not possible to rule out these competing hypotheses.

Methodological Conclusions

The ambiguity in the assessment of the changes in business performance highlights the main conclusion we draw from this study. Impediments to learning about complex systems such as a systems thinking intervention in a large firm are well documented (Sterman 1994). But because the original intervention was not designed from the start to facilitate assessment and evaluation these impediments were intensified. The lack of baseline data taken in real time before and during the intervention severely hampered our ability to discriminate among competing hypotheses about the impact of the learning laboratory. Field study and action research are difficult enough; these difficulties are compounded when the evaluation is forced to be retrospective because there was insufficient attention to the requirements of evaluation at the start of the study. It is understandable that proper design and resources for evaluation and assessment suffer when action researchers begin a new project. The job of identifying potential partner organizations, negotiating entry, building trust, and working with the partner team to understand the business dynamics is demanding. Often, members of organizations seeking help do not appreciate the benefits they will gain from a commitment to evaluation and research. Yet this study shows how such behavior leads to a short-term/long-term tradeoff. While we uncovered tantalizing suggestions that the systems thinking intervention had a significant impact on the organization, the evidence is weakened by the fact that we were forced into a retrospective evaluation. Had the original intervention been designed from the start with an eye towards rigorous evaluation of its effectiveness, the evidence might be stronger, and the potential to learn about the dynamics of system thinking interventions in organizations might have been fully realized.

References

Bergin, R. and Prusko, G. (1990). The Learning Laboratory. *Healthcare Forum Journal*, 33(2), 32-36.

Forrester, J. and P. Senge, (1980) Tests for Building Confidence in System Dynamics Models. in Forrester, Legasto, and Lyneis (eds.) *System Dynamics*. TIMS Studies in the Management Sciences. New York: North Holland, 209-228.

Gagne, R. (1985). *The Conditions of Learning*, 4th edition. New York: Holt, Rhinehart & Winston.

Moissis, A. (1989). Decision Making in the Insurance Industry: A Dynamic Simulation Model and Experimental Results. Unpublished MS thesis, MIT Sloan School of Management, Cambridge, MA 02142

Morecroft, J. and J. Sterman, eds. (1994) *Modeling for Learning Organizations*. Portland, OR: Productivity Press.

Senge, P. (1990) *The Fifth Discipline*. New York: Doubleday.

Senge, P. and J. Sterman (1992). Systems Thinking and Organizational Learning: Acting Locally and Thinking Globally in the Organization of the Future. *European Journal of Operational Research*. 59(1), 137-150.

Senge, P. et al. (1994) *The Fifth Discipline Fieldbook*. New York: Doubleday.

Sterman (1989). Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Experiment. *Management Science, 35*(3), 321-339.

Sterman, J. D. (1994). Learning In and About Complex Systems. *System Dynamics Review, 10*(2-3), 291-330.