Chapter # (editors will provide this info)

Measuring Change in Mental Models of Complex Dynamic Systems

James K. Doyle and Michael J. Radzicki
Department of Social Science and Policy Studies
Worcester Polytechnic Institute
doyle@wpi.edu
mjradz@wpi.edu

W. Scott Trees
Department of Economics
Siena College
trees@siena.edu

As scientists who are interested in studying people's mental models, we must develop appropriate experimental methods and discard our hopes of finding neat, elegant mental models, but instead learn to understand the messy, sloppy, incomplete, and indistinct structures that people actually have.

Donald A. Norman [1983, p. 14]

1.1. Introduction

A complex dynamic system can be thought of as a collection of interrelated variables whose structure determines the behavior of the system over time. The more variables that are involved and the more highly interconnected they are, the greater the complexity of the system. In addition to tight coupling of variables and a tendency toward self-organization, the features that give rise to dynamic complexity include the existence of feedback mechanisms, nonlinear relationships between variables, irreversible processes, adaptive processes, and time delays [Sterman 2000]. Human failure to control the behavior of such dynamically complex systems in desired ways has been demonstrated by case studies and experiments on human subjects in a wide variety of domains, including urban planning [Forrester 1969; DÖrner, 1996], inventory control [Sterman 1989], resource management [Moxnes 1998], medical care [Kleinmutz and Thomas 1987], and forest fire control [Brehmer 1989]. In all of these domains experimental subjects have been found to have great difficulty anticipating side effects and predicting the long-term impacts of their decisions, often resulting in severe negative consequences for the human actors in the system. Another common finding in these studies is the inadequacy of subjects' internal mental representations of the systems they are trying to control, often referred to as "mental models."

It should be noted here that the term "mental models" has a long and varied history and has been described differently by various authors. It has also been used by different academic disciplines in various ways, sometimes to indicate the entire range of mental representations and cognitive processes, and sometimes to indicate a specific subset of mental phenomena (e.g., images, assumptions, generalizations, perceptions). The present work adopts a definition of "mental models of dynamic systems" suggested by Doyle and Ford [1998] for use in research on complex dynamic systems:

> "A mental model of a dynamic system is a relatively enduring and accessible, but limited, internal conceptual representation of an external system whose structure maintains the perceived structure of that system."

Thus, for the purposes of this chapter the term mental model indicates a particular type of mental representation that is an attempt to mentally replicate the structure and relationships of a complex dynamic system. It is proposed that such mental models are developed naturally through experience with systems and that they play an important role in guiding dynamic decision making [Doyle et al. 2002]. Further, it is assumed that people have conscious access to these mental models and can articulate them through a knowledge elicitation process. However, it is important to distinguish between the output of such an elicitation process and the mental models themselves, which exist only in people's minds and cannot be directly observed. The phrase "measuring change in mental models" in the title of this chapter was carefully chosen to recognize the difficulty in externalizing mental models with any degree of certainty. You can never be sure that the external representation, whether in verbal or diagrammatic form, is a complete and accurate representation of the internal mental model on which it is based. In fact, the external representation obtained is likely to be highly dependent on the particular knowledge elicitation procedure that was employed, and the process of measurement is subject to error. But it is assumed that a reasonably accurate measurement of *change* in a mental model can be obtained so long as identical procedures are used each time knowledge is elicited and the procedures are valid and unbiased.

Although different researchers may disagree about how to define mental models, there is substantial agreement on their characteristic shortcomings. Many researchers conclude that mental models are often oversimplified and ill-defined, typically fail to adequately incorporate important features (e.g., feedback mechanisms, nonlinear relationships, and time delays) of the real system, and are subject to various forms of bias and error [Doyle and Ford 1998].

Acknowledging these shortcomings of mental models and the important role they play in dynamic decision making, the system dynamics computer modeling tradition founded by Jay Forrester [1961] (see Sterman [2000] for an overview of the field) adopts the goal of improving mental models as one of its primary aims. It is proposed that the act of building a system dynamics simulation model (or learning by interacting with a model built by someone else) can help decision makers overcome cognitive limitations as well as various external barriers to learning in the real world (e.g., time delays, irreversible actions), allowing mental models to become more dynamically complex. A wide variety of system dynamics and systems thinking interventions have been developed and implemented over the years, many with the explicitly stated goal of changing the mental models of participants to make them more complete, complex, and dynamic in order to improve their ability to manage a complex system [e.g., Vennix 1990, 1996; Cavaleri and Sterman 1997; Huz et al., 1997].

Of course, in order to judge the effectiveness of these interventions in promoting learning, participants' mental models must be elicited, organized, represented externally, and compared before and after the intervention, whether formally or informally. Toward this and other ends system dynamics and systems thinking researchers typically apply one or more of a wide variety of formal techniques for organizing and representing mental model information, including system flow diagrams [Forrester 1961; Morecroft 1982], causal loop diagrams [Richardson and Pugh 1981], various forms of influence diagrams [Axelrod 1976; Coyle 1977; Eden and Jones 1984], hexagons [Hodgson 1992], and social fabric matrices [Gill 1996]. Researchers have also developed several distinct sets of procedures and methods for eliciting or mapping mental models, typically during facilitated group sessions, including the Strategic Options Workshop [Eden and Huxham 1988], the Strategic Forum [Richmond 1987], the corporate system modeling policy session [Roos and Hall 1980], and the group model building approach described by Vennix [1996].

However, these established techniques were not originally designed primarily to measure mental models but to facilitate change and improvement in mental models. In fact, the very features that make them valuable for changing mental models (the introduction of new, systematic ways of thinking about mental models, the assistance and direction provided by the facilitator, and the consensus achieved during group processes) simultaneously make them unsuitable for measuring that change in an accurate and unbiased way. For example, the introduction of new and unfamiliar ways of thinking about mental models may cause participants to change their mental models during the elicitation procedure, masking their pre-intervention content and structure. And, the involvement of the facilitator and other group members during elicitation procedures introduces a host of potential ways in which the mental models of others can interfere with the elicitation of the mental model of any particular individual.

These criticisms by no means imply that systems interventions based on existing methodologies for eliciting and representing mental models are not effective in promoting learning; it simply means that their effectiveness cannot be demonstrated beyond a reasonable doubt by current practice. The resulting inability to judge the relative effectiveness of different interventions with a high degree of confidence likely inhibits the ability of individual researchers to learn from experience and the ability of the research field as a whole to learn by comparing the experiences of different research teams. (Indeed, the very existence and use of so many different methods and procedures for eliciting, representing, and changing mental models of systems suggests that the difficult work of documenting their comparative advantages and disadvantages has yet to be done.)

The goal of accurate, unbiased measurement of changes in mental models of complex dynamic systems can only be fulfilled by a program of controlled, rigorous experimental research designed to supplement and support more realistic field studies [Doyle 1997]. Such an effort will require the development and testing of new methods and procedures that emphasize accuracy of measurement of mental models rather than methods and procedures that emphasize facilitation and improvement of mental models and that can be adapted to test multiple hypotheses related to the relative effectiveness of alternate intervention protocols. The purpose of the present paper is to define the necessary features of <u>any</u> methodology that aims to rigorously measure change in mental models of complex dynamic systems (see also Doyle et al. [2002]); to describe the development and implementation of one specific new methodology designed to fulfill these criteria; and to present the results of an exploratory application of the method to measuring changes in mental models due to a system dynamics intervention based upon the simulation game of the economic long wave, or Kondratiev cycle, developed by Sterman and Meadows [1985].

1.2. Experimental Design and Procedure for Measuring Change in Mental Models

The most appropriate and accurate techniques for measuring change in mental models of complex dynamic systems have yet to be established by the research literature [Vennix 1990], and there is a demonstrated need for research programs that will "make more precise and less artful the process of eliciting and mapping knowledge" [Richardson et al. 1989, p. 355]. Thus, it is not yet possible to prescribe the use of specific measurement instruments or protocols, and researchers should be encouraged to conduct exploratory work that tests alternate measurement techniques drawn from different literatures and research traditions. However, the more general requirements for the design and conduct of rigorous research on mental models and learning are well known in the psychology and education literatures and are largely agreed upon. Although there is room for researchers to exercise choice in how to operationalize these requirements, we believe that any method for measuring change in mental models that aims to be rigorous must strive to achieve at least the following nine goals:

1. *Attain a high degree of experimental control*. In designing any study of human cognition or behavior, choices must be made that affect the degree to which the study emphasizes experimental control (the ability to hold variables other than the one under examination constant) and external validity (the extent to which the observed results also apply to realistic settings outside the context of the study). Usually, but not always, a methodological choice that increases external validity decreases experimental control, and vice versa. For example, a study examining the effect of a system dynamics intervention that emphasizes realism might want to engage managers in a thorough, perhaps months-long examination of an important, real problem that affects the future of the company, in a setting that incorporates the incentives for performance, time pressures, and accountability of real business settings. In such a study, however, one would have great difficulty controlling important variables in the face of the other priorities of the company and the unpredictability of external events and would not be able to rule out alternate possible explanations for results. In a study where the emphasis is on experimental control, one might instead choose to engage a convenient sample of subjects (e.g., undergraduate students) randomly assigned to experimental conditions in a simplified, brief examination (perhaps lasting a day or a week) of a problem in a somewhat artificial setting devoid of the complications of realistic intervening variables. In this study, at the expense of raising questions about the applicability of its findings to realistic settings, the researcher would be in a much better position to unambiguously determine the causes of observed changes in thought and behavior. Of course, both types of study are valuable, important, and necessary. But rigorous studies that emphasize experimental control are particularly lacking in the systems thinking field and are unavoidable if questions about the ability of systems thinking interventions to change mental models are to be answered with a high degree of confidence.

2. *Separate measurement and improvement*. Any study that intends to assess the cognitive effects of a system dynamics intervention must attempt to both measure and improve mental models. However, it is important that these goals be separated; for example, if the first technique participants use to express their mental models is one that is thought to increase the degree of organization in mental models or to encourage completeness, then the first mental models elicited will not represent a true pre-intervention benchmark. Measurement and improvement of mental models should occur through distinct and separate procedures that take place during different experimental sessions.

3. *Collect data from individuals in isolation*. Group sessions coordinated by a facilitator are an important part of many systems interventions. However, there are several problems that make it difficult to accurately elicit the mental models of individuals in such settings. First, group discussions tend to be dominated by a few individuals and participants may fail to share ideas and opinions due to the effects of social loafing [Latane' et al. 1979], evaluation anxiety [Guerin 1986], or cognitive distractions [Baron 1986]. Second, in public forums people often comply with the views of others, while keeping their private opinions to themselves, in order to obtain rewards or avoid punishments [Kelman 1958]. Third, facilitators can inadvertently give participants clues about what ideas they believe to be better than others or lead

discussions in a direction they favor rather than the direction the participants would choose on their own, resulting in what psychologists call "experimenter bias" [Rosenthal 1966]. To avoid these problems, measurement procedures should be conducted in a setting that ensures confidentiality and effectively isolates participants from the influence of each other and any facilitators.

4. *Collect detailed data from the memory of each individual.* Mental models reside in the minds of individuals, and it is not possible to unerringly infer the contents of individual mental models without a detailed examination of the memory of each individual participant. For example, although an individual during an intervention may express agreement with statements made by other participants, or may indicate acceptance of a mental model representation developed by a group, it is possible that the relevant ideas are only held in memory in a temporary state. If so, the ideas may be forgotten or may be replaced by prior or subsequent information rather than become incorporated into mental models held in long-term memory. To control for this possibility, each individual participant should be asked to generate completely from memory their full mental model, in all of its messy, often fragmented detail, both before and after an intervention.

5. *Measure change rather than perceived change*. It is often tempting for researchers evaluating systems interventions to assess mental changes simply by having participants look inward and describe the effect the intervention has had on their mental models. However, there are serious problems with accepting such evidence at face value, including the possibility that participants may simply report what they think the researcher wants to hear, a phenomenon which psychologists call "subject bias" [Orne 1962]. Over-reliance on self-evaluation should be avoided: changes in mental models should instead be inferred by the researcher from a comparison of controlled pre- and post-intervention measures.

6. *Obtain quantitative measures of characteristics of mental models*. Efforts to improve systems thinking often fail to define the specific changes in mental models they hope to bring about. When this occurs, researchers are forced to judge the magnitude of observed changes in subjective, unsystematic, and possibly idiosyncratic ways. In addition, when dependent variables are defined post hoc, bias may result from focusing only on those measures that confirm expectations. To avoid these problems, researchers should explicitly define a priori such characteristics of mental models as detail complexity and dynamic complexity [see Senge 1990] and precisely how they will be quantified.

7. *Employ a naturalistic task and response format*. Research on human cognition suggests that memory and decision making are largely constructive processes. Which information is recalled from memory depends to a substantial degree on precisely how memory is measured [Roediger et al. 1989], whether the task being performed during retrieval is similar to the task performed during learning [Moscovitch and Craik 1976], whether external aids are used [Intons-Peterson and Newsome 1992], and subtle characteristics of the surrounding environment [Tulving 1983]. Similarly, the mental models and processes used in decision making are often highly variable

depending on task characteristics, goals, response modes, and even seemingly inconsequential changes in the way questions are ordered, worded, or framed [Kahneman and Tversky 1984; Hogarth 1982; Payne et al. 1992]. This implies that researchers that elicit mental models with a new or unfamiliar task run the risk of measuring different mental models than the ones participants use when free to follow their more typical habits of thinking, deciding, and problem solving. The degree to which an elicitation task encourages participants to think systematically or to exhaustively examine all of the relationships between relevant variables is also important and should be appropriate for the level of expertise of the participants. If, for example, the task encourages more effortful thinking than participants normally engage in, the study may end up measuring new, transient mental representations constructed during the elicitation procedure rather than the more durable mental models participants formed before the intervention. To avoid these problems, elicitation procedures designed for measurement rather than improvement should use tasks and response modes that approximate as closely as possible the way participants naturally go about representing and conveying their knowledge.

8. *Avoid bias*. Ideally, the elicitation procedures should be designed without making any a priori assumptions about the form or content of subjects' mental models. For example, asking how variable A is related to variable B may induce subjects to add variables and relationships that didn't exist in their mental model prior to answering the question. Asking subjects to identify feedback loops assumes that there are feedback loops in their mental models, which may or may not be true. Although the systems in which people operate are often complex and dynamic, their mental models of these systems may or may not be. The elicitation procedure should allow for the possibility that subjects' mental models, particularly novices, are relatively simple and static.

9. *Obtain sufficient statistical power*. The paradigm of controlled research on human subjects requires that sufficient numbers of participants be studied to allow hypotheses to be tested for statistical significance. Given that the magnitude of the effect of system dynamics interventions on various characteristics of mental models is not yet known, data should be collected independently from enough participants to allow the detection of even moderate to small changes in mental models.

1.3. Prior Research

The only prior research program conducted within the system dynamics community that meets all nine of the identified criteria for rigorous research on measuring change in mental models was described by Vennix [1990]; see also Verburg [1994]. In an ambitious, well-conceived, and thoroughly documented study, Vennix conducted a controlled experiment testing the effect of an intervention based upon a computer simulation of the Dutch social security system on several quantifiable features of mental models. One hundred and six college students participated in one of two sequences of experimental sessions involving a 40-hr. commitment over a 7-week period. Pre- and post-intervention mental models were elicited by asking subjects to prepare individually a two-page written policy note addressing a problem involving social security and the economy. These policy notes were subsequently coded into "cognitive maps," or directed graphs, following the procedures developed by Axelrod [1976]. Results showed that the intervention resulted in the following statistically reliable changes in mental models: an increase in the number of relationships that were quantified; an increase in the proportion of computer model concepts included (subjects' models, however, remained quite simple compared to the complexity of the computer model with which they interacted); an increase in the number of relationships between concepts; and an increase in the number of mentions of time delays. The intervention had no statistically reliable effect on the average length of paths or the number of feedback loops in the cognitive maps.

The present work applies the same general experimental approach to address a subset of the questions about the effect of a simulation-based intervention on mental models explored by Vennix [1990], and therefore can serve as a conceptual replication that may corroborate or qualify some of the conclusions of that work. However, to facilitate comparisons between the two studies, it is worth noting the following important differences:

1. Vennix [1990] tested the effects on mental models of interaction with an econometrics-based simulation. The present study tests the effects of interaction with a system dynamics-based model, which might be expected to better promote feedback thinking.

2. The intervention we tested is much briefer and simpler than the one tested by Vennix. While this limits the potential impact of the intervention and decreases external validity, it allows a higher degree of experimental control than longer, more complicated interventions. For example, unlike Vennix [1990], in the present work all data collection procedures were supervised and the time investment of subjects was kept constant. One of the goals of the present research is to develop a more practical, less time-consuming, yet rigorous methodology for measuring changes in mental models that will allow research results to accumulate more quickly. Toward this end we are interested in determining if statistical reliable changes in mental models can be obtained by a relatively brief intervention.

3. Both studies rely on a detailed content analysis of written documents that subjects are asked to produce in response to a set of questions. However, the present study borrowed its elicitation and coding procedures not from Axelrod [1976] but from a research tradition in cognitive psychology that holds that knowledge and beliefs are organized in memory in narrative or story-like structures that are variously termed narrative models [Bower and Morrow 1990], scripts [Schank and Abelson 1977], schemas [Fiske 1993], or, simply, stories [Schank 1990].[1] Research by Pennington [1981] and Pennington and Hastie [1986, 1988] has confirmed that these

[1] The assumption that the mental models of the participants in the present study will be based upon the specific, concrete events and relationships characteristic of stories rather than more abstract concepts and variables is consistent with the results of several studies that have reported the mental models of novices to be more representational and less abstract than those of experts [Larkin 1983; Chi et al. 1988].

story-like structures are spontaneously constructed and used to guide decision making when judgments are based on large amounts of interrelated information or experience that must be reviewed and organized. Thus the present study attempts to achieve naturalism by asking participants to convey their mental models the way they are typically conveyed in conversation: by creating a causal explanation or scenario that explains the available evidence.

4. In the Vennix [1990] experiment, subjects conveyed their mental models by writing a two-page essay that presumably took several hours, during which time subjects could refer to an introductory text on the topic at hand. Although this task is realistic and has the advantage of encouraging thoroughness, it is not clear whether the mental models derived from it represent durable models held in long-term memory. For example, a subject could include in the essay ideas and concepts he or she has read just minutes before but has not really learned and incorporated into mental models. The present study takes a different approach, giving subjects much less time to write a briefer essay without access to any reference materials, in order to ensure that the ideas being conveyed are coming from long-term memory.

5. In the Vennix [1990] experiment, subjects were asked to read a 25-page introductory text before pre-intervention mental models were measured. This allows for a much stricter test of the effectiveness of the intervention, as pre/post differences will reflect changes in mental models over and above any changes caused by reading the text. However, this also means that the more naive mental models subjects held before reading the text were not measured. Such naive mental models can be important, as several empirical studies have found that they tend to persist and influence behavior even after instruction in "correct" models [e.g., DiSessa 1982; Clement 1983; McCloskey 1983]. In the present study, we chose to compare post-intervention mental models with the naive mental models, based on life experience, that subjects held before engaging in any activities related to the experiment.

1.4. Method

1.4.1. Subjects

Sixty-four undergraduates enrolled in an introductory social science course at Worcester Polytechnic Institute participated in the experiment in order to fulfill a course requirement. Forty-six of the 64 students completed the experiment; the other 18 were dropped from the study due to failure to attend one or more of the experimental sessions or to participation in pretests of the experiment. The students were almost exclusively science and engineering majors taking the course to fulfill a breadth requirement and they had little or no prior exposure to economics, management, or system dynamics at the college level. Forty-six percent of the students were female. Thirty-eight percent of the students were seniors; 38% were juniors, 17% were sophomores, and 4% were freshmen. The students were assured that their responses would be completely confidential and that, although they were required to participate, their performance in the experiment would not affect their course grade in any way.

1.4.2. Design

Subjects participated in a system dynamics intervention structured around their experience playing STRATEGEM-2, a simulation game of the Kondratiev cycle, or economic long wave, developed by Sterman and Meadows [1985] and employed in experiments on dynamic decision making by Sterman [1989]. The purpose of the game is to help students and managers learn about and gain confidence in a simplified behavioral theory of the causes of long-term cycles of overexpansion and depression in the economy [Sterman 1985]. According to the theory, which focuses on the capital-producing sector of the economy, management investment decisions lead to overexpansion due to time delays in production and the reinforcing nature of capital self-ordering. This overexpansion leads to a subsequent depression as excess capital slowly depreciates over time. The goal of the intervention was to encourage participants to develop mental models that are more sophisticated in terms of both detail complexity and dynamic complexity [see Senge 1990] and that include important elements of the expert model, for example, the recognition that (1) the causes of the long wave are largely internal to the economic system rather than external; (2) it is the structural characteristics of the system (capital self-ordering, time delays) that largely determine the behavior of the system; and (3) periods of economic expansion and depression are causally connected.

Mental models of the causes of the economic long wave were measured by administering identical survey instruments before and after the intervention. [2]

---

[2] It should be noted that, as in Vennix [1990] and most studies of systems thinking interventions, the present study is limited by lack of access to a traditional control group. That is, there was no group of subjects studied concurrently who did not participate in the intervention. Therefore the possibility that any measured changes in mental models are due to events external to the experiment, although remote, cannot be completely ruled out.

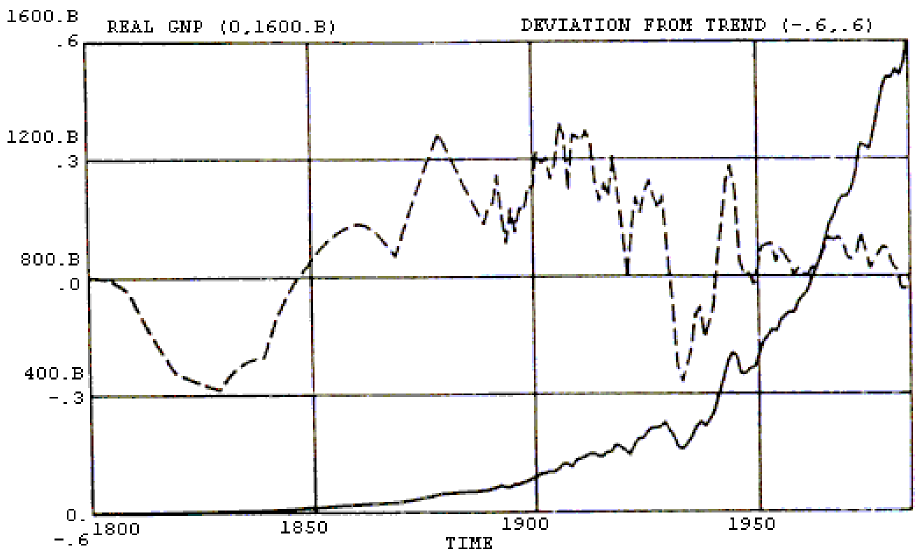

Figure. 1. Percentage deviation from the exponential growth trend of real GNP (adjusted for inflation) for the U. S. economy from 1800 to 1984 (1972 dollars). Reprinted from Sterman [1986].

However, two different experimental conditions were created: only half of the subjects were assigned, at random, to receive the preliminary survey in order to control for the possibility of pretest effects, that is, the possibility that the act of taking a pretest can itself change mental models and behavior independently of the effects of any intervention. For example, those subjects who take a pretest are made more cognizant of the fact that they are being studied and tested and may therefore exert more effort than others, increasing the effectiveness of the intervention. A more likely possibility is that subjects who express their mental models during a pretest may feel compelled to defend them later on and dismiss new ideas, decreasing the effectiveness of the intervention. This issue was chosen for experimental study due to its implications for future research. If significant pretest effects are found and confirmed, for example, then costly controls for pretest effects may have to become a standard part of methodologies designed to document change in mental models.

The mental models surveys began by introducing subjects to the concept of gross national product (GNP) and presenting reference mode data: a graph displaying deviations in the trend of real U. S. GNP from 1800 to 1984 (see Fig. 1). After receiving instruction in how to interpret the graph, subjects were told that some researchers had identified in these and other economic data a cyclical pattern of expansion (e.g., in the 1850s, 1900s, and 1940s) and depression (or recession) (e.g., in the 1830s, 1880s, and 1930s) that recurs about every 50 years. Subjects were then given the following prompt to elicit from them a causal narrative or "story" that would explain the data:

> Explain, in a paragraph or more, your best theory of the causes of the 50-year cyclic pattern in the GNP data. What do you think caused these changes in GNP? Use the space below to "tell the story" behind the pattern in the data, including important events, factors, and variables and the relationships between them. Rather than simply labeling the depressions and expansions, try to explain them by thinking back to the ultimate basic factors in the economy or society that caused them.

Subjects were given two additional prompts to elicit further information. They were asked, "What do you think caused the period of depression between the years 1929 and 1933?" and "What do you think caused the period of expansion between the years 1933 and 1944?" The prompts were kept simple and open-ended to avoid providing clues to subjects about which events or variables might be relevant and to discourage subjects from reaching beyond their knowledge to "guess" at how variables might be related. Within the allotted time, subjects could decide for themselves how much or how little to write in answer to the prompts. Each prompt was followed by a question asking subjects to indicate, on a 1 to 7 scale, how confident they were that their explanation was correct.

1.4.3. Procedure

The experiment was conducted during 5 separate class sessions spread over a two-week period. On the first day, half of the subjects were randomly selected to complete the mental models pretest. During this time the remaining subjects completed a different survey that was unrelated to the present experiment. Subjects were allotted 25 minutes to complete the surveys individually under strict supervision.

On the second day, subjects received verbal and written instructions, based on the recommendations of Sterman and Meadows [1985], covering the purpose and operation of the simulation game, which was implemented in Powersim. The instructions included definitions of all economic terms, a detailed presentation and explanation of the structure of the game (see Fig. 2), and explicit definitions of the player's goal and how performance would be measured.

Subjects, in small groups of 3 to 10, participated in a 1-hr. session in a microcomputer laboratory on the third day. As in Sterman [1989], the game began in equilibrium and there was a simple step-function change in exogenous orders from the consumer goods sector from 450 to 500 units; the object of the game was to respond to this external shock and return the system to equilibrium. Working individually, subjects completed one play of the simulation game (36 periods) and submitted their printed data. During the game subjects had access to the structure diagram; graphs showing changes over time in their orders, production capacity, and desired production; and information on changes over time in all variables in tabulated form. Two monitors were present at all times to answer questions about the structure and operation of the game The monitors also closely supervised the session to ensure that there was no communication between subjects and that subjects turned in the

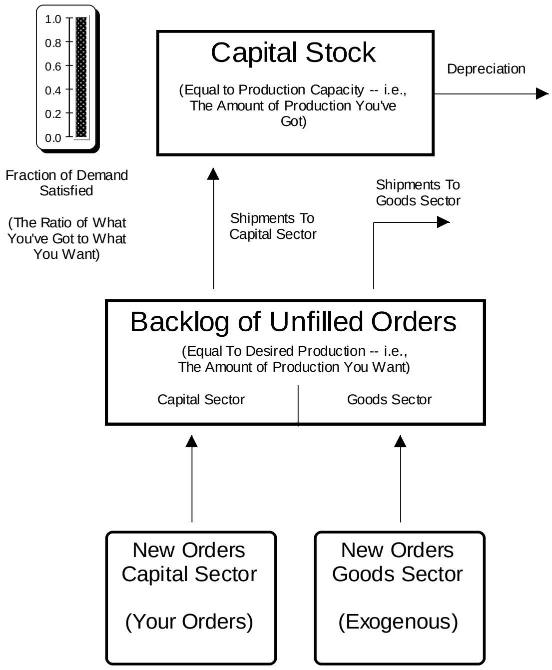

Figure. 2. Diagram of the structure of STRATEGEM-2 presented to subjects. Adapted from Sterman and Meadows [1985].

results from their first play of the game.[3] As previously reported by Sterman [1989], subjects' performance in the simulation game was quite poor compared with the optimum possible performance. The mean score was 1806 (SD = 1085),[4] compared to a mean of 591 (SD = 1,176) reported by Sterman [1989] and an optimal score of 19.[5] Eighty-one percent of the subjects generated an oscillatory wave pattern in production capacity, despite the simple step-function change in exogenous orders; only 17% of subjects were able to reestablish equilibrium before the game was over. The game thus achieved its goal for the great majority of subjects, namely, to illustrate that the long-wave pattern could result solely from decisions made by managers of the capital sector of the economy.

The fourth day consisted of a 50-minute debriefing session led by a facilitator that followed, with some modifications, the procedure outlined in Sterman and Meadows [1985]. Subjects were shown typical examples of results from their own experimental

---

[3] Pretests of the experiment determined that this level of supervision was, in fact, necessary. As first reported by Sterman [1989], in these pretests several subjects were so highly motivated to perform well that they attempted various forms of cheating.

[4] Four clear outliers (scores in excess of 20,000) were removed from this analysis.

[5] The difference in mean scores between the present study and the Sterman [1989] study is likely due to the differing degrees of subject matter expertise held by participants in the two studies.

sessions and were engaged in a group discussion concerning the thoughts and feelings they experienced while playing the game. The facilitator emphasized that poor scores on the game were not due to factors outside the players' control, since the structure and rules of the game and the state of the system were fully known. Subjects were asked to guess the pattern of orders from the consumer goods sector, and most suggested that there was probably a cyclic pattern in the exogenous orders. They were then shown the simple step-function that the simulation employed in order to emphasize the point that it was their own decisions that produced the cyclic wave pattern. At this point, the topic of the economic long wave was introduced and related to the simulation game, and subjects were shown examples of long-wave patterns in several different types of economic data. Summary results from the mental models pretest were presented to stimulate discussion about the group's pre-intervention mental models concerning long-term patterns in the economy. The facilitator presented data that contradicted some of the assumptions of subjects' pre-intervention mental models and asked the group to discuss the data. The facilitator then presented and explained, via causal loop diagrams, the simplified expert model of the economic long wave described by Sterman [1985], and led the group in a discussion of it. Finally, the facilitator closed the session by presenting an argument that the causes of the wave patterns in the simulation game were also a plausible explanation of similar patterns in the real economy.

On the fifth day the mental models post-test was administered to all of the subjects. This session was scheduled several days after the debriefing session in order to reduce recency effects (i.e., to reduce the chance of eliciting transient, unstable mental models). The post-test was administered in the same setting and employing the same procedures and time limits as the pretest.

1.4.4. Data Analysis

1.4.4.1. Content Analysis

The methodology employed in this study creates large amounts of verbal data which are often messy, incomplete, redundant, and idiosyncratic and which must be reduced, organized, and coded in a consistent and unbiased way. There are many different existing techniques for coding such data into diagrams or “cognitive maps” composed of concepts (or “nodes”) and the relationships between them (“links”). The present study adopted a simplified version of a “causal chain analysis” coding scheme developed by Pennington [1981] based on Schank’s [1972, 1975] conceptual dependency theory of causal relationships in natural language. In this type of analysis the nodes in a cognitive map of a mental model are not represented as abstract variables but as the concrete events that comprise stories or narratives. The nodes are organized temporally and connected with unidirectional, unsigned links to indicate that one event causes, enables, results in, or can lead to a second event. [6]

The coding process began by having an expert coder read through the entire data set to create a comprehensive list of all of the events described by subjects, resulting

---

[6] Unlike Pennington [1981], in order to simplify the analysis no distinctions were made between different types of events or links in the coding process.

in a list of 103 events.[7] The experimental data were then coded while blind to experimental condition. The resulting lists of linked events were compiled into causal scenario diagrams to facilitate the coding of quantitative variables.

1.4.4.2. Quantitative Variables

Several quantitative variables were created based on the characteristics of the causal scenario diagrams, following the recommendations of Vennix [1990] when applicable. The content of mental models was quantified by calculating the percentage of subjects in each experimental condition who included each event in their narrative. Pre/post differences in these percentages were then subjected to a chi-square analysis. Since one of the goals of most systems interventions is to move the participants toward a "shared understanding" or "shared mental model," a measure of the degree to which mental model content was shared by the participants was created by computing the average percentage of subjects who included the most often-mentioned events in their narratives. This measure indicates the extent to which subjects are converging on a small number of important events versus holding competing mental models that include widely divergent events.

According to Senge [1990], mental models can exhibit two different kinds of complexity: detail complexity and dynamic complexity. Detail complexity is simply the amount of content, for example, the number of nodes and links. In contrast, dynamic complexity indicates the presence of feedback thinking and an understanding of other important system dynamics concepts (e.g., that the causes of events are often remote in time and space from their effects). In this study detail complexity was assessed by counting the number of events and links in subjects' causal scenario diagrams. In addition, the number of links per event was calculated to control for the fact that an increase in the number of events can increase the number of links without increasing the extent to which the diagram is interconnected. As an additional measure of detail complexity, the average length of causal paths extending back from the primary to-be-explained events (changes in GNP and the occurrence of economic expansions or depressions) was calculated. Since subjects varied widely in the number of events they included, the average causal path length was divided by the maximum possible length for each subject (the number of events in the diagram – 1).

Three variables relevant to dynamic complexity were created. First, the number of feedback loops contained in each diagram was counted as an indicator of the degree of "feedback thinking." This number was divided by the number of events in the diagram and also by the average length of the feedback loops (since as the length of a feedback loop increases the chance that additional loops will be created by adding small variations to the single loop increases). Second, the percentage of subjects who described a causal link between economic depressions and expansions was noted for each experimental condition. Third, as a measure of the "remoteness in time and space" of initiating causes of events, the maximum causal path length extending back

[7] This approach, in which the coding categories are created from the experimental data itself rather than an independent source, is exploratory rather than confirmatory [see Carley and Palmquist 1992]. Such an approach is necessary when, as in the present case, the full set of concepts used by subjects cannot be predicted a priori.

from the primary to-be-explained events, divided by the maximum possible length, was calculated for each subject.

For all of the continuous variables, it is possible to ask if an increase in the variable is due to subjects simply writing longer narratives rather than including more information in the same number of words. To control for this possibility, all of the continuous variables were divided by the number of words in the narratives from which they were drawn. [8]

1.5. Results and Discussion

1.5.1. General Characteristics of the Causal Scenario Diagrams

As suggested by Norman [1983], the causal scenario diagrams reported in this study, both pre and post, indicate that the subjects indeed held mental models that were “messy, sloppy, incomplete, and indistinct.” The diagrams were relatively modest in both size (containing 11 events and 9 links, on average) and complexity (the mean average path length back from the primary to-be-explained events was 1.5, and the mean maximum path length was 3). In addition, evidence of feedback thinking was rare overall. For example, the average number of feedback loops in each diagram was only 0.5). And, most subjects did not even consider the possibility that expansions and depressions could act as causal agents, as is evident from the high number of links pointing toward these events compared with almost no links pointing away from them.

The diagrams were also highly variable across subjects in both size (the smallest contained only 4 events, while the largest contained 20) and content (subjects described a total of 103 unique codable events). One way of reducing this bewildering variety so that similarities between subjects can be more easily perceived is to create from the set of individual diagrams a “composite” diagram that includes only the events and links mentioned by a substantial number of subjects. This is done in Figs. 3 and 4 for pretest and posttest data, respectively.[9] The most obvious feature of these diagrams, both pre and post, is how greatly simplified they are compared with expert explanations of economic systems and the long wave. For example, the detailed chain of events by which technological innovation leads to economic growth is reduced in both diagrams to a single link. The diagrams are also very nonspecific: when events from the expert theory of the long wave are included, they are not

---

[8] Actually, post-test narratives proved to be reliably shorter, on average, than pretest narratives ($t = -2.66$, $p < .05$, mean number of words 158 vs. 180).

[9] The causal scenario diagrams in Figs. 3 and 4 are closely related to but distinct from causal loop diagrams. The main difference is that increases and decreases in variables are treated as separate “events” and therefore the “sign” of relationships is incorporated into the nodes rather than the links, as they are in causal loop diagrams. Causal scenario diagrams can easily be converted into causal loop diagrams if desired. However, the assumption that subjects in this experiment are thinking at the level of abstraction represented by causal loop diagrams is not supported by the data.

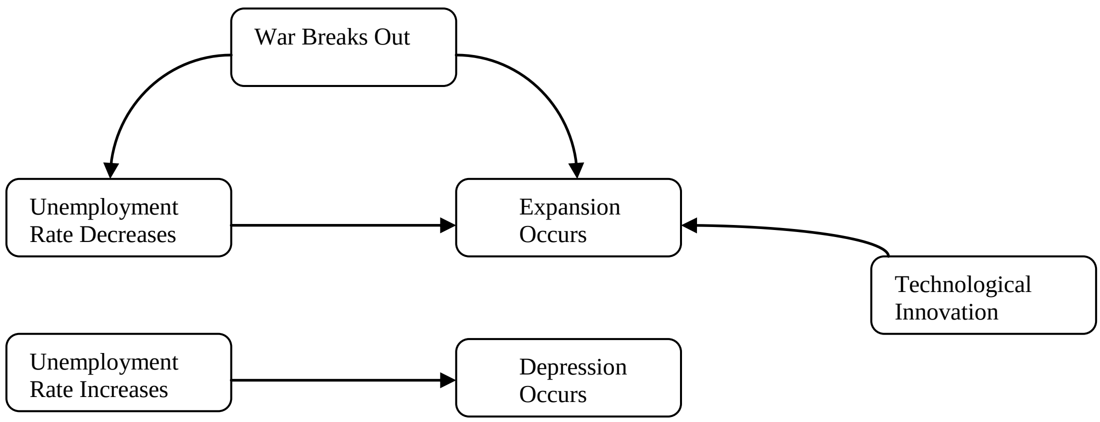

Figure. 3. Pretest composite causal scenario diagram containing events and links mentioned by at least one-third of subjects.

precisely correct. For example, when subjects mention "management decision errors," they do not typically specify that they take place in the capital sector of the economy – they just know that some manager, somewhere ordered too much of something.

Even though the composite diagrams are highly sanitized, they still show evidence of mental sloppiness. For example, in more than one case both simple and more complicated explanations of the same causal chain exist simultaneously (e.g., "war causes economic expansion" and "war reduces unemployment, which causes expansion"). And, on occasion, as in Fig. 4, "dead ends" are included, that is, events or chains of events are included if they are thought to be relevant, even if it is not known how they relate to the rest of the events. In summary, these diagrams, as should be expected given that the subjects typically had no formal training in the domains relevant to the intervention, reflect the exceedingly simple, occasionally confused thoughts of complete novices.

1.5.2. Pre/Post Differences in Causal Scenario Diagrams

1.5.2.1. Content of Mental Models

Tables I and II list the events contained in at least 19% of the pre- and post-test causal scenario diagrams, respectively. Table II also displays the χ^2 statistic and associated significance level for the pre/post difference in percentage of mentions for each post-test event. The most significant pre/post difference is the marked increase (from 5 to 43%) in the percentage of subjects mentioning the occurrence of a "poor management decision," a key element of the expert theory of the long wave. Two additional events important to the expert theory, "overproduction of goods (actually, capital)" and "time delays occur," also show a substantial increase after the intervention, although they are only marginally significant statistically. Thus there is reliable evidence that, post-

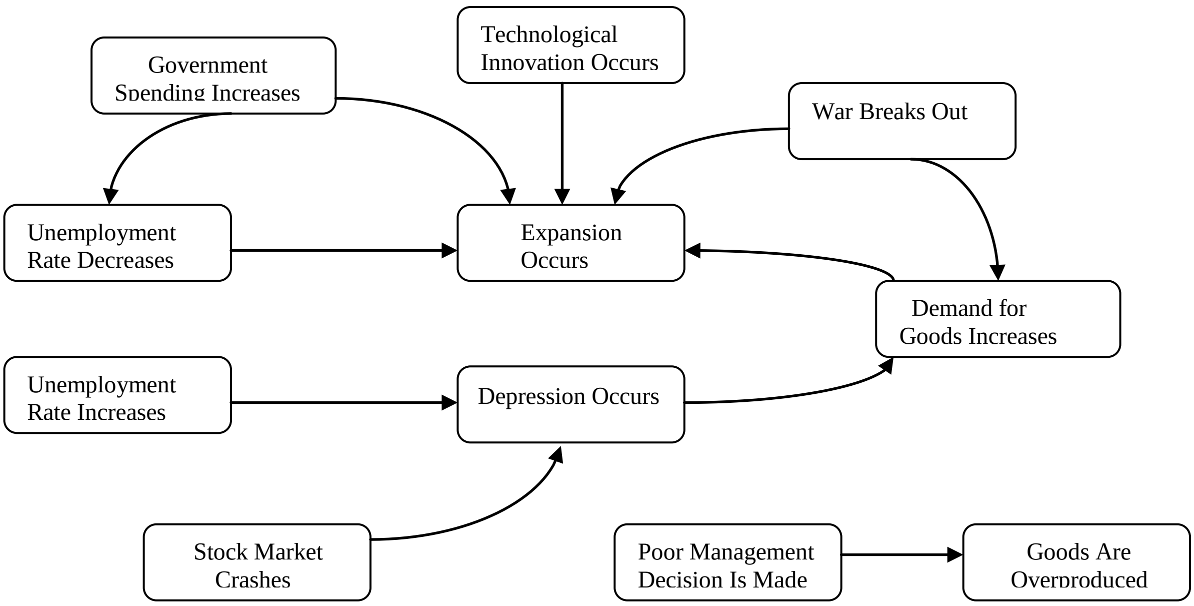

Figure. 4. Posttest composite causal scenario diagram containing events and links mentioned by at least one-third of subjects.

intervention, subjects are incorporating at least some of the expert concepts into their mental models.

However, this does not mean that these new events are replacing the events in the pretest models. Instead, subjects seem to have integrated the new events from the expert theory into their existing naïve mental models. This is apparent, for example, in the fact that there is no statistically reliable post-test change in the percentage of subjects mentioning many of the most common pretest events (e.g., “war breaks out,” unemployment rate increases”). In fact, for two events unrelated to the expert theory of the long wave, “desire to innovate increases” and “technological innovation occurs,” there is a statistically significant and a marginally significant increase, respectively. This suggests that during the intervention subjects are not only learning some elements of the expert theory but are also learning about and accepting elements of the naïve mental models of their fellow subjects.

Overall, the method seems to be capturing a transitional state as subjects move from novices toward a level of somewhat more expertise. The pretest causal scenario diagrams suggest mental models in which the primary causes of the long wave are thought to be external to the economy (e.g., war, technological innovation). The post-test diagrams include these elements essentially unchanged while also incorporating aspects of the expert model. This pattern of results likely represents a problem endemic to the enterprise of trying to change mental models: old mental models are

Table I

Most Often Mentioned Events in Pre-Test Causal Scenarios of the Economic Long Wave (N = 21)

Event	Percentage of Subjects
War breaks out	81
Technological innovation occurs	52
Unemployment rate decreases	43
Unemployment rate increases	33
Consumer spending increases	29
Consumer spending decreases	29
Government spending increases	24
Demand for goods increases	24
Consumer morale increases	24
Consumer morale decreases	19
Stock market crashes	19
Savings are depleted	19
Amount of trade increases	19
Incomes decline	19
War ends	19

not easily gotten rid of, even after new mental models have been learned and accepted.

Although some aspects of the content of the diagrams changed due to the intervention, this did not result in subjects converging on a "shared mental model:" there was no reliable pre/post difference in the average percentage of subjects who mentioned either the top 5 (χ^2 (df = 1) = .38, n.s.) or top 15 events (χ^2 (df = 1) = .30, n.s.). This is not particularly surprising since the intervention did not incorporate group consensus building elements.

1.5.2.2. Complexity of Mental Models

Detail Complexity. Post-intervention causal scenario diagrams contained more events ($t = 2.92$, $p < .01$) and links ($t = 3.74$, $p < .002$) than pre-intervention diagrams. It should be noted, however, that this increase in detail complexity, although statistically reliable, was relatively modest (the mean number of events per 100 words of text increased from 6.2 to 8.1 and the mean number of links per 100 words increased from 4.7 to 6.9). The pre/post difference in the number of links per event, however, was only marginally significant (means .49 vs. .59, $t = 1.72$, $p = .10$). This suggests that,

Table II

Most Often Mentioned Events in Post-Test Causal Scenarios of the Economic Long Wave (N = 25)

Event	Percentage of Subjects	χ^2 (df = 1) test for pre/post difference
Technological innovation occurs	76	3.60 ($p < .10$)
War breaks out	72	.52 (n.s.)
Demand for goods increases	52	1.62 (n.s.)
Poor management decision is made	43	18.0 ($p < .001$)
Unemployment rate increases	43	.40 (n.s.)
Unemployment rate decreases	38	.10 (n.s.)
Government spending increases	38	1.02 (n.s.)
Goods are overproduced	38	3.08 ($p < .10$)
Stock market crashes	38	1.88 (n.s.)
Desire to innovate increases	29	4.28 ($p < .05$)
Time delays occur	24	3.10 ($p < .10$)
Capital depreciates	19	2.04 (n.s.)
Consumer spending increases	19	.53 (n.s.)
Consumer spending decreases	19	.53 (n.s.)

although subjects' mental models increased in size due to the intervention, they did not become substantially more intricate or interconnected. This conclusion is supported by the finding that there was no reliable pre/post difference in the average path length back from the primary to-be-explained events, divided by the maximum possible length ($t = -.49$, n. s.)

Dynamic Complexity. Post-intervention causal scenario diagrams contained significantly more feedback loops than pre-intervention diagrams ($t = 2.45$, $p < .05$, mean number of loops/number of events/average length of loops/100 words .013 vs. .045). This result is particularly noteworthy because subjects received no training in describing, identifying, or constructing feedback loops. In addition, subjects were not asked to think diagrammatically; instead, the loops were implicit in the verbal narratives they were asked to write. Further evidence of an increase in feedback thinking is apparent in a statistically reliable increase in the number of subjects who described a causal link or chain connecting economic expansions and depressions, a key element of the expert theory of the long wave (χ^2 (df = 1) = 7.48, $p < .01$). However, there was no reliable evidence of a change due to the intervention in the second component of dynamic complexity, the "remoteness in time and space" of initiating causes of events: the pre/post difference in the maximum causal path length extending back from the primary to-be-explained events (divided by the maximum possible length) was not statistically significant ($t = .70$, n.s.).

1.5.2.3. Confidence

Few interventions attempt to measure the degree of confidence participants have in their mental models. However, people are remarkable explanatory creatures and are often quite willing to construct plausible-sounding explanations on the spot that they don't necessarily hold much confidence in. This leaves open the possibility that any measured changes in mental models due to an intervention may be illusory, since confidence may not have increased. To rule out this possible interpretation of results, the present study included both pre- and post-measures of participants' confidence in their mental models. After each prompt for verbal data, subjects were asked to indicate on a 1 to 7 scale, where 1 was not at all confident and 7 was extremely confident, how confident they were that their explanation was correct. The results indicate that subjects were reliably more confident in their explanations of the long wave after the intervention than they were prior to the intervention ($t = 2.55$, $p < .05$, means 3.5 vs. 4.1), although the average level of confidence remained near the middle of the scale.

1.5.2.4. Pretest Effects

Chi-square analyses were conducted to determine if the 25 subjects who did not participate in the mental models pretest were more or less likely to include in their narratives the top 15 events listed in Table II. For thirteen of these events the percentage of subjects did not reliably differ between the two groups, and were often quite similar. However, there were two statistically reliable differences: subjects who took the pretest were significantly more likely to include the events “desire to innovate increases” (χ^2 (df = 1) = 5.3, $p < .05$, percentages 29 versus 4) and “technological innovation occurs” (χ^2 (df = 1) = 17.0, $p < .001$, 76 versus 16) than subjects who took only the post-test. In addition, analysis of variance tests were conducted to determine the effects of participation in the mental models pretest on the quantitative variables related to detail complexity, dynamic complexity, and confidence described above. Two marginally significant effects were found: both the average ($t = 1.89$, $p < .07$, mean avg. path length/maximum possible length/ 100 words of text .14 vs. .24) and maximum ($t = 1.65$, $p = .11$, mean maximum path length/maximum possible length/100 words of text .34 vs. .22) length of paths extending back from the primary to-be-explained events were longer for subjects who did not participate in the mental models pretest.

Thus, while the pretest had no effect on the majority of the variables related to change in mental models, there were a small number of significant or marginally significant effects. Apparently, the mere act of answering the pretest survey made subjects much more likely to include at least two of the popular pretest variables on the post-test and also somewhat more likely to write narratives that were similar to their pretest narratives in terms of detail and dynamic complexity. This tendency for post-test models to be more similar to pretest models than they would have been if no pretest had been conducted represents another example of how the goals of measurement and promoting learning and change can be in conflict. While pretests are unavoidable for rigorous assessments to be conducted they may at the same time reduce the effectiveness of the interventions they are designed to assess.

1.5.2.5. Moderating Variables

Several moderating variables relating to subjects' experience playing the Kondratiev game and to subjects' general background characteristics were examined as possible predictors of how much their mental models changed due to the intervention. These variables included subjects' Kondratiev game scores (log-transformed since they varied over more than 2 orders of magnitude) as well as the timeshape of production capacity generated during the game.[10] In addition, subjects filled out a post-intervention survey in which they were asked to indicate, on a 1 to 7 scale, how much they enjoyed playing the game, how hard they tried to get a good score, and how carefully they thought about each decision before submitting it, as well as to report the number of economics classes taken prior to the experiment and self-rated computer skill. Finally, the average exam score in the class in which the experiment was conducted was included as a proxy variable for general academic ability.

These variables were included together in ordinary least squares (for continuous dependent variables) and logit (for categorical dependent variables) regression models.[11] The dependent variables included all of the quantitative variables related to detail and dynamic complexity described above as well as the content-related variables relevant to the expert theory of the long wave. Several of the variables were significant or marginally significant predictors of pre/post differences in the number of events and links in subjects' causal scenario diagrams. The pre/post difference in number of events and links decreased as enjoyment of the game increased (events $t = -1.9$, $p < .10$; links $t = -2.3$, $p < .05$), as computer skill increased (events $t = -2.3$, $p < .05$; links $t = -3.2$, $p < .01$), and as exam score increased (events $t = -2.5$, $p < .05$; links $t = -3.2$, $p < .01$). Thus the causal scenarios of those students with more computer skill and who like computer games more were less likely to change in size as a result of the intervention, perhaps because these students had less interest or skill in writing the verbal narratives through which mental models were assessed. The models of those students with higher academic ability were also less likely to change in size, perhaps because these students had larger models to begin with. Exam scores were, however, related positively to increases in the average and maximum causal path length of links in the causal scenario diagrams (average $t = 2.4$, $p < .05$; maximum $t = 2.0$, $p < .10$). How hard students tried during the game was also positively related to increases in average and maximum path length (average $t = 2.8$, $p < .05$; maximum $t =$

[10] The timeshape variable was examined because it can be unrelated to the game score, it may be a better indicator of performance than the game score (i.e., did the subject bring production capacity and desired production back into equilibrium or not?), and it is a good indicator of whether subjects experienced the simulated long-wave the game was designed to induce.

[11] In these models the statistical tests are for partial regression coefficients: the test asks whether the given variable reliably explains a portion of the variation in the dependent variable after controlling for all the other variables included in the model. With covariation among the predictor variables, this can produce conservative conclusions about the importance of a variable.

3.4, $p < .01$), perhaps because effort in playing the game is correlated with effort during the debriefing sessions and post-test. None of the predictors were significantly related to changes in the number of feedback loops or to changes in the percentage of subjects including expert concepts in their diagrams. Finally, Kondratiev game score and timeshape did not reliably predict any of the measures of change in mental models, after controlling for the other predictors. This may mean that what happens during game play is less important for fostering change in mental models than what happens during the subsequent debriefing session.

1.6. Conclusion

Measuring change in the mental models of the participants in systems thinking and system dynamics interventions is unavoidable if the relative effectiveness of different interventions in promoting learning about complex dynamic systems is to be assessed. However, few efforts have been made to design and implement rigorous methods that emphasize measurement of mental models rather than improvement of mental models. As a step toward encouraging such efforts, this paper has described the general features of rigorous methods designed to measure change in mental models of complex dynamic systems and provided a detailed example of the design, implementation, and analysis of one such method.

The experimental results produced by this process have in the present case been generally encouraging about the effectiveness of system dynamics interventions in promoting learning. Although the intervention tested was quite modest, involving simply a single play of a management flight simulator and related preparatory and debriefing sessions comprising about 5 hours of time spread out over a two week period, several statistically significant changes in participants' mental models resulted, including: an increase in the size of subjects' mental models, a change in the content of mental models toward the expert model of the problem posed by the intervention, and an increase in the degree of feedback thinking contained in subjects' models. [12] Certainly it would be expected that longer interventions or interventions that focus on building systems thinking or system dynamics modeling skills would elicit even stronger changes in mental models.

However, the present results must be interpreted with a great deal of caution: the paradigm of controlled laboratory experimentation does not allow the luxury of drawing grand conclusions from a single experiment, but instead relies on systematic

[12] It is worth noting that these results are similar in several ways to the results reported by Vennix [1990], despite the important differences between the interventions and methods applied by the two studies. For example, both studies reported an increase in the number of expert concepts included in post-test cognitive maps, an increase in the number of links between concepts, an increase in the number of mentions of time delays, and no change in the average lengths of causal paths. One major difference between the results of the two studies is that the present work reported a statistically reliable increase in feedback thinking, whereas Vennix [1990] did not. This is most likely because the present work studied a system dynamics-based intervention, while the Vennix [1990] study examined an econometrics-based intervention.

replication that inches toward truth one small step at a time. In fact, the experiment reported herein has several important limitations that can only be addressed (and should be addressed) by future research. For example, the study demonstrated positive effects due to the intervention, but does not address the possibility that some other type of intervention, unrelated to system dynamics, might be even more effective. The study also does not address what aspects of the intervention are important for producing the observed results: for example, were the changes in mental models due primarily to subjects' experience with the management flight simulator or to the information presented during the debriefing session? In addition, the limited time frame covered by the intervention did not offer the opportunity to assess the stability of the measured changes in mental models: it is entirely possible that the gains achieved by the intervention disappeared or at least decayed in the weeks and months after the intervention. The time constraints on the experiment also did not allow for the study of the correlation between the measured mental models and decision behavior as represented, for example, in post-intervention plays of the management flight simulator: as described in Doyle [1997], it cannot simply be assumed that this correlation is strong and positive.

Given these limitations on interpreting the experimental results, the main contribution of this research lies not in resolving questions related to the effectiveness of systems thinking interventions, but in demonstrating how they can most appropriately be studied. The described method proved to be both practicable and capable of capturing and quantifying even subtle changes in mental models due to the intervention, and it can be adapted in a straightforward manner to resolve the above-stated questions as well as a wide variety of other questions related to the effectiveness of systems interventions in promoting learning. Finally, it is hoped that the present work, which documents the messiness and sloppiness characteristic of mental models, as well as the problems faced by those who would attempt to measure and change them (including the persistence in memory of old mental models in the face of new information and the existence of pretest effects) will lead to an increased appreciation of the high degree of difficulty inherent in studying mental models of complex dynamic systems.

References

Axelrod, R. (ed.), 1976, *The Structure of Decision: The Cognitive Maps of Political Elites*, Princeton Univ. Press. (Princeton, NJ).

Baron, R. S., 1986, Distraction-conflict theory: Progress and problems, in *Advances in Experimental Social Psychology*, edited by L. Berkowitz, Academic Press (Orlando, FL).

Bower, G., & Morrow, D., 1990, Mental models in narrative comprehension, *Science*, **247**(5), 44-48.

Brehmer, B., 1989, Feedback delays and control in complex dynamic systems, in *Computer Based Management of Complex Systems*, edited by P. Milling and E. Zahn, Springer-Verlag (Berlin).

Carley, K., & Palmquist, M., 1992, Extracting, representing, and analyzing mental models, *Social Forces*, **70**(3), 601-636.

Cavaleri, S., and Sterman, J. D. 1997, Towards evaluation of systems thinking interventions: A case study. *System Dynamics Review*, **13**(2), 171-186.
Chi, M. T. H., Glaser, R., & Farr, M. J. (eds.), 1988, *The Nature of Expertise*, Erlbaum (Hillsdale, NJ).
Clement, J., 1983, A conceptual model discussed by Galileo and used intuitively by physics students. In *Mental Models*, edited by D. Gentner and A. L. Stevens, Erlbaum (Hillsdale, NJ), pp. 325-340.
Coyle, R. G., 1977, *Management System Dynamics*, Wiley (London).
DiSessa, A. A., 1982, Unlearning Aristotelian physics: A study of knowledge-based learning, *Cognitive Science*, **6**, 37-75.
Dörner, D., 1996, *The Logic of Failure: Recognizing and Avoiding Error in Complex Situations*, Addison-Wesley (Reading, MA).
Doyle, J. K., 1997, The cognitive psychology of systems thinking, *System Dynamics Review*, **13**(3), 253-265.
Doyle, J. K., & Ford, D. N., 1998, Mental models concepts for system dynamics research, *System Dynamics Review*, **14**(1), 3-29.
Doyle, J. K., Ford, D. N., Radzicki, M. J., & Trees, W. S., 2002, Mental models of dynamic systems, in System Dynamics and Integrated Modeling, edited by Y. Barlas, from *Encyclopedia of Life Support Systems (EOLSS)*, developed under the auspices of the UNESCO, EOLSS Publishers, Oxford, UK (http://www.eolss.net)
Eden, C., & Huxham, C., 1988, Action-oriented management, *Journal of the Operational Research Society*, **39**(10), 889-899.
Eden, C., & Jones, S., 1984, Using repertory grids for problem construction, *Journal of the Operational Research Society*, **35**(9), 779-790.
Fiske, S. T., 1993, Social cognition and social perception, in *Annual Review of Psychology*, **44**, edited by L. Porter & M. R. Rosenzweig , pp. 155-194.
Forrester, J. W., 1961, *Industrial Dynamics*, MIT Press (Cambridge, MA).
Forrester, J. W., 1969, *Urban Dynamics*, Pegasus Communications (Waltham, MA).
Gill, R., 1996, An integrated social fabric matrix/system dynamics approach to policy analysis, *System Dynamics Review*, **12**(3), 167-181.
Goodman, M. R., 1974, *Study Notes in System Dynamics*, MIT Press (Cambridge, MA).
Guerin, B., 1986, Mere presence effects on human beings: A review, *Journal of Personality and Social Psychology*, **22**, 38-77.
Hodgson, A. M., 1992, Hexagons for systems thinking, *European Journal of Operational Research*, **59**, 220-230.
Hogarth, R. M. (ed.), 1982, *Question Framing and Response Consistency*, Jossey-Bass (San Francisco).
Huz, S., Andersen, D. F., Richardson, G. P., and Boothroyd, R. (1997). A framework for evaluating systems thinking interventions: An experimental approach to mental health systems change. *System Dynamics Review*, **13**(2), 149-169.
Intons-Peterson, M. J., & Newsome, G. L., III. , 1992, External memory aids: Effects and effectiveness, in *Memory Improvement: Implications for Memory Theory*, edited by D. Herrmann, H. Weingartner, A. Searleman, & C. McEvoy, Springer-Verlag (New York), pp. 101-121.
Kahneman, D. & Tversky, A., 1984, Choices, values, and frames, *American Psychologist*, **39**, 341-350.
Kelman, H. C., 1958, Compliance, identification, and internalization: Three processes of attitude change, *Journal of Conflict Resolution*, **2**, 51-60.
Kleinmutz, D., and Thomas, J., 1987, The value of action and inference in dynamic decision making, *Organizational Behavior and Human Decision Processes*, **39**(3), 341-364.
Larkin, J. H., 1983, The role of problem representation in physics, in *Mental Models*, edited by D. Gentner & A. L. Stevens, Erlbaum (Hillsdale, NJ). pp. 75-98.

Latane', B., Williams, K., & Harkins, S., 1979, Many hands make light the work: The causes and consequences of social loafing, *Journal of Personality and Social Psychology*, **37**, 822-832.

McCloskey, M., 1983, Naive theories of motion, in *Mental Models*, edited by D. Gentner and A. L. Stevens, Erlbaum (Hillsdale, NJ), pp. 299-324.

Morecroft, J. D. W., 1982, A critical review of diagramming tools for conceptualizing feedback structure, *Dynamica*, **8**, 20-29.

Moscovitch, M., & Craik, F. I. M., 1976, Depth of processing, retrieval cues, and uniqueness of encoding as factors in recall, *Journal of Verbal Learning and Verbal Behavior*, **15**, 447-458.

Moxnes, E., 1998, Not only the tragedy of the commons: Misperceptions of biofeedback. *Management Science*, **44**(9), 1234-1248.

Norman, D. A., 1983, Some observations on mental models, in *Mental Models*, edited by D. Gentner and A. L. Stevens, Erlbaum (Hillsdale, NJ), pp. 7-14.

Orne, M., 1962, On the social psychology of the psychology experiment, *American Psychologist*, **17**, 776-783.

Payne, J. W., Bettman, J. R., & Johnson, E. J., 1992, Behavioral decision research: A constructive processing perspective, *Annual Review of Psychology*, **43**, 87-131.

Pennington, N., 1981, Causal Reasoning and Decision Making: The Case of Juror Decisions, Ph. D. Thesis, Harvard University, Cambridge, MA.

Pennington, N., & Hastie, R., 1986, Evidence evaluation in complex decision making, *Journal of Personality and Social Psychology*, **51**, 242-258.

Pennington, N., & Hastie, R., 1988, Explanation-based decision making: Effects of memory structure on judgment, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**(3), 521-533.

Richardson, G. P., & Pugh, A., III, 1981, *Introduction to System Dynamics Modeling with DYNAMO*, MIT Press (Cambridge, MA).

Richardson, G. P., Vennix, J. A. M., Andersen, D. F., Rohrbaugh, J., & Wallace, W. A., 1989, Eliciting group knowledge for model building, in *Computer-Based Management of Complex Systems*, edited by P. M. Milling & E. O. Zahn, Springer-Verlag (Berlin).

Richmond, B., 1987, *The Strategic Forum: From Vision to Strategy to Operating Policies and Back Again*, High Performance Systems (Hanover, NH).

Roediger, H. L., III, Weldon, M. S., & Challis, B. H., 1989, Explaining dissociations between implicit and explicit measures of retention: A processing account, in *Varieties of Memory and Consciousness*, edited by H. L. Roediger, III & F. I. M. Craik, Erlbaum (Hillsdale, NJ), pp. 3-41.

Roos, L. L., & Hall, R. I., 1980, Influence diagrams and organizational power, *Administrative Science Quarterly*, **25**(1), 57-71.

Rosenthal, R., 1966, *Experimenter Effects in Behavioral Research*, Appleton-Century-Crofts (New York).

Schank, R. C., 1972, Conceptual dependency: A theory of natural language understanding, *Cognitive Psychology*, **3**, 552-631.

Schank, R. C., 1975, The structure of episodes in memory, in *Representation and Understanding: Studies in Cognitive Science*, edited by D. G. Bobrow and A. Collins, Academic Press (New York).

Schank, R. C., & Abelson, R. P., 1977, *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*, Erlbaum (Hillsdale, NJ).

Schank, R. C., 1990, *Tell Me a Story: A New Look at Real and Artificial Memory*, Macmillan (New York).

Senge, P., 1990, *The Fifth Discipline: The Art and Practice of the Learning Organization*, Doubleday (New York).

Sterman, J. D., 1985, A behavioral model of the economic long wave, *Journal of Economic Behavior and Organization*, **6**(1), 17-53.
Sterman, J. D., 1986, The economic long wave: Theory and evidence, *System Dynamics Review*, **2**(2), 87-125.
Sterman, J. D., 1989, Misperceptions of feedback in dynamic decision making, *Organizational Behavior and Human Decision Processes*, **43**, 301-335.
Sterman, J. D., 2000, *Business Dynamics: Systems Thinking and Modeling for a Complex World*, McGraw-Hill (Boston).
Sterman, J. D., & Meadows, D., 1985, Strategem-2: A microcomputer simulation game of the Kondratiev cycle, *Simulation and Games*, **16**(2), 174-202.
Tulving, E., 1983, *Elements of Episodic Memory*, Oxford Univ. Press (New York).
Vennix, J. A. M., 1990, *Mental Models and Computer Models: Design and Evaluation of a Computer-Based Learning Environment for Policy-Making*. CIP-Gegevens Koninklijke Bibliotheek (Den Haag).
Vennix, J. A. M., 1996, *Group Model Building: Facilitating Team Learning Using System Dynamics*, Wiley (New York).
Verburg, L., 1994, *Participative Policy Modeling Applied to the Health Care Insurance Industry*. CIP-Gegevens Koninklijke Bibliotheek (Den Haag).