National Institute for Science Education

University of Wisconsin-Ma&on · National Center for Improving Science Education

Research Monograph No. 8

The Modeling of Systems and Macro-Systemic Change: Lessons for Evaluation from Epidemiology and Ecology

James Ridgway

Funded by the
National Science Foundation

National Institute for Science Education (NISE) Publications

The NISE issues papers to facilitate the exchange of ideas among the research and development community in science, mathematics, engineering, and technology (SMET) education and leading reformers of SMET education as found in schools, universities, and professional organizations across the country. The NISE Occasional Papers provide comment and analysis on current issues in SMET education including SMET innovations and practices. The papers in the NISE Research Monograph series report findings of original research. The NISE Conference and Workshop Reports result from conferences, forums, and workshops sponsored by the NISE. In addition to these three publication series, the NISE publishes Briefs on a variety of SMET issues.

The research reported in this paper was supported by a cooperative agreement between the National Science Foundation and the University of Wisconsin-Madison (Cooperative Agreement No. RED-9452971). At UW-Madison, the National Institute for Science Education is housed in the Wisconsin Center for Education Research and is a collaborative effort of the College of Agricultural and Life Sciences, the School of Education, the College of Engineering, and the College of Letters and Science. The collaborative effort is also joined by the National Center for Improving Science Education, Washington, DC. Any opinions, findings, or conclusions are those of the author and do not necessarily reflect the view of the supporting agencies.

Research Monograph No. 8

The Modeling of Systems and Macro-Systemic Change: Lessons for Evaluation from Epidemiology and Ecology

James Ridgway

National Institute for Science Education
University of Wisconsin-Madison

August 1998

About the Author

James Ridgway is Reader in Psychological Aspects of Education in the Department of Psychology at the University of Lancaster, England, and a National Institute for Science Education (NISE) Fellow. He is director of the Center for the Study of Advanced Learning Technology (Lancaster University); Co-director of the National Science Foundation-funded Balanced Assessment project (based at Berkeley, Harvard, Michigan State, Nottingham and Lancaster) and the Mathematics Assessment Resource Service (based at Michigan State). He directs a 3-year project funded by the UK Economic and Social Research Council into the cognitive processes involved in task design for assessment and curriculum. His research interests center on the problems of educational change.

Acknowledgments

A number of people have contributed to the ideas in this monograph, including Cliff Adelman, Rose Asera, Constance Barsky, Steve Bauman, Charles Bruckerhoff, Hugh Burkhardt, Jeanne Century, Donald Chambers, Daryl Chubin, Bill Clune, Susan Cohen, LeAnne Dickey, Richard Elmore, Susan Gross, Pe.irce Hammond, Dan Heck, Joan Kenney, Mark Leach, Susan Millar, Mary Murray, Eric Osthoff, Marvin Patterson, Andy Porter, Senta Raizen, Alice Ridgway, Anne Ridgway, Tom Romberg, Ingrid Rosemeyer, Pat Rossman, Judah Schwartz, Patrick Shields, John Supovitz, Norman Webb, Paula White, Ken Wilson, John Witte, John Wright, and Andy Zucker.

I am grateful to each person for some stimulating discussions; the faults in the monograph are all my own.

Contents

Figures and Tables

Abstract

The research monograph begins with a discussion of the roles of science in the political world and, borrowing from Shakley and Wynne (1996), describes some different relationships which can exist between science and policy. It argues that education systemic reform (ESR) constitutes a novel approach to educational reform, about which little is known, and about which much is yet to be discovered. ESR requires "abracadabra" science, in the language of Shakley and Wynne. To be consistent with the philosophy of ESR, the evaluation of ESR must be an evaluation of systems undergoing change, and so evaluation itself also requires a good deal of abracadabra science.

The monograph describes three styles of modeling in science: analytic modeling, exemplified by eighteenth century physics; systems modeling, exemplified by biology; and macro-systemic modeling, exemplified by studies of ecologies undergoing change. Each modeling style depends on and incorporates its predecessor. The dominant intellectual traditions in education have been analytic, rather than systemic. The emergence of systemic reform as a paradigm for educational change has created a need for approaches to educational evaluation that set out to judge the functioning of systems; this will require attention to each of the major phases of evaluation-the evaluation of plans about some new system, the evaluation of the implementation of the plans, and the provision of summative feedback about its success or otherwise. However, systemic reform requires more than just an understanding of systems; rather, it requires an understanding of systems undergoing change. It follows that the evaluation of education systemic reform requires the evaluation o:f macro-systemic change. Several disciplines outside education have systemic and macro-systemic approaches as their dominant intellectual traditions. This monograph considers the approaches taken to evaluation and inquiry in some of these disciplines, notably epidemiology and ecology, and the central roles that evaluation plays in planning and monitoring change.

From a description of the methods used in other disciplines, a specification of the evidence base needed to conduct evaluations of ESR is derived. Attention is given to some of the research styles from a number of different academic disciplines (including physics and earth sciences) that face the same problems as those faced by education in terms of handling complexity. Some ideas on data gathering, modeling, and strategies for further research into educational evaluation are presented.

The monograph points out the importance of making full use of existing knowledge and the knowledge that the evaluation community is rapidly creating. It endorses arguments made by Wilson (1994) and Scriven (1993) that there is a pressing need for an intellectual community to emerge that addresses the issues of the management and evaluation of systems undergoing change.

Introduction

The National Science Foundation (NSF) has spent very large sums of money on education and has promoted a vision of systemic reform. The 1993 Government Performance and Results Act (GPRA; P.L. 103-62) decreed that, by 1999, all agencies show that they have measurable objectives and that these are being met. It follows that NSF must evaluate the whole program of education systemic reform (ESR), using objective measures, a task that may prove to be problematic. ESR is a complex business with long term goals; it is not at all clear how such an initiative might be evaluated.

ESR is a relatively new activity. The underlying assumption of ESR is that important elements in education do not have simple additive effects. Providing a better textbook will not necessarily improve student attainment unless teachers know how to use the new text; introducing a standards-based curriculum will not necessarily improve attainment unless teachers and the community understand and value the standards.

The theory that underpins ESR makes a number of assumptions:

- no education system should be viewed as a set of independent elements;
- all the elements of an education system (such as texts, teacher competencies, school-community relations, state policies, and the actions of different funding agencies) should be seen as interdependent elements of a "system";
- changing a single element is unlikely to result in changes in the performance of the overall system;
- if educational reform is to be effective, a concerted effort is needed which changes several elements in unison, and in such a way that their effects'are compatible and mutually supportive.

For any particular attempt at ESR (such as a specific statewide systemic initiative) to be successful, it is necessary to understand the elements in the current system and their interconnections. On the basis of this understanding, changes can be planned that take into account the mutual effects of interacting elements. Feedback will be necessary to see how well plans have been implemented and to monitor the successes of these plans. Some summative feedback will be necessary to judge the success of the whole initiative.

Evaluation is concerned with determining the value and worth of something. A useful distinction can be made between the evaluation of plans, formative evaluation, and summative evaluation (Stevens, Lawrenz, & Sharp, 1993). In contexts such as evaluating the impact of a new curriculum on student attainment, or a program to increase minority students' and women's enrollment in science, mathematics, and engineering courses, the evaluation community has a number of techniques that can be applied routinely. For example, an evaluator might start by eliciting the aims and objectives of the program, then derive relevant measures of performance, and then identify suitable benchmarks against which to judge the success of the new program. Although considerable intellectual effort is required to get the details right, the process itself is unproblematic, because evaluators have a considerable body of knowledge to draw upon.

The evaluation of ESR is a relatively new challenge. Some aspects of evaluation, such as summative evaluation, might be seen as unproblematic. A particular Systemic Initiative (SI) might be viewed as a "treatment" that can be compared with other treatments or with no treatment at all. However, summative evaluation of ESR is difficult for a number of reasons. Educational goals have changed to reflect new standards, and measuring the attainment of these new goals is difficult (Ridgway, 1998; Zawojewski, Hoover, & Ridgway, 1997). There are conceptual issues related to the attribution of the cause of changes that are detected; there are technical issues about the appropriateness of many research methods for determining change (Manski, 1995); and there are practical issues such as the time frame over which one might expect to see some change. The evaluation of plans and the provision of formative feedback in the context of ESR raises even more difficult problems than does summative evaluation (Heck & Webb, 1998; Webb, 199'7). They both take the evaluator into unknown territory. Rather little is known about how to evaluate plans for ESR or about how to describe systems that are undergoing change in such a way as to inform directors of ESR on possible effective courses of remedial action.

Education is not alone in facing "systems problems." A number of other disciplines tackle the problem of trying to change complex systems in particular ways. Medicine and ecology, for example, both deal with systems characterized by a large number of interacting variables that change over time, which are subject to outside influences, and where there are time lags between actions and observable effects. It makes sense to look to these disciplines to see how they conceptualize their subject matter, how they describe systems and systems undergoing change, and to consider the role evaluation has in managing complex systems.

The focus of this monograph is to explore the evidence bases, models and research styles used in a range of disciplines in order to inform the development of methods for evaluating education systems.

Science and Social Policy

Shakley and Wynne (1996) offer a fascinating account of the conflicts that scientists face when they enter the arena of public policy. The dilemmas arise because of a clash between two cultures. In the scientific world, it is legitimate to confess ignorance and to announce time lines measured in decades before knowledge will be available. In contrast, in the political domain, leaders are expected to solve problems or at least to be active in solving them within the span of their term of office. Any scientist who enters the public domain has to reconcile these conflicting positions.

Shakley and Wynne (1996) describe a number of relationships that can exist between scientists and policymakers and identify the following styles:

- the "monastery" model, where the scientific community is supported by the community around it and is expected to contribute to the spiritual well-being of everyone in the community (and not much else);

- the "attic" model, where the scientific store is sufficiently full so that a policy request can be addressed by getting groups of scientists to hunt around in what is already known to find a solution;
- the "gopher" model, where a (seemingly) researchable question is identified, relevant to policy, and scientists set off to find the answer;
- the "abracadabra" model, where the policy issue is so pressing that a new branch of science has to be invented.

Clearly, this simple classification system does violence to the spectrum of policy-oriented research activities that can be conducted; nevertheless, it makes some useful distinctions. It is interesting to speculate on the focus of some of the components of educational research within this framework. Any academic community can exist using the monastery model. The monastery model requires no engagement with practical problems, and so the issue of evaluation does not arise. The attic model works well when the task is to generalize from one well-understood situation to another. Educational researchers and teachers have a great wealth of craft skills that can be generalized from one situation to another situation that closely resembles it. For the attic model, the existence of an appropriate body of knowledge means that evaluation methods are likely to be available, or very easy to construct, because scientists are working from what is already well known. Education Systemic Reform (ESR) is a new venture for educators and is likely to be an example of either the gopher or the abracadabra model. Similarly, the field of evaluation of ESR is also likely to be an example of either the gopher or the abracadabra model.

It is important to distinguish between the evaluation of an individual systemic initiative (SI) and the evaluation of ESR. The evaluation of an SI will need to ask:

- Is the plan for the SI "'systemic"?
- Is the implementation of the SI "systemic"?
- Has the SI been a success?

The last question-which addresses the success of the SI—can be seen as gopher research. An SI can have well-defined goals, specified in terms of student outcomes, and evaluators can have a variety of methods to judge how well these goals have been met. Even so, there are conceptual problems in attributing changes in student performance to the activities of the SI, rather than to other causes. These problems are discussed at length by Manski (1995), but are beyond the scope of this monograph..

The answers to the first two questions depend on a view of systemic reform and so must be related to some theory of ESR. Theories differ in terms of the assumptions they make, the representations they use, and the sorts of evidence they consider. A theory of ESR might be cast in any one of a number of distinct theoretical frameworks. It is important for evaluators of SIs to be familiar with a variety of theoretical styles in order to contextualize particular approaches and to help frame appropriate questions for evaluation. It may well be the case that a particular SI has a theory of ESR that is quite inadequate and that will doom the SI to failure.

The evaluation of ESR as a whole is even more problematic. ESR is a theory of change. A thorough evaluation of ESR must evaluate the theory along with judging the reform. One might evaluate a theory by asking:

- What body of knowledge does it summarize?
- Is it internally consistent as a theory?
- What predictions does it make, and how do they stand up to tests?
- How useful is the theory in guiding practice?

To answer these questions, an evaluator must have a view of the key features of the body of knowledge, some critical flair, access to field test results, and a keen sense of whether or not different SIs have been implemented in systemic ways (clearly, one cannot judge the success of a theory if the implementation of that theory has failed), and the generative value of ESR.

Styles of Science: Evaluations of Systemic Reform Require an Appropriate Theory of Systemic Reform

Make things as simple as possible - but no simpler.
— Albert Einstein

Science is concerned with answering a few direct questions: about structure (what is there?); about function (how does it work?); and about evolution (how do things change over time?). Members of the scientific community share a number of common assumptions and approaches: evidence should be collected systematically; evidence should be reported in such a way that others can comment on the appropriateness of the methods used and (in many cases) can repeat the study themselves; theories should account for the available data.

Testing any theory — scientific or not — can also be guided by two simple principles:

- Is the theory internally consistent?
- Does it fit all the evidence?

A major goal of scientific activity is to tell a story about a range of phenomena in such a way that phenomena are neatly summarized, and in such a way that future events can be predicted, and in a way that provides a plausible explanation for what is happening.

The first stages of exploration are likely to pay attention to phenomena-What interesting things happen? What needs to be explored and perhaps explained?

A second stage is likely to explore effects-under what conditions do things occur? Under what conditions can certain things be made to occur?

The discovery of effects is simplified if:

- variables can be manipulated systematically (easy to do in school chemistry and physics, but hard to do in astronomy and anatomy);
- there are few interactions between variables, so that the effects of several variables acting together can be deduced by simply adding together the effects of each one acting alone.

A third stage is likely to combine studies of effects into *models of data.* With models of data, a large collection of results can be summarized by a few mathematical summary statements, such as the combined gas laws, Newton's equations of motion, or the Lotka-Volterra model of predator-prey relationships (e.g., Lotka, 1956).

A fourth stage involves the *creation* of theories-explanations of what is observed. These might claim the existence of objects or agents that are unobservable when the theory is created, such as electrons, viruses, or phlogiston.

Each of these stages produces things of real value to the individual and to the scientific community. Descriptions of phenomena can be a guide to practical behavior (e.g., when the ball game ends, traffic congestion is increased; Florida is warmer than Wisconsin in January). Studies of effects and modeling data allow one both to summarize large quantities of information which have been gathered, and to make predictions about future events either on the basis of direct past observation, or on the basis of interpolation or extrapolation from existing evidence. Theory building can lead to an understanding of phenomena, to predictions that go beyond simple extrapolation and interpolation using existing data, and to guiding scientific and practical endeavor.

It is useful to identify a number of distinct approaches to modeling that different scientific communities employ in their attempts to understand the world. The choice of model is colored by the phenomena of interest. It is also a function of scientific culture — scientific communities can be characterized by their topics of interest, by the range of research tools they use, and by the sorts of models they employ.

Analytic Modeling

Analytic modeling is quite familiar to the education community. Analytic modeling depends on experiment and quasi-experiment; some variables are controlled, others are manipulated (or observed at different levels when manipulation is impossible), and the effects on some variable of interest are noted. This approach is implicit in studies that depend on correlation or regression analysis (including recent techniques such as structural equation modeling). The approach is likely to be successful when:

- a small number of variables is involved;
- effects of positive and negative feedback are negligible; and
- effects of variables can be accumulated in straightforward ways.

The gas laws provide a good example of an analytic research style and of analytic modeling.

A range of *phenomena* concerning the expansion of gases was noticed.
Three *effects* were described after carefully controlled experimentation. For a fixed mass of an ideal gas:

> Boyle's law states that the volume is inversely proportional to the pressure, at a fixed temperature; i.e., PV = constant
> Charles' law states that the volume is proportional to the temperature (in degrees absolute), at a fixed pressure; i.e., V = T*constant
> The Pressure law states that the pressure is proportional to the temperature (in degrees absolute) at a fixed volume; i.e., P = T*constant

These three "laws" (actually, idealized generalizations from data) can be combined into a *model of data* — the ideal gas equation; i.e., PV = T*constant.
A *theoretical account* is offered in terms of the actions of molecules.

Challenges to analytic modeling. Consider the second law of thermodynamics. The second law states that entropy (i.e., uncertainty or randomness) increases over time. A classical example is a warm drink left in a cold room. At the start, the distribution of energy can be predicted; after a while, it cannot. The second law of thermodynamics applies to all physical systems, but is unhelpful in understanding the thermodynamics of living systems. For example, the entropy associated with a warm mouse in a cold room stays pretty much the same, over several hours. Living systems are characterized by an increase in structural complexity from conception to maturity, and (in many animals) by homeostatic feedback that maintains a relatively constant body temperature despite heat gain from and loss to the outside. Both of these characteristics contradict the second law of thermodynamics and show that its application is limited to physical systems.

Systems Modeling

The underlying mechanisms for defying chaos are a blueprint for action, appropriate resources in the environment, and a good deal of feedback. Models that use feedback and involve a large number of variables are called "Systems" models. Systems models are far more common in school biology than in school physics and are essential to understanding everyday problems studied by biologists (such as enhancing plant growth) in ways that are not essential to understanding everyday problems in physics (such as choices of bicycle gearing).

Systems approaches are useful in situations that involve feedback loops, and in situations where a large number of variables interact in nonlinear ways. Computer-based modeling commonly is used in systems approaches because the complexity of the interactions between elements means that predictions about future behavior in the model can only be made using computers, which bear the computational load. Examples are simulations of power stations, of the economy, and of world weather. The elements (e.g., the furnace, generator, valves, fuel supply, in the case of a power station) are relatively stable over time, but the state of the system can change a good deal. In terms of a theoretical account of any system, one needs to specify the elements of the system, the functional links among the elements, and the levels of particular resources.

Predator-prey relationships provide a simple example of a dynamic system. The phenomena are the large swings in sizes of the hare population and the lynx population, as observed in records of pelts kept by the Hudson Bay Company. The effects are cycles in the hare and lynx populations, which are out of phase with each other.

A dynamic systems model (created in STELLA II from High Performance Systems) is shown in Figure 1. This model specifies that the hare population is:

- increased by births (births = number of hares * hare natality)
- decreased by death (death = number of lynx * kills per lynx)

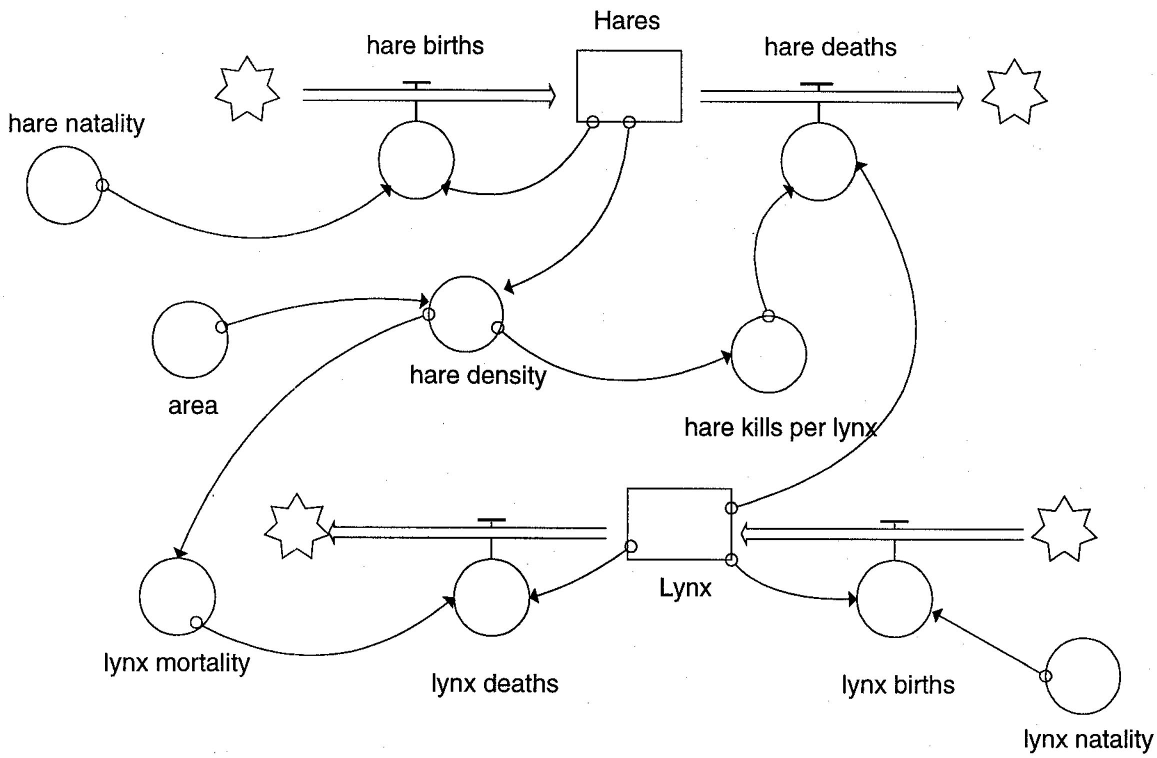

Figure 1. ***A systems model of predator–prey relationships***

The lynx population is:

- increased by births (births = number of lynx * lynx natality)
- decreased by deaths (deaths = number of lynx * lynx mortality)

(Note that "natality" and "mortality" are rates.)

The number of hares killed by each lynx is a positive linear function of the density of hares in the ecosystem—the more hares, the more are killed by each lynx.

The lynx mortality is a negative linear function of the density of hares in the ecosystem—fewer hares leads to a higher proportion of lynx deaths.

The model, when run on a computer, produces the cyclic fluctuation in hares and lynx shown in Figure 2. In this model, there are initially 50,000 hares and 1,250 lynx living in a 1,000-hectare ecosystem. The model shows marked cycles in the sizes of the two populations. It can be adjusted to explore the effects of changes in the parameters, or in the starting values of the numbers of hares and lynx. The model could be made more complex by adding more layers in the food chain, or by adding other predators and other prey.

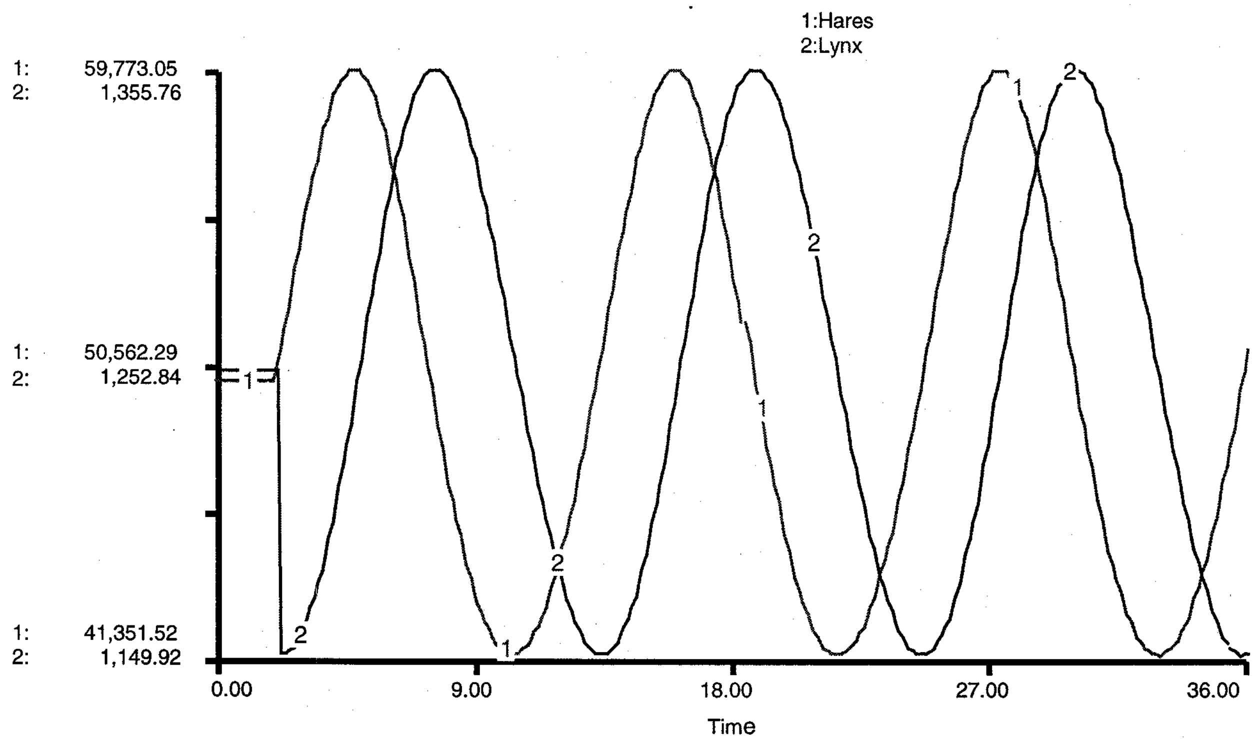

Figure 2. ***Cycles in the populations of predators and their prey***

After this extended example, it is appropriate to consider the uses and limitations of systems modeling. Systems modeling is essential when:

- a large number of variables are involved;
- effects of positive and negative feedback are significant.

Modeling is at the heart of all science, and systems modeling inherits all the problems of other sorts of modeling. These include:

- specifying elements;
- specifying the connections between elements;
- estimating parameters;
- specifying functional relationships between variables;
- fitting the predictions of the model to data;
- making and testing predictions for situations beyond ones where data are already available.

Challenges to systems modeling. Systems modeling assumes that the phenomena being studied, the effects, and the models of data stay stable over time. If the system is modified by introducing another element or a new relationship between existing elements, then a new model will have to be created.

In ESR, such changes can be a major goal. For example, an SI that adopts mentor teachers, or Web-based resources, is deliberately changing the system. In the case of Web-based resources, the education system might be considered to change each time a significant new element is added, such as integrating *The Why Files* (from NISE) into science lessons, or posting new forms of assessment for schools to download.

Systems modeling, then, is well suited to the depiction of stable systems, but not well suited to representing systems that are undergoing changes of the sort that characterize ESR. Theoretical models need to be developed that facilitate the description of systems undergoing change. For the purposes of this monograph, the depiction of systems undergoing radical change will be called macro-systemic modeling.

Macro-Systemic Modeling

The macro-systemic approach (e.g., Wilson, 1994) sets out to account for the evolution of systems. The macro-systemic approach accepts the complexities of modeling dynamic systems, and addresses the added challenge of describing ways in which systems themselves change over time in terms of the elements that are added to, or that become irrelevant in, the system, and in terms of the changes in the functional relationships between elements.

Consider the changes in the biosphere over the course of the history of the earth. In the initial stages, the planet cooled and condensed. The early atmosphere contained large amounts of carbon dioxide. Around 2500 million years ago, the level of oxygen began to rise (plausibly) as the result of oxygenic photosynthesis (the conversion of water and carbon dioxide to hydrocarbons and oxygen) by algae. Increased oxygen made it possible for other life forms to evolve, notably the invertebrates, then fish, amphibians, reptiles, birds, and mammals. Each stage set the scene for future development; however, the nature of that future development could not have been predicted from one stage to another.

Another example of macro-systemic development is provided by Wilson (1994), who describes the evolution of the air transportation industry. In 1903, the Wright brothers built an aircraft that flew about 100 yards. Less than 100 years later, there are systems in place that transport millions of people around the world. The transition from the first powered aircraft to modern transportation systems has not been an unrolling of a singlesystem; rather, it has been the creation and recreation of new systems. Each new system developed because the previous system set up conditions that allowed it to develop; in turn the new system makes future systems possible. Once a new system is in place, it makes a whole new set of systems possible, which in turn facilitate the emergence of other systems.

Challenges to macro-systemic modeling. Macro-systemic modeling approaches inherit all the problems of systems modeling. However, a macro-systemic model has to account for the emergence of new elements that arise in the system, the developments made possible from these new elements, the new relationships formed among the elements, and changes that occur in the fundamental nature of the system.

Table 1 summarizes the differences among the models.

Table 1
Characteristics of analytic, systems and macro-systemic models

Models	Elements	Relations between elements	Interactions	Stability of the system
Analytic	fixed	Fixed	Modest	stable
Systems	fixed	Fixed	extensive	relatively stable
Macro-systemic	changing	changing	extensive	unstable, evolving, or n-stable

Table 2 provides illustrations of analytic, systems, and macro-systemic modeling from the natural and human-engineered world. Macro-systemic modeling applies to the evolution of transport systems described earlier.

Table 2
Illustrations of analytic, systems, and macro-systemic models

	Model		
Illustrative example	Analytic (cf., classical physics)	Systems (cf., biology)	Macro-systemic (cf., evolution)
Airplanes	model for choosing a plane for the NY to LA route	model of the design of the 747	modeling the evolution of transport systems 1903 to 2003
Learning theory	B. F. Skinner's theory of conditioning	models of memory	Piagetian and Vygotskian theories of cognitive development
Epidemiology	statistical models used to analyze drug trials	models of the mechanisms of cholera transmission	modeling the evolution of public health, 1800 to 2000
Ecology	statistical models used to analyze field trials	models of predator–prey relations	models of ecology changes
Education systemic reform	statistical models used to analyze experiments and quasi-experiments	?	?

Airplanes. To illustrate the differences among the three types of modeling, consider airplanes. Analytic modeling applies to the "choice problems" of an airplane for a particular route. This modeling usually requires listing desirable features, rating objects (airplanes) on each, and applying some weighted combination of the ratings.

Systems modeling applies when considering the design of interacting component parts of an airplane. For example, increasing the number of passengers to be carried has obvious effects on the fuselage in terms of accomodating extra seats. It has slightly less obvious effects on the provision of "hotel" facilities such as food and restrooms, on safety provisions, and on luggage handling. Increased weight requires increased lift, which has implications for engine and wing design (which interact with each other). Computer models are used to map these interactions.

Learning theory

- *Analytic modeling.* Skinnerian conditioning (e.g., Skinner, 1953) was explored by careful control of situations and an exhaustive analysis of individual variables on learning outcomes.
- *Systems modeling.* Models of memory (e.g., Baddeley, 1976) often propose a number of discrete components such as sensory buffers, a short-term working memory store, and a long term memory store. Different components place limits on human functioning, such as the number of digits in an unfamiliar telephone number that can be remembered when dialing, and the rate of learning new information.
- *Macro-systemic modeling.* Piaget (1929) proposed a model where children go through a number of distinct stages in the same order; their rate of progress is a function of the environmental stimulation they receive, along with their genetic inheritance. The different stages reflect qualitatively different worldviews and so correspond to macro-systemic changes. Much of the work in the Piagetian tradition has focused on documenting these stages. In the terms used here, the work sets out to produce a macro-systemic account described in terms of the transitions between well-specified systems. For Vygotsky (198 1), the course of development is less like the unfolding of a flower in response to external and internal triggers; rather, its course of development is determined largely by the culture the child is brought up in. So the language and the intellectual tools of a culture such as its mathematics and science have a profound effect on the cognitive development that ensues. Vygotsky would argue that, unless one has studied human development in a particular culture in detail, one would not be able to predict the course of development that will take place.

Epidemiology

- *Analytic modeling.* The statistical analysis of data from drug trials models the data in terms of additive effects and their interactions (e.g., drug or no drug; young or old persons; high or low blood pressure; etc.).
- *Systems modeling.* Analysis of the transmission of cholera requires a model of interacting systems involving human waste, water systems, cholera itself, and public hygiene measures.

- Macro-systemic modeling. Analysis of public health changes might trace the introduction and impact of measures such as sewage collection and treatment, improved nutrition, the development of new drugs, and changes in medical provision on the health of a nation.

Ecology

- Analytic modeling. Field trials examine the conditions that facilitate the growth of certain plants via experimental plantings to explore the effects of shade, moisture, and soil in carefully controlled ways.
- *Systems modeling.* The predator-prey model described earlier provides an example.
- *Macro-systemic modeling.* Modeling changes in ecological systems that can result from changes in water provision, soil erosion, natural disaster, or human intervention all require macro-systemic accounts.

Education systemic reform

- *Analytic modeling.* Analytic models abound in education research. Describing the effects of teaching interventions or the introduction of new curricula are usually explored by controlled experiment, in order to determine the "effect size" of particular changes.
- *Systems modeling.* Informal systems models can be created simply by drawing "box and arrow" diagrams connecting elements of an educational system (teacher competence, initial teacher education, professional development, school resources, etc.). Zucker and Shields (1997) use an informal representation of major elements in education systems as the basis for describing the focus of work by individual SIs.
- *Macro-systemic modeling.* Macro-systemic models can be created by considering the evolution of educational systems over time. For example, the introduction of computer-supported learning in a school might begin with two enthusiastic geography teachers who use departmental funds to buy computers, sensors, and software and who rewrite the geography curriculum. It might evolve into a schoolwide system with laboratories and laptops, technical support, and cross-curricular planning to coordinate student learning of word processing, spreadsheets, and uses of the Web.

Although macro-systemic models relate to the evolution of systems models, the absolute time scales need not be long. The examples of human development (say over 10 years), engineered ecological changes such as the creation of gardens (say 2 to 200 years), and the introduction of computer-supported learning into a school (say 5 years) show that the time scale need not be great.

Table 2 shows that most scientific disciplines make use of each form of modeling, although the extent to which they use formal (e.g., computer based) models differs a great deal.

In education, there are few formal systems or macro-systemic models on which to base the planning and evaluation of systems reform. The next section considers an example of a systems model from epidemiology and an example of a macro-systemic model from ecology. The purpose of these examples is to show how they might be used in ESR and to describe the uses of such models for the purposes of evaluation.

Looking Outside Education for Evaluation Models

If I were you, I wouldn't start from here.
— Lewis Carroll

There is an extensive literature on educational evaluation. Why should one look outside this literature for new ideas? An argument can be made on a number of grounds.

First, the challenges presented by systemic reform are new, and one should not make the assumption that assessing a new kind of educational venture can be done by a simple extension of existing methods. This might be like applying the evaluation methods associated with preparing athletes for the 100 meter butterfly to a "new" Olympic event such as synchronized swimming. Many of the methods currently used to explore and evaluate issues in education are grounded in analytic approaches that dominate education and psychology; these methods and theories have evolved in a particular cultural setting, in response to a particular set of cultural pressures. The dominance of analytic methods in psychology can be illustrated by the uses of analysis of variance (ANOVA) over the past 60 years. ANOVA received a good deal of attention following the publication of Fisher's (1935) publication ***The Design*** *of Experiments*. By 1955 more than 80% of articles in four leading psychology journals used ANOVA and related methods for significance testing and the evaluation of hypotheses (Sterling as cited in Girgerenzer, 1992); by the early 1990s, Girgerenzer (1992) estimated that the figure was almost 100%. It can hardly be the case that almost all of the problems that psychology might address are best studied using investigative techniques of the sort suitable for analysis of variance. Systemic and macro-systemic models are highly relevant, but are rarely used.

A second reason to consider other disciplines is that many disciplines employ evaluation techniques when facing essentially the same problems as those faced in education. These problems include:

- exploring situations where there are a large number of interacting variables that change over time, both in terms of the variables that are relevant and in terms of their interrelations;
- making decisions about future practices that have profound effects upon human lives; and
- being accountable for these decisions in a very public way, and so needing not just a robust account, but also an account that can be communicated to nonexperts who are stakeholders.

It seems reasonable to believe that one might learn something about representing and evaluating complex evolving systems from intellectual domains such as medicine and biology, which have already addressed such matters via systemic and macro-systemic modeling with some success.

Characteristics of Educational Systems

Educational systems have a number of important characteristics:

- Educational systems involve a large number of interacting agents and agencies.
- The notion of one system is fundamentally flawed; subsystems differ in so many ways that each needs to be modeled separately.
- Educational systems are open systems; outside influences are important (e.g., political decisions at local, state, and government levels; community concerns, via the media; client concerns, via students, teachers, and employers).
- Educational systems are loosely coupled, unlike tightly coupled systems such as those found in the human body, or in a car, where changes in one element of the system (heart, lungs, tires, electronics) can have dramatic and immediate effects on the functionality of the whole system, so it is uncertain how changes in one part of a system will affect other parts.
- Agents in the educational system are self aware, so ideas themselves (and the act of evaluating) can transform the nature of the system and many of its properties.
- The system is subject to great time lags in terms of educational effects, so, for example, a decision is made to reform basic teacher education, it will take a great deal of time before the effects will be visible in the education system.
- There is no single “right” level of analysis; one can view each human as a self-contained system or as an element within a social group, or as a member of some broad community.

If one is to look for models that might guide the evaluation of ESR, it is important to find scientific domains that share the characteristics of education, yet which are more advanced in terms of developmental methods and conceptual models. Two domains have been chosen as exemplars here, namely disease control and ecology. Both share many of the characteristics of education (although the elements in neither system, diseases or plants, are self-aware). Both are domains where there is a great deal of human intervention in the system’s management, and this management is effective. These two domains will be used to illustrate different lessons for evaluation in education. A systemic model of the spread of disease is adapted from epidemiology to illustrate the creation of simple dynamic models. A more elaborate (and less well-specified) model is borrowed from ecology to illustrate macro-systemic modeling. In both cases, an attempt is made to show how each model might be transferred to education. Later sections of the paper offer an analysis of how educational evaluation might be conducted in the intellectual traditions of both systems modeling and macro-systemic modeling.

A Systems Model from Epidemiology: The SEIR Model

The key questions to be asked of any attempt to model a system are:

- What are the elements in the system?
- What are the interconnections?
- What are the functional relationships among different components of the system?

To build and validate even simple models, one needs a theory of the underlying processes, some reasonable estimation of the model parameters, and some evidence from realistic settings so that the model can be tested.

A great deal of effort has been devoted to modeling (and to preventing) epidemics. The SEIR (Susceptible, Exposed, Infectious, Recovered) model is often used as a generic starting point to model the likely transmission rates of specific diseases. The model is expressed as a set of three nonlinear ordinary differential equations, which are easy to simulate iteratively via computer.

Every population is composed of collections of individuals who are susceptible (that is, in certain circumstances, they can contract the disease), who are exposed (that is, are placed in circumstances where the contraction of the disease is likely), who are infectious (that is, when in certain kinds of contact with others who are susceptible, are likely to infect them) or who are recovered (that is, they have developed antibodies that render them immune to infection by the same disease). In simple cases, the history of infection for an individual runs through each stage in turn. Consider the simplest of epidemics such as a common cold in Wisconsin; some gross simplifying assumptions will be made, for didactic purposes. In this society, there are public meetings every day, and seating is allocated at random, subject to the constraint that people are not permitted to sit next to anyone they have been seated next to before. The epidemic starts with the arrival of Jim, who flies in from Britsville to a population that is entirely susceptible. Assume that Jim, and each subsequently infectious person, infects two other people each day; the number of infectious people in the population each day grows by 2, 4, 8, 16, 32, 64, 128, 256,512, 1024, etc., so that within 10 days there are over a thousand new infections each day, in 20 days a million, and in 30 days, a billion new infections.

The model needs to specify the recovery period (which dents the power function, above). It is common to assume (and often true) that recovered people are no longer infectious; and, of course, the population is finite.

An example of a computer simulation is shown in Figure 3 where it is assumed that the whole of the susceptible population is exposed (called the Non Infected Popul in the diagram) to some infection.

The time course of the disease is shown in Figure 4. It is characterized by little apparent influence of the disease in its early stages, then by a dramatic rise in the number of infected people, which declines as people recover,

Models written as programs have the virtue that all sorts of “what if?” conjectures can be explored by changing the parameters. What if there are more contacts per infected person? What if people are infectious even when they have recovered? What if there are subpopulations who behave differently (e.g., consider the transmission of AIDS in male and female homosexual communities)?

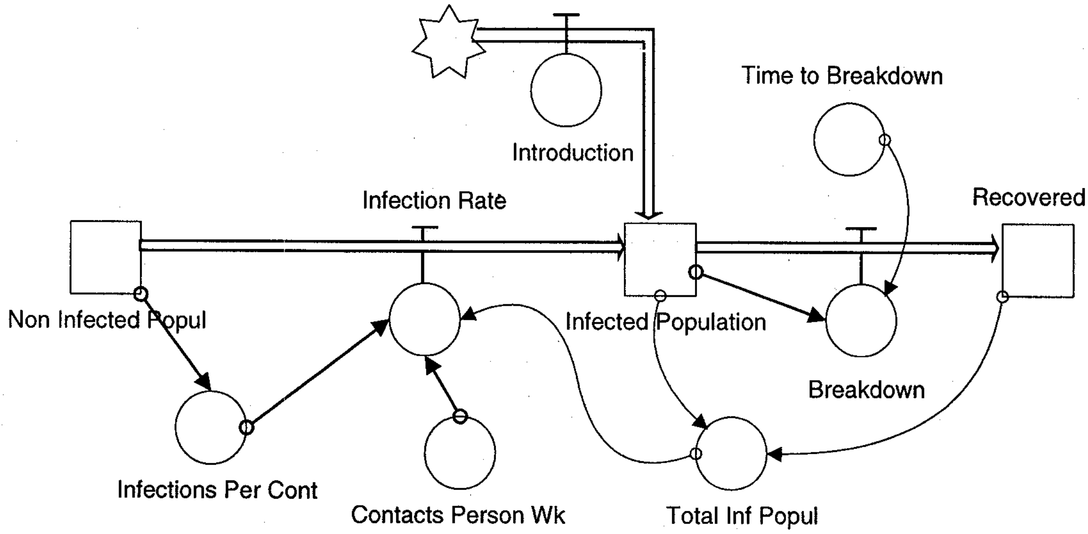

Figure 3. ***A systems model of an infectious disease***

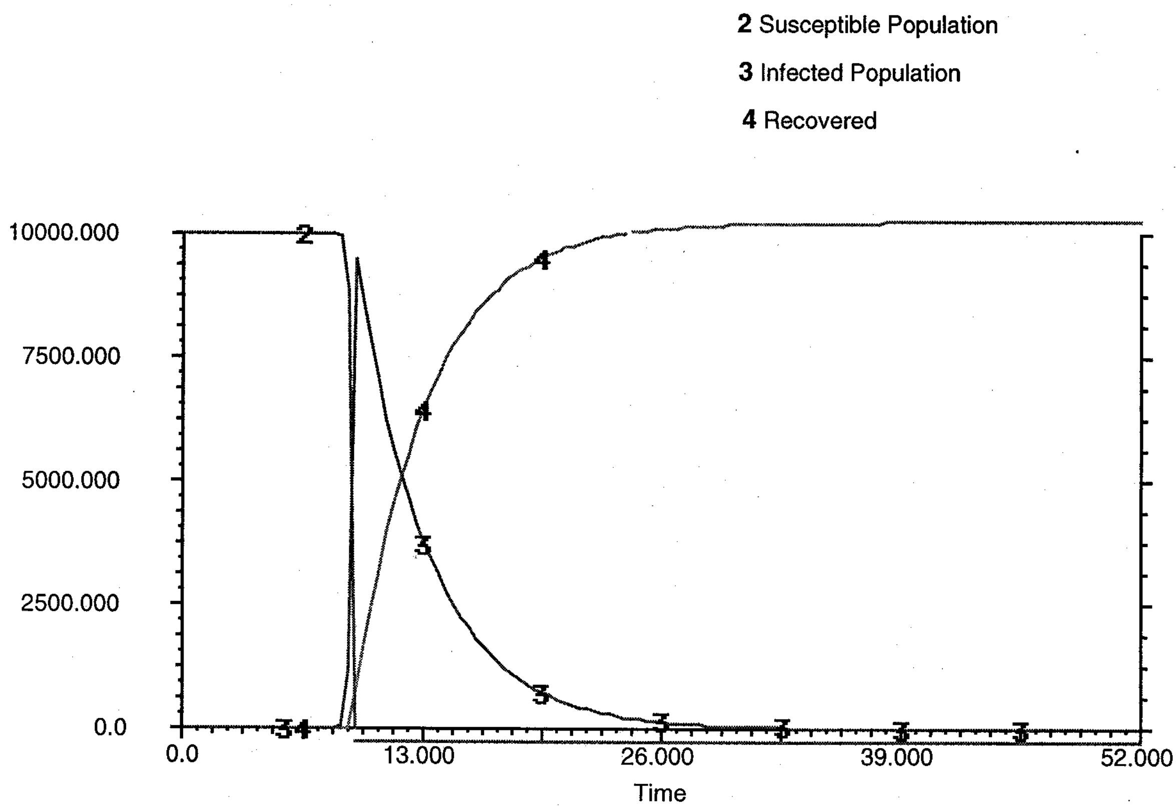

Figure 4. ***The output from a model of an infectious disease***

Some diseases, like malaria, remain dormant within individuals, and can recur. Others, like AIDS, are asymptomatic for a long time, yet are infectious (and, of course, the recovery rate is very small). For others, like gonorrhea, no body defenses build up, and exposed individuals can be re-infected. Each of these different diseases could be modeled by a variant of the simple EIR model in Figure 3.

Applying the Model to Education

In the context of education, one might adapt the same mathematical model to describe the functional form of the impact of professional development on'classroom practice. A mapping of elements between epidemiology and education is shown below.

Epidemiology	Professional Development
Susceptible	Susceptible
Exposed	Exposed
Infected	Changed classroom behavior
Recoverd	Classroom behavior relapses

A variety of versions of the model can be considered that reflect different forms of professional development (e.g., Ridgway, 1997). Pyramid models have the same structural form as the model in Figure 3. (Conceptually, they differ in that the nature of what is transmitted-classroom behavior-is far more likely to suffer mutation than is a disease passed from one person to another.) In a model to simulate change in classroom practices by teachers who attend summer schools, there would be no effect of the Total Changed Population on the Influence Rate; and so on. The evaluation of ESR plans (and indeed the whole engineering science of ESR) can benefit from some direct modeling of subprocesses, such as the process of professional development.

The process of building models need not be difficult. However, difficulties do arise because of ignorance about key features of education, such as the likelihood of changed classroom behavior given exposure to different sorts of professional development, or the likelihood of certain kinds of classroom practices reverting to old forms. The act of thinking about exactly what information is essential to inform the model is an important component of evaluation and is one of the benefits that derives from modeling activities.

Even informal systems modeling can serve a valuable role in the evaluation of SI plans. For example, using just the simple model here, an evaluator might ask:

- How many teachers need to change their classroom behavior?
- What opportunities do they have to be exposed to new practices?
- What is the probability that teachers will change their classroom behavior after exposure?
- What is the time horizon for remission?

The answers to many of these questions lie in the existing (analytically derived) literature that relates to descriptions of classroom behaviors and the effectiveness of different professional development experiences in changing classroom behavior over different periods of time.

Different gradations of formal modeling (from hand calculations to computer simulations) can provide demonstrations that an SI plan can (or cannot) in principle have an impact on classrooms throughout the whole system, in the time scale specified.

Comparing the Knowledge Bases in Epidemiology and Education

In modeling the epidemiology of common diseases, all the required information is available, including:

- a well-developed description of diseases in terms of the cycle of symptoms in humans, direct observations of viruses or bacteria, and effective and ineffective means of transmission;
- the "natural" time course of a disease within an individual (so "infectious periods" and "recovery rate" can be identified);
- reasonable estimates of the infection rate; and
- data on the time course of diseases through populations to validate models.

To build a model of the dissemination of professional development, one needs:

- a description of the target behaviors;
- knowledge of the "natural" time course of skill and knowledge acquisition, and of their breakdown;
- reasonable estimates of the rate of change; and
- data on the time course of the changes in classroom behavior through the population as a whole.

Existing literature can act as a guide when evaluating SI plans. From the viewpoint of formative evaluation, detailed studies of specific interventions on the desired classroom behaviors are necessary to inform the model that then can be used to provide formative feedback. Again, the purpose of systems-model-based evaluation is to predict the likely success or failure of current practices. The features of formal models that allow "what if?" conjectures to be explored are critical. Formal models can be used to calculate the minimum total amount of time that must be spent on professional development, using the methods adopted by a particular SI, that will be required to reach all the teachers in that SI, for example. This result can be used to make judgments about the value and worth of the initiative.

Professional development has been used here to provide an example of the roles that might be played by systems models in evaluation. Any aspect of an SI could be the focus of a systems model, at the level of evaluating plans, or for making decisions. At present, it seems unlikely that systems models of an entire SI would be worthwhile, because of the likely complexity of the model and the problems of parameter estimation.

An Introductory Analogy — Breeding Butterflies

Imagine that one faces the task of evaluating new programs designed to breed butterflies. In order to breed butterflies, the proposer/breeder needs a detailed knowledge of the life cycle of the butterfly. Butterflies go through a number of distinct stages — egg to caterpillar to pupa to butterfly. At each stage, different environments are necessary (a leaf to stick to, leaves to feed on, twigs to hang from, environments to fly in with pollen to feed on, and places to meet fellow butterflies). Different creature behaviors are to be expected in order to promote population growth (sticking, browsing, hanging, feeding, and mating). What intellectual tools might benefit the evaluator of new breeding programs? The evaluator needs:

- a knowledge of the stages of development;
- a knowledge of the conditions that are appropriate at each stage; and
- ways to describe the stages, signs of development within each stage, and appropriate environments.

Armed with this information, the evaluator can make informed judgments about:

1. Plans for breeding
- Does the breeder have an account of the life cycle stages?
- Are appropriate environments being created?
- Will procedures be put in place to monitor the stages of development, to provide appropriate environments, and to monitor them carefully?

2. Formative evaluation
- At what stages are the different butterflies (how are stages conceived and described)?
- Has the appropriate environment been created for each cycle of butterfly life? How is the environment monitored and modified? How are environments conceived and described?
- What is the breeder doing to discover how things can be changed to make them more effective? What mechanisms are in place to enable the breeder to improve on current breeding practices?

3. Summative evaluation
- How many butterflies are produced?
- What varieties of butterflies are produced? How can butterflies be classified?
- How healthy are they? How can the state of health of a butterfly be determined?

It is clear that the stages of development (if they exist at all) in changing educational systems are far harder to describe than the stages of butterfly development. Schools are unlikely to be as similar to each other as are different kinds of butterfly. In education, the knowledge base associated with systemic change is at an early stage of development. Information about the cycles of change and the conditions that trigger these changes is only just beginning to emerge. If the notion of macro-systemic change is to be taken seriously, an essential target for research in evaluation is the development of a knowledge base about stages of reform, critical factors in the reform process, and the like, which will complement existing knowledge about the evaluation of more familiar change derived from program evaluations.

Analogies with imaginary evaluations (here, butterfly breeders) can be useful for scene setting. However, a detailed account of actual evaluation practices during a socially important macro-systemic reform is more likely to highlight issues critical to educational evaluation.

A Detailed Example — Prairie Restoration

Aldo Leopold (1949) emphasized the importance of studying the whole ecology of a landscape: plants, animals, and the physical setting. His pioneering work in the 1930s on prairie restoration at the University of Wisconsin-Madison arboretum led to a rich body of knowledge about the restoration of natural environments.

Ecology might provide some good analogies for education because the interacting elements are themselves complex systems (e.g., an individual animal or plant can be viewed as a system in its own right, comprising a variety of subsystems (blood circulation and systems for nutrition in animals, food creation and fertilization in plants, for example); subsystems exist with different degrees of coupling (such as dry soil communities, wet lands, etc.); systems are affected by external conditions, some of which are relatively stable (such as climate and soil), and some of which are relatively unstable (such as fire and flood); changes occur over time, sometimes via natural shifts in environmental conditions, or sometimes via deliberate or accidental human intervention. Research methods are well established, although models of change are poorly developed. Nonetheless ecology has a number of key concepts and methods that can inform practices in education for data collection, data display, and descriptions of phenomena, and for planning, implementing, and monitoring change. Ecologists have studied a range of situations in order to build their current state of knowledge:

- systems in relative stasis;
- systems that are restarted from a relatively undeveloped state (for example, after some disaster, such as a massive flood in a canyon, that sweeps almost everything away; or after fires, volcanoes, or nuclear testing);
- systems undergoing change as a result of nonintentional changes (for example, in response to changes in water provision, or nutrients, or the emergence of some new predator, e.g., starfish eating the coral on the Great Barrier Reef); and
- the active management of ecosystems, both to maintain stasis (e.g., preserve wetlands) and to create “new” (actually, often “old”) ecosystems from existing systems.

Studying Systems in Relative Stasis

Ecosystems in “relative stasis” are recognizably the same over periods of years or decades. Ecologists have devoted a great deal of time to the detailed description of abiotic factors (temperature, exposure, water, nutrients, wind, and the like), and of assemblages, communities, and guilds of plants and animals in the field. They have described in detail individual plants and animals (descriptions include both form and behavior) under both natural and laboratory conditions. In addition, they have conducted controlled experiments in both the laboratory and the field on the conditions that favor and inhibit growth.

The resulting bodies of knowledge from research on ecosystems serve a number of distinct functions. They enable ecologists to identify certain types of communities (such as wet grasslands, prairie, Savannah, etc.) for inventory and mapping. They show which conditions are favorable for the growth of different plants, animals, and communities. They show which communities of plants and animals coexist necessarily or easily. These data are useful because they suggest some symbiotic relationships-l-for instance, between pollinating insects and plants with flowers. They offer a view about what is common and what is rare locally, nationally, and internationally. This knowledge is important for informing possible future actions on changes in land uses. For example, actions that will destroy rare plant communities are viewed as having a higher cost than actions that destroy common plant communities. This information also offers pointers to the type of ecological systems easily recreated, given particular abiotic factors.

Evaluators of education systems in relative stasis would find it extremely useful to have access to information about education that is analogous to the information available to ecologists. For example, when evaluating SI plans, useful information includes:

- descriptions of different types of communities such as classrooms, schools, neighborhoods, school districts, and states;
- causal relationships between classroom practices and student attainment; and
- identification of common and rare activities.

In the short term, this information is unlikely to be available, but developing this knowledge base would be valuable both for evaluators and for those engaged with ESR.

Studying Systems Undergoing Change

Ecologists make detailed studies of systems undergoing change. Consider, for example, the recolonization of desert after nuclear testing is stopped (e.g., in Nevada). A sequence of changes can be observed:

- blasts kill all life;
- a year after the last blast, some spring annuals appear, such as desert pincushion and stickleaf, from seed brought by the wind or by birds;
- as these plants die, nutrients are added to the soil, their roots increase water retention, which in turn reduces soil temperature, thereby changing the environment in important ways;
- this new environment now permits other plants to grow, such as wild buckwheat and foxtail chess.

Recovery from nuclear devastation is a dramatic example of phenomena that occur more commonly (such as, recovery after the eruptions on Mount St. Helens). Most ecological systems can be judged to be in a state of change, if a long enough time scale is considered. The community that ends the succession (for a long period of time, at least) is called the climax community. An important lesson for education is that, over a long time scale, the same climax community can evolve from quite different starting points.

For example, woodlands of spruce trees can derive from different initial conditions:

- rock to lichen to meadow to aspens to spruce; or
- pond to marsh to meadow to aspens to spruce.

The knowledge base that supports macro-systemic modeling comprises descriptions of systems undergoing change where significant changes take place in the system itself. The knowledge base associated with systems in relative stasis is used here, too.

The knowledge base can serve a number of different functions:

- identifying conditions favorable to certain kinds of communities;
- planning desirable changes;
- implementing desirable changes; and
- monitoring ongoing changes.

Ecological Restoration as Systemic Reform

The settlement by Europeans created marked changes in the ecology of North America. Corn and wheat replaced prairies; cities covered plant and animal habitats; waterways were created, and a great deal of land was drained. This pattern of increased agriculture and urbanization reflects changes globally and is associated with a decrease in biodiversity. At the time of the settlement of Wisconsin, about 42% of the land was covered in oak savannahs, which are areas of scattered trees with some ground level vegetation (Kelley, 1997). Oak savannahs now account for around 0.0 1% of land cover.

A number of initiatives are underway to recreate ecological systems in Wisconsin and elsewhere that have been destroyed by farming or other sorts of cultivation:

- large scale restorations of oak savannahs in arboretums and other locations;
- prairie restoration in school grounds;
- schemes to promote gardening with native plants; and
- European initiatives to “decommission” agricultural land.

These initiatives have strong parallels with systemic reform in education. Each ecology provides a classical example of “a system.” There are well-articulated views about the nature of the changes that are sought. Deliberate attempts are made to bring about particular sorts of change. Ecology has an advantage over education because of its “engineering base”: there are clear descriptions of the elements in the system (plants, animals, and their behavior over their life cycles); phenomena are well documented; the outcomes of different environmental changes can be predicted with a reasonable degree of accuracy; techniques exist to monitor and adjust the course of systemic reform; and attempts at system change have a reasonable track record of success.

Ecological management typically follows this sequence:

- A site is analyzed via a baseline survey.
  - What communities have existed here before?
  - What is here that should be preserved?
  - What will have to be removed?
  - What could exist on this site?
- Goals are set and plans are made.
  - Why does this site need renovation?
  - What are the new goals for the site?
  - Does a particularly rare species need to be preserved?
  - Should things that grow well locally be reestablished?
- The biotic community is selected.
  - What species will be used?
  - What needs to be implemented in order to attain the final goal state (eg., in terms of abiotic factors and interim plant cornmunitites)?
- The site is prepared.
- The site is managed.
  - The attainment of goals is monitored.
  - Radical approaches are devised to support desirable development.

Site analysis — Describing the current system. Different ecologists approach the problem of classification in different ways. Some begin by identifying distinct plant communities (e.g., Rodwell, 1991-1995), others by describing the abiotic conditions that favor the growth of individual plants. For the purposes of this analysis, the approach pioneered in Wisconsin will be used. Natural prairies are commonly classified (e.g., Curtis, 1971) as wet, wet mesic, mesic, dry mesic, or dry. Some species of plants are found predominantly in one sort of environment, and not in others, while some plants can be found under a great variety of conditions. Precise definitions of ecological systems are not always possible. For example, in the definition of "oak Savannah" there is agreement on the nature of the tree canopy (mainly oaks), but the nature of the ground layer is less certain, since it comprises nearly all the plant species in the savannah community.

Ecologists first analyze the available land in terms of site, soil, drainage, and light. Then they identify those families of plants and plant communities that will grow well in those settings. Curtis (197 1) examined the vegetation in over 1400 examples of prairie, wetland, and forest and related species composition to environmental factors, such as the nature of the soil (nitrogen, phosphorus, potash, pH, moisture content, moisture retaining properties, organic matter concentration, permeability, soil components) and local climate (e.g., Wisconsin mesic prairies have an average precipitation of 3 1.3", a growing season of 152 days, and monthly mean temperatures that range from about 16 degrees to 72 degrees

Fahrenheit).‘Prairies develop in full sun and require at least 12 hours sunshine during the growing period. Prairie types are determined by the qualities of the soil and by drainage.

Similar environments produce similar ecologies. Places that lie between clearly distinguishable ecosystems are described as “tension zones” or “buffer zones.” Ecologists have a variety of well-developed techniques for describing plant communities. These include systematic sampling by inspecting communities lying along particular lines or by random sampling using quadrats. These sampling methods are complemented by detailed systems for describing exactly what is present. However, ecologists differ markedly in how they define the variety they encounter. A description of an ecosystem is likely to require visits many times during a year, so that changes in the plant community can be observed.

- *Goal setting — Considering possible plans.* An archive of knowledge developed by research aids those interested in change. Analysis of a site using abiotic features allows an ecologist to identify individual plants and plant communities that could thrive and those that are unlikely ‘to succeed. Novices can easily access this knowledge. A beginner can gain a great deal of information to identify individual plants likely to succeed in the site conditions that prevail locally. A school interested in restoring a prairie, for example, can call upon a considerable body of knowledge (Murray, 1993). Design work should begin by looking at natural communities living in settings as similar to the target site as possible. Seeds for planting should be collected from sites with conditions as similar as possible to the site to be planted. Schools and gardeners can buy mixtures of native plants from plant collections indexed in terms of the characteristics of the most commonly occurring local conditions.

  It is important at the outset to map out the evolution of the target ecology or, in the language of this monograph, to specify the stages of macro-systemic change. A central idea is that certain conditions have to be created in order to allow later developments. For example, by planting oaks at the outset, the way is paved for a savannah at a later time, once shade is established.

  Species selection is constrained by the site. Within these constraints, planning should address the visibility and visual essence of different plants at different times. This will be a function of the distribution of species. Species grow to different heights, bloom at different times, and reach maturity over different periods of time. Schools restoring prairies are advised to plant both fast maturing species and some slow maturing ones (Murray, 1993).

- *Site preparation.* “Proper preparation of your site is probably the single most important factor in.. .success...” (Murray, 1993). This advice is based on evidence from the early days of prairie restoration, when native plants were planted into degraded pasture. Native plants failed to compete well with the pasture plants. A good deal of empirical work identified plants that should be avoided, or removed if they are discovered.

- *Management.* Prairie plants, like most perennials, do not flower the first year they are planted. Rather, they spend most of the first year developing a root system designed for surviving drought. A prairie planting often does not look like a prairie until the fifth year after planting. A great deal depends on weather and how effectively weed competition has been controlled.

Murray (1993) distinguishes short-term management (the first two years) and long-term management. In the early years, the central problem is weed control. The treatments are familiar to most gardeners. Over the longer term, different management techniques are necessary. The central problem in the long term for marginal prairies such as those in Wisconsin is that they revert to woodland without interventions such as burning, grazing, or mowing. Curtis (197 1) discovered that fire was essential to the vigor and spread of prairie plants, and that most prairie plants show significant increased blooming after fires.

Prairie management techniques include prescribed burning; controlling exotic plants and pest plants; collecting and distributing seeds; propagating plants; and protecting sites. In restoring prairies in school grounds, many sites are similar and have used a common set of plants and seeds as their starting points. Nevertheless, quite different plant communities develop; all are recognizably prairie-like, but the dominant grasses differ, as do other ecological features (Murray, 1997, personal communication). In general, there is increasing diversity as the prairie matures.

If the long term plan is to create Savannah, then the management activities are also concerned with steady evaluation of the land, as well as with the establishment of a relatively steady state.

Applying the Model to the Evaluation of ESR

The sequence of ecological management has relevance for evaluators concerned with the evaluation of an SI plan or the evaluation of the way an SI approaches school planning. It is worth considering each aspect in turn. The purpose of the analysis is to identify the nature of the knowledge used to support ecological restoration in order to identify some research targets for the evaluation community concerned with studying systemic reform in the context of ESR. The final section of this monograph will describe ways in which the requisite knowledge base for the evaluation of ESR might be created.

Site analysis-Describing the current system. Schools and school systems vary in a great many ways. Establishing ways to classify schools as alike or dissimilar on successful educational activities poses an interesting challenge. It would be useful for an evaluator to be able to classify a specific educational setting using a broad classificatory framework and to make informed judgments about the likely success of the proposed set of educational activities in that setting. Simply knowing what educational environments are common and what activities are tolerant of a wide range of environments would be useful for both planning and evaluation. Knowing that some kinds of activities can only take place in a narrow range of circumstances would be powerful information for evaluators to have. It is unlikely that precise classifications of educational communities will ever be possible. This is not fatal to the argument-ecologists face exactly the same problems. Fuzzy knowledge can be very useful. The scale of the research necessary to describe ecologies is impressive (see

Curtis, 197 1; Rodwell 199 1- 1995). Similar comparable levels of investment probably are needed to achieve similar levels of description in education. Short-cut methods that might be suitable for short-term purposes are described in the final section of the monograph.

Goal setting-Considering possible plans. Analysis of a site allows the ecologist to identify those individual plants and plant communities those that could thrive and that are unlikely to succeed. Murray (1993) recommends that prairie restorers identify sites that are as close as possible to the target site and use them as models for their restoration. In the educational context, this has a number of parallels. Evaluators might critique plans for SIs by asking about the base of evidence that justifies the proposed scheme. A good deal of evidence has been developed by the education community on which ESR can build (e.g., Grouws, 1992). Evaluators might review working examples established in settings that match the target settings reasonably closely. Plans that set out to extend local good practice might be more likely to succeed than plans that promote good practice taken from contexts that are considerably different from local conditions.

Murray stresses the importance at the outset of mapping the evolution of the target ecology or, in the language of this monograph, specifying the stages of macro-systemic change. A central idea is that certain conditions have to be created in order to allow later developments. Within the constraints imposed by the site, planning should address the visibility and visual essence of different plants throughout the year and over the course of the restoration. She recommends planting some fast maturing species and some slow maturing ones. There are analogies with education. Evaluators can judge plans for macro-systemic change conceived by an SI. Timeframes for effects of proposed interventions need to be judged. Plans without any "fast maturing" effects are likely to be less successful than those that include a mixture of more immediate and long term effective changes.

Site preparation. From the early days of prairie restoration, native plants planted into degraded pasture failed to compete well with the pasture plants. Site preparation requires identification of plants that should be avoided and removed if they are discovered. Again, there are useful analogies for education. Teachers are often influenced by the ways in which they were taught. Teaching methods have often been practiced over many years, and new approaches that are planted on top of these practices are unlikely to persist for a very long time. Evaluators need to understand and judge how the new is to fit in with the old. It is necessary for evaluators to check that new methods will receive appropriate resources so that they can compete with well-established methods. They need to determine how undesirable forms of teaching and learning will be eradicated.

Management. Prairie plants spend most of the first year developing a root system designed for surviving drought. A prairie planting can be rather unimpressive for as long as five years. Murray (1993) distinguishes short-term management, over the first two years, and long-term management. In the early years, the central problem is weed control. Over the longer term, different management techniques such as prescribed burning, control of exotic plants and pest plants, seed collection and distribution, plant propagation, and site protection are essential to the vigor and spread of prairie plants. Again, there are interesting analogies with education. A key issue for SI evaluators is to identify a reasonable time scale over which an educational innovation should be judged. It is hardly appropriate to keep digging up each plant to see whether its roots are growing; conversely it is irresponsible not to do some form of evaluation to ensure that planned growth is on track. From the viewpoint of evaluating plans, an evaluator should ask about SI plans to inform stakeholders about the likely timescale over which negative and positive signs might be detected. Plans for recognizing and controlling undesirable teaching and learning activities over both the short and long term can be reviewed. There is at least one example in the UK where the incoming head of mathematics burnt all the mathematics books in school as his starting point in curriculum reform (John Mason, 1997, personal communication). Evaluators might reflect on earlier discussions of the need for careful site analysis before recommending this approach as a panacea for educational ills.

The Knowledge Base

By way of a summary, tables 3, 4, and 5 identify important aspects of the knowledge base that make ecological restoration possible. Analogies are drawn with education. The final section of the monograph makes suggestions about how this knowledge base might be built up.

Table 3
Illustrations of the knowledge bases in ecology and education: Classification and description

Ecology	Education	
	Individual level (e.g., students and teachers)	Organization level (e.g., schools and SIs)
Methods to classify individual plants; detailed descriptions of phenomena such as plant heights, colors, blooming times	Methods to describe students and teachers; methods used to describe standards and curricula; methods to describe student achievement	Methods to describe classrooms and schools; methods to report student outcomes
Methods to recognize healthy and unhealthy growth	Methods to recognize healthy and unhealthy growth, e.g., appropriate measures of changes in student performance	Methods to recognize healthy and unhealthy growth, e.g., of organizations
Methods to describe environments	Methods to describe environments, e.g., in class	Methods to describe environments, e.g., in school; or the school in the community

Table 4
Illustrations of the knowledge bases in ecology and education for understanding learning and change

Ecology	Education	
	Individual level (e.g., students and teachers)	Organization level (e.g., schools and SIs)
Studies of seasonal cycles of growth	What is the course of learning under different treatments ? What are the cognitive characteristics of students at different ages? How does professional skill develop over a course of years? If teachers hit a "steady state," how do you pull them out?	How do school structures change over time?
Analysis of reproductive mechanisms: conditions, time lines, methods of dispersal (e.g., good times for seed collection, ideas on appropriate seed collection methods, germination treatment)	How do students and teachers learn? How is knowledge disseminated? How are good ideas passed on without damage?	What are the mechanisms of school change? What are the time lines? How is knowledge disseminated? How are good ideas passed on without damage?
Analysis of environment plant interactions	What student-by-treatment interactions exist? What teacher-by-treatment interactions exist? What new classroom practices would work well in a particular context?	How does school structure affect classroom practice? What SI actions are effective in bringing about changes in different types of schools?

Table 5
Illustrations of the knowledge bases in ecology and education for understanding systems and systemic change

Ecology	Education	
	Individual level (e.g., students and teachers)	Organization level (e.g., schools and SIs)
A classification of ecosystems and descriptions of the variability within each	What sorts of classroom organization are there? How does classroom style interact with curriculum? What school-home links exist? What are the processes in each?	How can the processes within schools and school systems be described? Are there "clusters" of each that have the same essential features?
Descriptions of the evolution of ecosystems over time	How does classroom practice change over time? How will student attainment change over time? What sequence of changes is appropriate to show improvements over both the short and long term?	How do schools and districts change over time? How can they be planned? What resources have to be put in place to create an appropriate environment for new developments?
Details of invasive weeds, and methods for their removal	What restricts learning for individual students? What types of regression occur in teaching practices after professional development? Can poor practices be recognized and changed?	What undesirable organizational developments occur? How can they be recognized and eradicated?
Management via diagnosis and remediation	How can progress be monitored and misconceptions remedied? How can regression in classroom practice be detected and corrected?	How can organizational progress be monitored and undesirable changes be corrected?
Ways to communicate with the local community about what to expect over a 5-year time course	Ways to communicate with students and teachers over a short period of time	Ways to communicate with the local community about what to expect over a 5-year time course

Evaluation of ESR in education needs to consider the ESR design, the implementation and management of ESR, and the outcomes of ESR. Successful evaluation of each of these components of ESR needs to call upon an evidence base that is as rich as the evidence base in epidemiology and ecology. The education community already has a considerable assembly of evidence about the conditions of learning at the level of the individual, the classroom, and the school. However, ESR is a new venture. Strenuous efforts are necessary to. learn as much as possible, and as quickly as possible, so as to maximize the effectiveness of current initiatives. The final section of the monograph offers suggestions on how an appropriate knowledge base might be constructed.

Building the Evidence Base for the Evaluation of Systemic Reform

There is a pressing need for a research community of practitioners and academics devoted to the evaluation of ESR that contributes to and draws upon a common pool of knowledge. Research cultures are not set in stone. There are a number of examples of disciplines that have emerged because of social need (e.g., statistics: Hacking, 1990) or academic need (e.g., molecular biology, neurophysiology, cryogenics, geophysics). People concerned with ESR might reflect on the things that have to be put in place to establish ESR and the evaluation of ESR as a viable academic discipline with its own distinctive features.

The underlying philosophy of ESR requires that evaluation must become an integral part of the whole system. Evaluation must not stand apart from the activities of SIs or ESR. The evaluation community has an important role in assembling, acquiring and disseminating knowledge about effective practice. (See the work of the Consortium for Policy Research in Education (CRPE), for example, Goertz, Floden, and O'Day, 1995.) One can hardly evaluate a plan and the techniques for managing that plan without some knowledge of what is likely to work. Nor can one evaluate the success or failure of a course of action without some definition of what desirable effects are, and how these desirable outcomes might be assessed. In the case of ecology, the whole design, management, and evaluation cycle of SR is based on a great deal of knowledge gleaned from different sources. The education community needs to continue building a body of knowledge about the process of educational change, making extensive use of information gathered from the evaluation community. This knowledge will complement the large body of work that has already been conducted in educational research on processes of learning, classroom practices, and school effectiveness.

Evaluation is not a neutral activity. The act of observing can result in profound changes in what is being observed. For example, an interview for gathering data to evaluate the plans for an SI might ask about aspects of ESR that the SI directors did not consider. The result of the initial evaluation interview is likely to be a revised plan, rather than a poor score on "planning" followed by the unfolding of a failing SI. Similarly, asking about how performance will be measured, how feedback on progress will be obtained, and what multiplier effects will be called upon can change the design of the SI. It follows that evaluators can serve a role in the dissemination of information about effective SR.

The levels of scientific knowledge described in an earlier section-phenomena, effects, models of data, theories-provide a framework for conceptualizing different kinds of knowledge. Each SI, past and present, can be seen as a set of educational experiments that can provide evidence at every knowledge level. Iris Weiss (1997) offered vivid illustrations of ways in which the knowledge accumulated by the evaluation community can be used to inform both the day-to-day pragmatics of change and the local theory of ESR. As this evidence accumulates and is collated, it will be possible to offer more detailed theories of ESR, as well as techniques for the evaluation of individual SIs and ESR as a whole.

Describing phenomena. Each SI evaluation has collected evidence about a whole range of phenomena related to educational change. As this evidence accumulates (e.g., Massell, Kirst, & Hoppe, 1997; Breckenridge, Goldstein, Zucker, & Adelman, 1996; Clune, Millar, Raizen, Webb, Bowcock, Britton, Gunter, & Mesquita, 1998), the collection of examples of phenomena will make it possible for evaluators embarking on new evaluations to identify situations similar to the ones that they are investigating. When close matches are found, evaluators can examine the relevant case histories for evidence about likely outcomes, to inform the evaluation of plans and to help design formative feedback.

Discovering *effects*. Gathering descriptions of phenomena, along with detailed descriptions of the circumstances under which the phenomena occurred, can provide the basis for conjectures about effects. That is, factors that are often found together may be causally related.

Creating models of data. Modeling requires evidence about effects and some mathematical tools. Analytic tools, such as analysis of variance and structural equation modeling, provide relatively simple models of static data. Tools such as dynamic modeling packages provide ways to describe data that allow complex feedback between elements.

Creating models and theories. At present, ESR has many of the hallmarks of an intellectual field in its earliest stages of development (e.g., Knapp, 1997):

- some definitions are absent or contradictory;
- some ideas are conflated (e.g., although "systemic" refers to "influencing the whole system" it is often confused with "standards-based reform." Logically, one could have a systemic approach to a back-to-basics curriculum and an analytic approach to standards-based reform);
- accounts of the elements in the system, specifications of their interconnections, or the functional relations between pairs of variables are patchy;
- there are few attempts at formal modeling;
- there is little analysis of what a theoretical account might look like, or of what the appropriate level of specificity might be; and
- despite widespread use of the term "systemic," there appears to be little use made of the large literature on systems theory (e.g., Banathy, 1992; Beer, 1976; Bertalanffy, 1968; Checkland, 1981) or the literature on the management of change (e.g., Asch & Bowman, 1989; Kanter, 1984; Wilson, 1992).

Dynamic models can be seen as theories of system change. The development of dynamic models can help the development of the intellectual field of ESR as a whole. The creation of macro-systemic models of change require a description of systems (for example, classrooms, or schools, or school districts) at different times as they evolve. Any discovery of similar developmental patterns across systems then can provide the basis for macro-systemic theories, which can guide both the evaluation of SI designs and development.

As with any attempt to build theories, there is a need for a critical examination of the quality of the evidence that is available. In the context of evaluating ESR, it is essential to begin with a distinction between the intended curriculum, the implemented curriculum, and the attained curriculum. The success of an individual SI should be judged at each of these levels. In judging the success of the theory of ESR (whatever it might be!), one must be careful not to confuse a barren theory with a poor implementation. It would be foolish to form a negative view of the utility of the theory of ESR on the basis of evidence from SIs that had failed to introduce ESR. A related point should also be borne in mind. Because a particular theoretical approach has been shown to be valuable in one instance, one cannot conclude that it can be easily applied across a wide range of situations.

Michael Faraday, the nineteenth century English physicist, once claimed that there is nothing so practical as a good theory. It also is clear that there is nothing so impractical as waiting for a good theory before any actions are taken. Evaluation of SIs can support the engineering of future SIs and can inform the development of the theory of ESR. The next section offers some ideas about how the evaluation community can help build a knowledge base that can promote effective ESR.

Strategies to Build the Evaluation Evidence Base

This section continues the theme of the monograph by identifying research techniques used in a range of academic disciplines outside education. Disciplines were chosen that face problems similar to those faced in education, notably, that the information flow is essentially infinite compared with our ability to record and analyze. Ideas are presented in the form of strategies that might be used to accelerate the development of the knowledge bases necessary for the evaluation of SIs and of ESR. Identifying people who might conduct the necessary work is not easy. Many of the strategies require efforts that go way beyond the resources provided to individual SIs for evaluation. A web site that collates information from different evaluations could provide a real service to the evaluation community. It could be used to guide the evaluation of future SI plans and could be useful in the development of formative feedback. Again, evaluating evidence and presenting it in a usable form are nontrivial tasks that would require the deployment of considerable resources.

It is logically impossible to draw conclusions about the critical factors in an SI and ESR on the basis of a single case history. Given several case histories, one can at least begin to piece a story together. However, the multidimensionality of systems still poses major problems for making inferences. Researchers schooled in conventional science commonly use experimental methods where the majority of variables are held constant and the interrelations between a to evaluate educational effects. Humans have considerable problems in handling large numbers of variables. One approach to the problem might be to store research results in a database that allows a very large number of descriptors of the system and the treatment to be stored. As individual case histories are added, the evidence will accumulate, albeit in patches. Conjectures about critical variables, triggering thresholds, cost-effectiveness, and the like can be explored and re-explored as more data accumulates (this strategy allows direct exploration of phenomena and the search for effects described above). A variety of ways to present complex data to make it more intelligible is described by Tufte (1983, 1990, 1997). Such methods are rarely used in education and might be of value.

Strategy 1: Compare Comparable Schools and Districts

An approach to the problem of identifying effective educational treatments which avoids statistical moderation (and the required strong assumptions) is to look for differences in performance between individual schools, or between whole school districts, that are roughly comparable, but have had different levels of involvement in the SI. A problem arises in defining "roughly comparable." One solution is to use multidimensional scaling (MDS). This is a statistical technique rather like factor analysis, but it allows more control over the necessary statistical assumptions. It allows a number of objects, each of which has a number of different attributes, to be related to each other in an object space. MDS allows schools (or school districts) to be related to each other in terms of their distance apart determined by some combination of the attributes that are available. For example, suppose data are available on school funding levels, some measure of family poverty, and local crime rates. A measure of similarity between two schools can be obtained by calculating the differences on each indicator and summing the absolute value of these differences. Distance measures can be as complex as one chooses. Factors can be scaled, can be weighted, and can be combined in all manner of ways. MDS can be used as a descriptive tool that facilitates the identification of "similar" schools. The search for effects begins by exploring practices in schools that are similar in terms of relevant descriptors, but that differ in terms of student attainment. It is clear that the measure of "similarity" will reflect one's theory of the key features that determine school performance. Few people identify the gender of the head teacher as a key variable, or the local weather conditions. More common choices might be school size, percentage of students from families defined to be economically disadvantaged, percentage of students from different ethnic groups, average educational attainment of parents, local crime rates, rural-urban location, poverty level, and prior student attainment levels in SMET. Refining these implicit theories about what makes schools similar or different will contribute to understanding more about educational processes and educational effects. Comparing schools that are similar in terms of these characteristics, but different in terms of student attainment, is valuable for evaluating the design of SIs.

MDS has the potential to be a powerful technique to support forming and testing hypotheses about ESR. Schools that seem to show considerable improvements can be judged against schools that were comparable initially. The use of matched controls is a powerful educative device for teachers, principals, and supervisors, as well as for evaluators. It is of little practical help to be told that the attainment of students in inner city schools is lower than that of students in middle class suburbs. This information is too coarse in grain size to be useful. Teachers can hardly be expected to reshape their city in order to improve the educational attainment of their students. Information about the relative attainment of students in schools that are comparable is far more useful.

Strategy 2: Treat the Attribution of Causality Seriously

The development of a theoretical framework for interpreting educational change is critical to success, as is amassing a large collection of results that hang together. Interventions that can be shown to affect student outcomes provide strong evidence about causality, if these effects are shown to be robust and unambiguous. ("I believe that A causes B. I change A in these sites, and B changes, but changes in B hardly occur at all if nothing is done about A.")

The problems of attributing causality using the conventional tools of social science research discussed eloquently by Manski (1995) has been addressed by seismologists (e.g., The Incorporated Research Institutions for Seismology, funded by NSF, *www. iris.washington.edu/*) who are concerned with monitoring seismic events to determine their likely causes. In particular, they are interested in distinguishing among nuclear explosions, mining blasts, earthquakes, and meteor impact. This work has achieved a new prominence with the recent United Nations resolution to end all nuclear testing. This challenge of making plausible inferences about the cause of some detectable change is directly analogous to the problem faced in education, where a need is seen to distinguish between alternative possible causes.

Seismologists use a number of distinct kinds of evidence when forming judgments about causality:

- the nature of the event, its "fingerprint" (e.g., shock waves from nuclear blasts begin with a distinctive spike as the ground is compressed violently, followed by rapid exponential decay; earthquakes typically begin with minor tremors that increase in strength; aftershocks are common);
- knowledge of local capability (e.g., nuclear tests are more likely in Pakistan than in Barbados);
- knowledge of local intent (e.g., a nuclear test is more likely in territory held by nations that are not signatories to the UN resolution than in territory held by strong advocates of the resolution);
- location (e.g., a seismic event located on a French island in the Pacific Ocean is more likely to result from a nuclear test than one located in Los Angeles); and
- size of the event (e.g., seismic recordings can eliminate mining blasts as a cause of major events; the size of the event can be used to judge the likely size of a meteor impact crater and hence the ease which it could be found).

Other triangulating evidence includes:

- eyewitness reports (e.g., lights in the sky before a seismic event suggest meteor impact as the most likely cause);
- air traffic data; and
- evidence on the ground, such as new craters.

Seismology has a number of specific lessons for education:

- Local intention and local capability to effect change are relevant.
- The locus of the change is relevant-one should expect change where there has been SI activity.
- It may be worth looking for "fingerprints" that one associates with ESR and not with other sorts of changes which affect education.

"Fingerprints" might be nonrandom change associated with SI sites. For example, there may be observed change in SMET subject areas, but not in other subject areas; perhaps weaker effects away from centers of change (e.g., as "cascade" models fail progressively).

The overall lesson from seismology is that a variety of sources of evidence needs to be brought to bear on the problem of attributing causality. Data, such as changes in the profile of student performance, need to be understood in the context of some interpretative framework, and conclusions need to be drawn on the basis of plausible inferences that relate data and theory.

Strategy 3: Learn from Failures

Fast prototyping and testing is a characteristic of successful research and development activities. For example, some research and development groups have sayings such as "ready, fire, aim"; "fail forward"; "fail fast, fail often."

In many areas of engineering, a great deal of effort is devoted to the study of failures of working systems (e.g., Levy & Salvadori, 1992). Disasters involving aircraft, power stations, bridges, cars, etc., are followed by detailed analysis and often by changes in legislation that governs safe working practices. Specific failures often can be viewed as individual symptoms of broader system failures (e.g., Fortune & Peters, 1995). In contrast, in education current traditions of focusing almost entirely on positive results provide a poor strategy for understanding the phenomena, or for theory testing and building. There is an urgent need to learn from current failures. For example, NISE might provide a strong lead by convening a conference on SI "effects" where participants are constrained to spend as much time describing what they have learned from failures as they spend describing seemingly positive effects.

In the case of some NSF-funded SIs, funding has been discontinued. It would be worth analyzing these SIs in some detail. If they have "failed" for reasons to do with implementation strategy rather than outside political influences, the knowledge about these failed strategies can be extremely valuable. Knowing what does not work can be as important as knowing what does work. The benefits to the evaluation community are twofold. The direct benefit comes from the search for informative indicators. For example, what performance indicators (with hindsight) give clear evidence that an SI is failing? The long-term benefit is a contribution to the emerging body of knowledge about what makes for successful and unsuccessful SIs.

Strategy 4: Spend Most Time Looking at the Most Informative Evidence

In earth sciences there is a strong emphasis on the detection of big effects (e.g., volcanoes) which is relatively easy to do. However, the study of big effects by extensive recording around single sites is not easy to do. Clearly, one wants to focus one's resources where they are likely to do most good. Distributing data gathering evenly across all possible sites is unlikely to be optimal.

Physicists face a number of problems when conducting experiments in particle physics. These relate to the total volume of data and the rate of data flow. In a typical experiment, a beam of particles will be directed at another beam, or at a stationery object, such as an atom. The purpose of the experiment is to cause a collision in which interesting things happen-atoms or particles might be split, for example. The occurrence of such events is relatively rare, so there are considerable advantages to be gained by developing methods that allow the experimenters to record information from a small set of events that are likely to contain interesting things. A number of experimental methods have been designed specifically to do this. One approach is to have detectors that record only the presence of particular particles. Another approach is to use such detectors as triggers that switch on a broader spectrum of recording devices.

These techniques illustrate two fundamental principles of scientific work on complex systems. First is the idea that one can best understand the dynamics of systems when they are undergoing dramatic changes. Second is that data are essentially infinite and the sooner one can eradicate irrelevant information from consideration the better.

The first principle suggests that detailed evaluation should be conducted on extreme cases. These might occur "naturally," as in the case of schools or school districts that perform exceptionally well or exceptionally badly (as in vulcanology). An alternative is to destablilize a school or school district deliberately and to observe the dynamics (as in physics). This is likely to require high levels of energy, in the form of added resources. To learn from the situation, a good deal of instrumentation is likely to be needed. Physics offers some suggestions here, too. One idea is to develop a set of indicators that are specialized to detect particular kinds of events. In an educational context, these might refer to events at a variety of layers in the system, such as changes in the behavior of school principals (e.g., fostering home-school links), changes in classroom practices (e.g., the introduction of collaborative working groups in science, or changes in student performance (e.g., improvements on tasks involving decimals). Some indicators already exist, while others will need to be invented.

At the level of detecting situations that ought to be investigated further, there are two distinct ways to acquire information. One is to locate schools that have been deemed to be failing; the other is to use published data on student attainment, such as those provided by state tests. State tests can have considerable weaknesses, for current purposes. The tests may not provide measures of academic performance that cover a broad set of SMET indicators or may not be relevant to new standards-related goals (see Ridgeway (1998) for a critique, and Schoenfeld, Burkhardt, Daro, Ridgway, Schwartz, & Wilcox (eds.) for approriate resources in mathematics). It might well be worthwhile establishing a Center (or a division within an existing Center such as NCES) with a special responsibility for detecting educational sites where extraordinary events are taking place.

Another useful idea that can be borrowed from physics is to make deliberate attempts to filter events of particular interest. Here, there are analogies with classroom observations. The key is to tailor the research instrument to study the specific phenomena of interest, such as student teacher interaction with males and females, classroom atmosphere, metacognitive remarks, or other occurrences. No attempt should be made to summarize "everything," a logically impossible task. Given the constructive nature of knowledge, new measures could be invented ad infinitum.

Strategy 5: Search for Big Effects, and Disseminate Them Quickly

The most useful sort of feedback an evaluator can provide in the early phases of a program is the rapid identification of large-scale effects. These can be large-scale effects that are positive or that are negative. Once they have been identified, then ways of permeating (or inhibiting) these effects throughout the system quickly can be sought. Techniques for ensuring that treatments do not suffer from dilution or corruption are beyond the scope of this monograph, but see Elmore (1996).

In New York, in 1983 there were 425 deaths attributed to AIDS. This rose steadily to 7,102 deaths in 1994. Chiasson (1997) reported that, in November 1995, deaths peaked at 20.9 per day, yet in November 1996 the number was 10.1 deaths per day. The number began to decline in March 1996, fell steeply over the summer and fall, and then leveled off. AIDS mortality fell for both sexes, all races, and all ages above 24 years old. Chiasson commented that the trend "appears to have occurred at a single moment in time starting around March 1996." There were about 20 deaths a day in January and February; by July they had fallen to 11.5 per day.

The new treatment is a cocktail of existing drugs that usually includes a protease inhibitor. This cocktail is capable of arresting virus growth in many patients and returning them to a better state of health than they have enjoyed for several years. The treatment costs more than $10,000 per year. Chiasson attributes the decline not just to the drugs, but also to a significant injection of funds to pay for the new treatments. In 1994, New York received $100 million through the Ryan White Care Act, compared with $44 million in the previous year. More patients could be treated with drugs and more patients had access to traditional treatments for the diseases that killed AIDS patients, such as pneumonia.

New York City's health department collects birth and death records on its own residents. In other places, a common route is for state health departments to collect data and then pass these data on to city health departments. News of the effectiveness of the triple drug therapy was learned early in New York City, illustrating the effectiveness of fast data collection, analysis, and dissemination.

Another example from epidemiology of rapid dissemination is provided by the Centers for Disease Control (CDC). CDC uses a number of channels for the rapid dissemination of information. For example, the Morbidity and Mortality Weekly Reports contains recent data on morbidity and mortality, and a daily summary of news clippings relevant to CDC is published and is available throughout the organization.

Strategy 6: Use Systems Models as Part of Design Evaluation

Relevant groups (states, urban areas, etc.) made bids to receive SI funding and submitted plans for their work. Such plans can be evaluated using a systemic framework. The evaluation of plans must address a number of key areas that relate to the "systemicness" of the plans that are proposed. These key areas cover:

A description of the existing system:

- human and physical systems;
- deployment of resources (Where are existing resources being spent?);
- demographics of students and teachers;
- existing assessment schemes and associated performance data.

Identification of areas of current system dysfunctionality:

- those inherited from conflicts among federal and national programs;
- local problems.

An account of the changes proposed:

- some schematic representation of key system functions and their interrelations;
- a description of the intended curriculum;
- a description of how the conceptual gaps among the intended curriculum, the implemented curriculum, and the attained curriculum will be addressed;
- predictions of the time scale over which different effects of the reform might be expected to emerge (e.g., When can improvements in student performance be expected?).

An account of management issues:

- a description of the plans for system monitoring, and the scope for corrective action.

When the amount of money being spent on each SI is compared with the amount of money being spent on the educational system to which it is being applied, it is quite clear that the money must be used as a catalyst, not as a primary source of energy for there to be any significant change. It follows that SIs should make it clear just how SI funds will be used to steer the educational system. It follows that there needs to be a justification of the changes that are proposed, including:

- an account of leverage issues and multiplier effects—what is planned? how are they to be observed?;
- the expected cross-impact matrix of planned changes.

Figure 5 shows a "model" of systemic reform produced by the SRI evaluation of Statewide Systemic Initiatives that provides a clear representation of a number of key elements in the education system. The model has two components. It identifies an education system as comprising two blocks, one of states, regions, and districts, and the other of schools, classrooms, and teachers. Each block has a primary responsibility for certain kinds of actions (e.g., incentives for reform; classroom experiences). The two blocks also influence each other. The model shows that SIs can influence either or both components of the system.

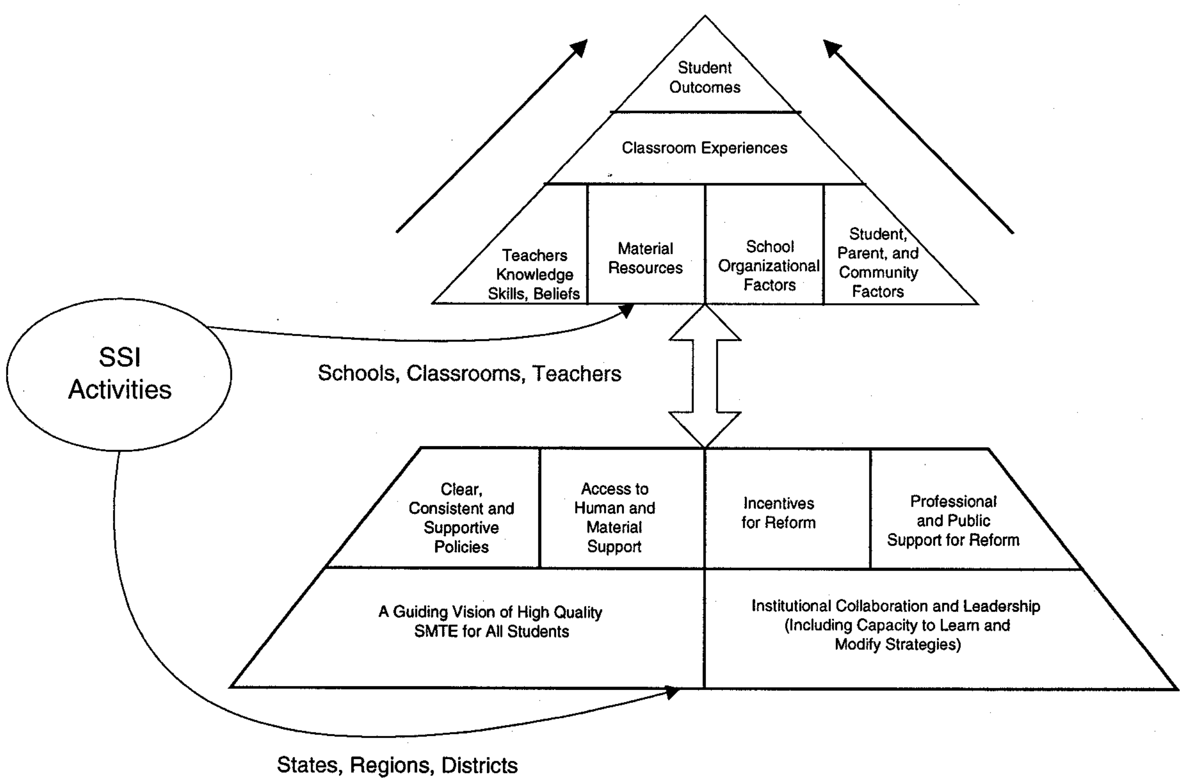

Figure 5. ***A "model" of systemic reform***

Source: Zucker, A., & Shields, P. (1997). *Evaluation of NSF's Statewide Systemic Initiatives Program.* Arlington, VA:SRI International.

This model is useful because it identifies key elements that SIs should take into account. Its primary use (Zucker & Shields, 1997 and personal communication with Zucker and Shields) was to help SIs articulate their approaches to ESR by, for example, mapping the locus of their major efforts onto the diagram, and talking about the ways that the identified system components were involved in the SI. The role of the model is to support a discussion and the evolution of ideas. It is to focus attention on the educational manipulations proposed in the plans to see whether they are consistent with the available evidence. It is not a model in the sense of the combined gas laws, or in the sense of SEIR in epidemiology, or in the sense of an ecologist's plan for prairie restoration.

One can conceive of a continuum of models that range from the SRI model (actually a useful representation that can be the basis for eliciting models stated verbally), through the "box and arrow" models favored by systems modelers such as Checkland (1981) and Banathy (1992), through to fully implemented systems models such as the SEIR model from epidemiology. Fully implemented models are challenging to create, because they require elements to be specified, along with the connecting links, and an account of the functions that relate variables to each other. Given the constraints on time and resources, the creation of a computer simulation of an entire SI at a particular moment in time would be quite unrealistic, and probably useless, given that SIs are constantly undergoing changes. However, the act of attempting to create partial models can serve a useful function in clarifying one's conceptions of the SI design.

Strategy 7: Attend to Parameter Estimation and Model Small Parts of the System

In the case of AIDS in New York, effective treatments could be repeated on each client group, given adequate quality control in the production of the treatments. The size of the effect appears to be dramatic. The good news for AIDS patients might not be good news for the healthy citizens of New York, for two related reasons. First is the cost. As patients are kept alive longer, the costs of care increase linearly with a very steep slope. Consider the crudest of models. Drugs for one hundred AIDS patients cost $1 million per year. One year's 3,500 saved lives added $35 million to that year's costs. If the trend continues, the costs will be increased by a further $70 million in the next year. A second problem relates to epidemiology. Increasing the number of cured patients is a good thing. Increasing the number of infectious people in the population is not. There is an interesting set of questions about how infectious patients are on new drug treatments and how much exposure the noninfected population suffers.

By now, evidence should be available within state SIs both about big effects and about some of the key parameters for modeling ESR, for example, the length of time required to produce changes. Constance Barsky (personal communication) reports that data from the Ohio SI—where a major goal is to help teachers incorporate significant amounts of discovery learning into their teaching styles-suggest that science teachers needed about 120 hours of professional development before visible differences in classroom behavior emerged after about 3 years. It is easy to use these data to calculate the costs of going to scale (Elmore, 1996) across the state.

One need not depend on computer models for systems modeling. Weiss (1997) gives an example of an SI whose main method of bringing about more investigative methods in elementary science was to have practicing scientists teach demonstration lessons in class. A total pool of 500 scientists was identified, all of whom were prepared to volunteer some of their time, at no cost to the project. There were several thousand teachers, but no analysis had been done at an early stage of the success or failure of the demonstration lessons. Also, no estimate was made of the number of visits required per teacher or the nature of the interventions that are effective in changing teacher behavior. One does not need a full simulation of the approach on a computer in order to decide that the model won't work. It is enough to do a "thought experiment" and ask about the model and its likely effectiveness to decide that it would not produce the desired results.

Strategy 8: Construct Macro-Systemic Models

If the notion of macro-systemic change is to be treated seriously, there needs to be an active research program that sets out to identify stages, possible transitions between stages, and the mechanisms whereby these transitions can be brought about. From the viewpoint of evaluation, such knowledge is important for the design of evaluation:

- How is change conceived?
- What stages are envisaged?
- How will they be recognized?
- What causal agencies will be deployed?
- What tools will be used to identify the current stage of development?
- What stages are found in practice?
- What transitions are possible between stages?
- What factors are relevant at each stage, and what precipitates the evolution of the school, the department, and the individual teacher?
- What mechanisms are in place to support a learning community (e.g. to gather evidence about classroom effects in order to inform policy, and to inform teachers and key change agents about what is effective)?
- How will information about possible stages be disseminated?

Ridgway and Passey (1995) describe a macro-systemic model of the development of computer use in schools. The model was derived from three sources of information: case histories of the evolution of computer use in individual schools, aggregation of patterns across "snapshots" of schools, and logical analysis. The model is macro-systemic because it identifies a number of stages of development and the need for different organizational features, and different behaviors, to be put in place at different stages. A simple representation of the model is shown in Figure 6.

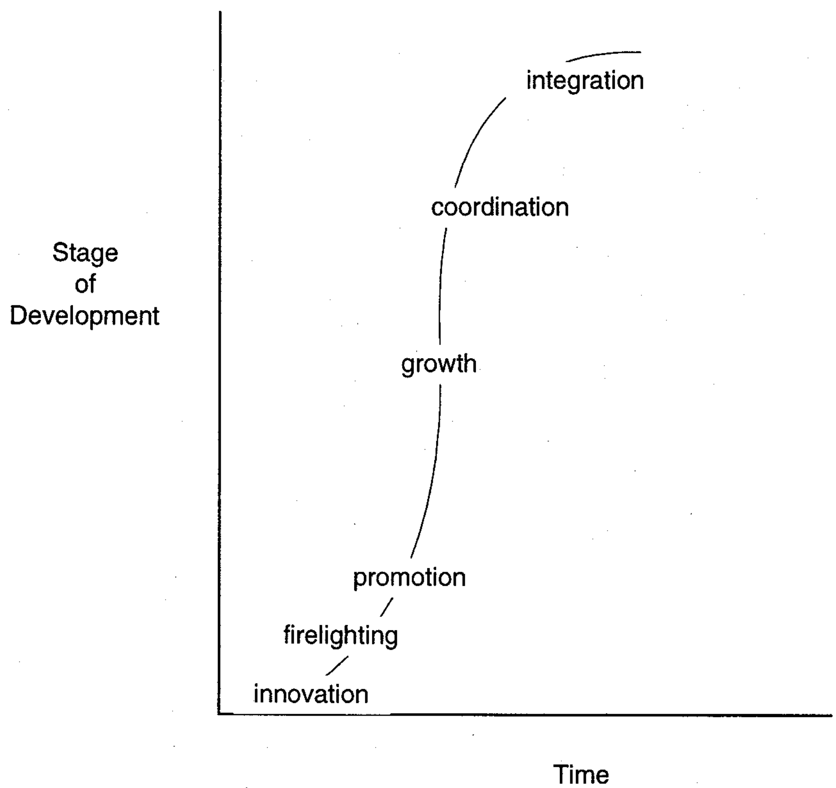

Figure 6. ***Stages of school development when using computers***

In the early stages of development—Innovation and Firelighting—the responsibility for the innovation rests with a few individuals, who make all the decisions about equipment provision, software, curricular ambitions, and their own professional development. When Promotion and Growth occur, school management need to be involved because of the implications for professional development and the need for a large increase in computer provision. At the stage of co-ordination, there is a need to ensure curriculum coherence from the viewpoint of students; provide technical support, equipment and routine maintenance; define school policy on software provision; establish ways to record student progress across their educational careers. Illustrations of the macro-systemic nature of the development are shown in Table 6.

Table 6
Macro-systemic stages in the evolution of schools' computer use

Factors	Stages of Development		
	Innovation	Promotion	Coordination
Persons responsible for computer use	individuals	supportive managers	the majority of teachers
Focus group for development	self-motivated individuals	departmental groups	whole school
Teaching styles	specified by an expert teacher	explored by small numbers of teachers	wide variety, known about within and across departments
Assessment and recording of student capability using computers	done by one person, if at all	some information about students shared across staff	schemes are in place; teachers understand what is to be learned and how this can be recorded

Review

This section began by drawing attention to the evidence base needed to support the reform movement. It argued that systemic reform necessarily must include a fuller integration of evidence from current evaluations into the processes of SR. SR is a new venture, at a relatively early stage; a great deal of information has been gathered by different evaluators in different places about the phenomena and the effects of different educational treatments. The evaluation community faces challenges in assembling this distributed wisdom in such a way that it can be shared and made useful to the community at large. Several ideas are proposed for knowledge sharing. Some suggestions are made about how systems models and macro-systemic models can provide intellectual frameworks to support SR.

Concluding Remarks

This monograph has offered a classification of the approaches to modeling science in terms of analytic, systemic, and macro-systemic styles. It has argued that the shift from an analytic style to a systemic style in education constitutes a paradigm shift (albeit one that encompasses all that has gone before) that is sufficiently great to justify a reconceptualization of the process of evaluation. Systemic reform actually requires a deep analysis of the processes of macro-systemic reform if it is to be successful. The "climax community" that matches current educational ambitions is unlikely to be attainable from the current educational system in a single systemic jump. Interim stages, which might have some temporary stability, need to be considered.

Scriven (1993) argued that evaluation should be established as a transdisciplinary subject, rather like statistics and logic. Wilson et al. (1996) argue that there is a pressing need for an intellectual community to emerge that addresses the issues of the management and evaluation of systems undergoing change. These views are endorsed strongly here. There is an urgent need to develop ways to share information around the evaluation community and to support the emerging field.

This monograph set out to review disciplines outside education to look for ideas that might inform the evaluation of systemic reform. A number of conclusions can be drawn.

1. Systemic reform has been adopted as if it were a natural extension of existing knowledge in education. A case is made that, while enough is known to support the design and evaluation of each individual SI, a new general field of inquiry needs to be promoted to support the evaluation of systemic reform in general, because of the need to treat systemic and macro-systemic issues seriously.
2. Educational reform should devote more attention to systems and macro-systemic modeling since these are closer to the core ideas of systemic change. Evaluators need appropriate methods of judging whether such models are in place and how well they have been designed and implemented.
3. Ecological restoration is a generic example of macro-systemic reform. The knowledge base needed to engage in systemic reform and the evaluation of systemic reform has a parallel with ecological restoration in terms of the need for:

    - clear definitions of desirable end points;
    - ways to describe critical aspects of educational systems;
    - knowledge of the timeline of different sorts of development;
    - knowledge of the conditions that need to be established before certain kinds of growth can occur;
    - knowledge of transition states that are necessary and sufficient to reach desirable end points from particular starting points;
    - ways to recognize undesirable developments, and knowledge of how to eradicate them;
    - ways to communicate effectively to stakeholders about the time lines of macro-systemic change.

4. A set of vignettes from sciences that face the same problems as those faced in education illustrate some of the research methods that might be used to build the requisite knowledge base. These research methods are important to construct the evidence on which the evaluation of systemic reform (as an entire program) and systemic initiatives (as individual case studies) can be based.

A key issue for the evaluation community is how the knowledge base relevant to research questions is assembled, stored, and accessed by the relevant communities. There is a clear role to be played by some coordinating group such as NISE.

References

Asch, D., & Bowman, C. (1989). *Readings in strategic management.* London: Macmillan and Open University Press.

Baddeley, A. D. (1976). *The psychology of memory.* New York: Basic Books.

Banathy, B. (1992). *A systems view ofeducation.* Englewood Cliffs, NJ: Educational Technology Publications.

Beer, S. (1976). *Platform for change. New* York: Wiley.

Bertalanffy, L. von (1968). *General systems theory.* New York: Braziller.

Breckenridge, J. S., Goldstein, D. S., Zucker, A. A., & Adelman, N. E. (1996). *Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program. The impact on students of the SSI Program: A pilot study of the impacts of the Louisiana and Montana SSIs* (Draft). Menlo Park, CA: SRI International.

Centers for Disease Control and Prevention. *(1997). Morbidity and mortality weekly reports.* Atlanta, GA: Author.

Centers for Disease Control and Prevention. (1997). *Today's news.* Atlanta, GA: Author.

Checkland, P. (1981). *Systems thinking, systems practice.* New York: Wiley.

Chiasson, A. (1997, January 25). Report presented at the Fourth Conference on retroviruses and opportunistic infections. *The Washington Post,* p. 1.

Clune, W. H., Millar, S. B., Raizen, S. A., Webb, N. L., Bowcock, D. C., Britton, E. D., Gunter, R. L., & Mesquita, R. (1998). *Research on systemic reform: What have we learned? What do we need to know? Synthesis of the second annual NISE Forum, (Workshop Report No.4).* Madison: University of Wisconsin, Wisconsin Center for Education Research.

Curtis, J. (197 1). *Vegetation of Wisconsin: An ordination of plant communities.* Madison: University of Wisconsin Press.

Elmore, R. (1996). Getting to scale with good educational practice. *Harvard Educational Review, 66*(1), l-26.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh: Oliver and Boyd.

Fortune, J., & Peters, G. (1995). *Learning from failure: The systems approach.* New York: Wiley.

Girgerenzer, G. (1992). Discovery in cognitive psychology: New tools inspire new theories. *Science in Context, 5*, 329-350.

Goertz, M., Floden, R., & O'Day, J. (1995). *Studies of education reform: Systemic reform. Vol.1: Findings and conclusions* (CPRE Report No. RR-035A). Philadelphia: University of Pennsylvania.

Government Performance and Results Act (GPRA) of 1993 (PL 103-62).

Grouws, D. A. (1992). *Handbook of research on mathematics teaching and learning. New* York: Macmillan.

Hacking, I. (1990). *The taming of chance.* Cambridge: Cambridge University Press.

Heck, D. J., & Webb, N. L. (1998). *Purposes and issues of systemic evaluation in education* (Research Monograph No. 8). Madison: University of Wisconsin-Madison, National Institute for Science Education.

High Performance Systems. (1990). STELLA II. Hanover, NH: Author.

Lotka, A. J. (1956). *Elements of mathematical biology. New* York: Dover.

Kanter, R. M. *(1984). The change musters.* London: Routledge.

Kelley, J. (1997). Arboretum research: determining the anatomy of a savannah. *NewsLeaf.* Friends of the Arboretum Newsletter; The University of Wisconsin-Madison Arboretum, March, 2, p. 23.

Knapp, M. S. (1997). Between systemic reform and the mathematics and science curriculum: The dynamics of implementation, and professional learning. *Review of Educational Research, 67*(2), 227-266.'

Leopold, A. (1949). *Sand County almanac.* Oxford, Eng: Oxford University Press.
Levy, M., & Salvadori, M. (1992). Why *buildings fall down. New* York: Norton.
Manski, *C.* F. (1995). *Identification problems in* the *social sciences.* Cambridge, MA: Harvard University Press.
Massell, D., Kirst, M., & Hoppe, M. (1997). *Persistence and change: Standards-based systemic reform in nine states,* (CPRE Policy Brief RB-21). Philadelphia: University of Pennsylvania.
Murray, M.F. (1993). *Prairie restoration for Wisconsin* schools. Madison: University of Wisconsin-Madison Arboretum.
National Center for Education Statistics. (1994). *NAEP 1994 trends in academic progress.* Washington DC: U.S. Department of Education.
Piaget, J. (1929). *The child's conception of the world.* New York: Harcourt Brace Javanovich.
Ridgway, J. (1987). *A review of mathematics tests.* Slough: NFER-Nelson.
Ridgway, J. (1998). *From barrier to lever: Revising roles for assessment in mathematics education,* (NISE Brief). Madison: University of Wisconsin-Madison, National Institute for Science Education.
Ridgway, J. (1997). Vygotsky, informatics capability, and professional development. In D. Passey & B. Samways (Eds.), *Information technology: Supporting change through teacher education* (pp. 3-19). London: Chapman Hall.
Ridgway, J., & Passey, D. (1995). Integrating IT: How to use existing evidence about individual teacher evolution to plan systemic revolution. In D. Tinsley & D. Watson (Eds.), *Integrating information technology into education* (pp. 59-72). London: Blackwell.
Ridgway, J., & Schoenfeld, A. H. (1994). *Balanced assessment: Designing assessment schemes to promote desirable change in mathematics education.* Keynote paper for the EARL1 Email Conference on Assessment.
Rodwell, J. S. (Ed.). (1991-1995). *British plant communities* (4 vols.). Cambridge, Eng.: Cambridge University Press.
Schoenfeld, A., Burkhardt, H., Daro, P., Ridgway, J., Schwartz, J., & Wilcox, *S.* (Eds.). (1998). *Balanced Assessment for the Mathematics Curriculum: Elementary Assessment (2* vols.)*; Middle School Assessment (2* vols.)*; High School Assessment (2* vols.)*; Advanced High School Assessment (2* vols.). White Plains, NY: Dale Seymore.
Scriven, M. (1993). *Hard-won lessons in program evaluation* (American Evaluation Association Monograph Series, New Directions for Program Evaluation, Number 58). San Francisco: Jossey-Bass.
Shakley, S., & Wynne, B. (1996). Representing uncertainty in global climate change science and policy; Boundary-ordering devices and authority. *Science, Technology and Human Values, 21*(3), *275*-302.
Skinner, B. F. (1953). *Science and human behavior.* New York: Macmillan.
Sterling, T.D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance or vice versa. *Journal of the American Statistical Association,* 54, 30-34. (cited in Girgerenzer, 1992).
Stevens, F., Lawrenz, F., & Sharp, L. (1993). *User-friendly handbook for project evaluation: Science, mathematics, engineering and technology education* (NSF 93-152). Arlington, VA: National Science Foundation.
Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.
Tufte, E. R. (1990). *Envisioning information.* Cheshire, CT: Graphics Press.
Tufte, E. R. (1997). *Visual explanations.* Cheshire, CT: Graphics Press.
Vygotsky, L. S. (1981). The genesis of higher mental functions. In J. Wertsh (Ed.), *The concept of activity in Sovietpsychology* (pp. 144-188). New York: Sharpe.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph No. 6). Madison: University of Wisconsin-Madison, National Institute for Science Education.

Weiss, I. (1997). *The role of evaluation in systemic reform* (Workshop Report No. 4, Vol. 2, pp. 73-75). Madison: University of Wisconsin-Madison, National Institute for Science Education.

Wilson, D.C. *(1992). A strategy* ***of*** *change.* London; Routledge.

Wilson, K. G. (1994). Wisdom-centered learning: Striking a new paradigm for education. *The School Administrator, 26-33.*

Wilson, K. G., Barsky, C. K., & Davis, B. (1996). A model development concept (MDC) for education. A framework for change. Unpublished manuscript, draft 3.

Zawojewski, J.S., Hoover, M.N., & Ridgway, J.E. (1997). *Analysis* ***of*** *student performance on two assessment instruments in the connected mathematics project.* Paper presented at the annual meeting of the American Educational Research Association.

Zucker, A., & Shields, P. (1997). *Evaluation* ***of*** *the NSF's SSI program.* Paper presented at the NISE Conference on Systemic Evaluation, University of Wisconsin-Madison, March.

Single copy price is *$7.25.* To order copies contact:

CENTER DOCUMENT SERVICE
Wisconsin Center for Education Research
1025 W. Johnson St., Room 242
Madison, WI 53706-I 796
608/265-9698

NO PHONE ORDERS. PREPAYMENT REQUIRED FOR ORDERS UNDER $20.00.

Price is subject to change without notice.

National Institute for Science Education
University of Wisconsin-Madison
1025 West Johnson Street
Madison, WI 53706

(608) 263-9250
(608) 262-7428 fax
niseinfo@ macc.wisc.edu
http://www.wcer.wisc.edu/nise