Unexpected Behaviors and Outcomes in STEM: Advancing Methodologies for Detection and Evaluation Response

Jonny Morell
j-morell@fulcrum-corp.com
Chris Coryn
chris.coryn@wmich.edu
Daniela Schroeter
daniela.schroeter@wmich.edu

Powerful Evaluation, Unpredictable Programs

Seldom do programs work out as planned. Their content may change. Their implementation processes may change. Their outcomes may vary from the intended. Unanticipated outcomes may emerge. In the face of this flux, how can powerful evaluation methodology be deployed to evaluate STEM programs?

> An example: STEM illustration of methodology in the face of program change:
>
> Imagine a program that was designed to help high school science teachers draw from material across many different scientific disciplines, and to transmit that material by setting students to work in small groups. It is not hard to imagine that beyond the obvious outcomes of better learning, a few other outcomes may arise. The teachers themselves may begin to self-organize into informal interest groups which may result in unanticipated collaborations. Or, the cross disciplinary approach in science may spur students to apply such thinking across their studies in literature and history, which would in turn spur their literature and history teachers to collaborate. Of course it is easy to play Monday morning quarterback and say; "Of course these kinds of changes should have been anticipated. If they were not included in the evaluation design, it means that the program planners and evaluators did a poor job." But the truth is that these unanticipated outcomes are but two of many possibilities. Maybe the real unanticipated change was an increase in students' interest in school, or excitement on the part of their parents which translated into more involvement in school affairs? It is easy to spin numerous other possibilities, all of which would seem obvious in hindsight. None of these unexpected changes would be problematic for evaluation if the design relied exclusively on open ended, rolling interviewing of teachers and students, with no inclusion of any comparison groups or pre-implementation data. But what would be lost in such a design? There would be no comparison of whether parents' involvement was changing as a result of other factors (e.g., a budget crisis that threatened cherished programs). There would be no way to tease out the reasons for increased teacher collaboration in the face of other programs and professional development activities that are ubiquitous in school systems. There would be no provision to determine whether history teachers noticed the change in their students approach to problem solving, or whether that translated into changes in students' grades.

The goal of this project is to expand the evaluation toolkit as broadly as possible, to make evaluation as powerful as possible, and still maintain its ability to adapt to unexpected program behavior. The essential difficulty in meeting this goal is that powerful evaluation requires that an infrastructure be implemented and maintained over time. To illustrate with just a few of many possible examples:

- There may be a need to develop validated instruments.
- There may be a need to negotiate for access to specific control groups or for archival data.

- There may be a need for processes that assure data collection during narrow time frames.

In each of these examples there is rigidity in the evaluation design. Validated instruments cannot be cheaply, quickly, or easily revised. It can be time and labor intensive to get access to parts of an organization for control purposes. Renegotiating for access to different parts of the organization is to say the least, nontrivial. Similarly, agreements to access one set of sensitive data may not carry over to other sets of data. Mechanisms that allow access to people during one time period may not translate into permission to talk to them at other time periods. Contrast these inflexible design elements to an evaluation design that relied solely on qualitative, posttest only interviewing with no comparison groups. A qualitative posttest design of this type is infinitely flexible and may be perfectly acceptable under certain circumstances. But to restrict oneself to reliance on designs like this is to hobble evaluation.

Review Criteria

Intellectual Merit and Broader Impact

The reasons for unexpected change in program behavior are well understood and have been thoroughly explored. The story has been told for university research (Behrens and Gray 2001), marketing (Fry and Polonsky 2004), tobacco restrictions (Hoek 2004), drinking age regulation (DiNardo and Lemieux 2001), speed and quality relationships in new product development (Lukas and Menon 2004), welfare (Courtney, Needell et al. 2004), national fiscal reform (Kildegaard 2001), teacher empowerment (Pugh and Zhao 2003), nongovernmental organization activity in developing countries (Stiles 2002), and workplace safety (Kaminski 2001). In addition to these domain-specific explanations, more general analyses based on the dynamics of open systems can be found, as for example in the works of (Dorner 1996), and (Tenner 1996).

What is less understood is how evaluators should deal with unexpected change. Bridging the gap from understanding unexpected change in programs to evaluation's response to those changes is weak from theoretical and practical points of view. Evaluators most certainly make that leap, and more often than not do quite a good job of it. But they do so only when forced, when the need springs upon them. Better responses would come from an intellectual understanding of how evaluations and programs work together, a practical understanding of all the tools that are available to them, and how a small number of particular tools should be chosen. The proposed research is designed to advance this intellectual understanding, and to do so in a way that supports both our theoretical understanding of evaluation, and our practical understanding of how evaluation theory can be applied in the service of developing evaluations that can describe, explain, and predict. Because STEM is the test setting, STEM programs will benefit directly from the research. STEM programs, however, are the same as any program in that they are assemblages of people and resources, set within political, organizational and economic contexts, and organized to achieve particular purposes. As such, lessons learned from this project are certainly going to be relevant in many other evaluation settings as well.

Integration of Research and Education

Graduate student involvement is a critical element in this proposal. Graduate students will be trained in the innovative evaluation approaches that are the subject of this proposal, will work closely with the developers of those approaches, will provide consultation to the STEM test sites, and will be deeply involved in data collection, analysis, and interpretation.

Integrating Diversity into the Project

While specific efforts to target particular ethnic or social populations are not part of this proposal, the analysis plan and the dissemination plans are designed to reach as broad a range of people as possible. This is because social networking plays a critical role in collaboration among study participants. While data collection and analysis are designed to be self contained within those participants, rigorous social networking efforts are planned to include as many other interested parties as possible. Details of these efforts are described later, in the section" STEM and Evaluation Advisory Groups: Continuous Feedback Through Social Networking"

The Theoretical Foundation

The proposed work is based on an approach to evaluation that has been developed (and continues to be developed) by Dr. Jonathan A. Morell, one of the investigators named in this proposal. Dr. Morell's approach begins with the belief that while the reasons for unexpected program behavior are well known (complex system behavior), evaluators' responses to those surprises are underdeveloped. At present evaluators react to change in a fire fighting mode. When surprised, they use whatever skill they have to deal with the situation. Those responses are often quite adequate. Adequate – but not optimal. In order to better respond to change, the evaluation field needs to step back from firefighting mode. The field needs to look at unexpected behavior in a systematic manner. They need to set evaluation behavior within a theoretical context of organizational and system behavior, and to use that understanding to develop and deploy tactics that will either anticipate change, or respond to it. To succeed, a community of interest needs to be developed within the evaluation community that will move thinking about the topic beyond the few people who are actively involved at present.

These ideas are set out in two publications: "Why are there unintended consequences of program action, and what are the implications for doing evaluation?"(Morell 2005), and *Evaluation in the Face of Uncertainty: Anticipating Surprise and Responding to the Inevitable (*Morell 2010). As developed at the time of this writing, the general approach is best understood in terms of three broad topics: (1) a continuum ranging from surprises that might have been reasonably anticipated, to surprises that are impossible to ever anticipate; (2) methodologies and tactics that are differentially useful along that continuum; and (3) guidelines for choosing tactics carefully in light of the fact that any corrective action may induce its own problems in the system.

Continuum

Figure 1: Continuum and Tactics

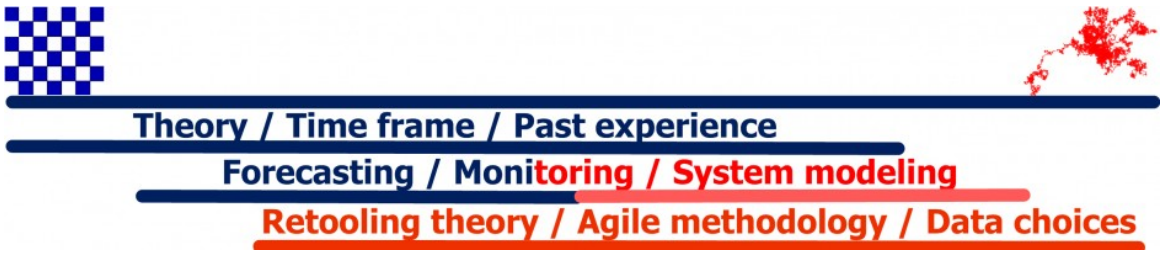


Figure 1 is the overview image used in workshops and presentations as a visual aid to explaining the theory. The extreme left represents scenarios in which the proper tactics could be reasonably expected to

provide a workable view of what might transpire. At the right are changes that for theoretical reasons, are impossible to foresee. (Those reasons are rooted in complex system behavior, complete with feedback loops of varying latencies, multiple interacting agents, self-organized behavior, emergence, and shifting fitness landscapes.) The vast middle presents those scenarios where foresight continues to dim as one moves from left to right. In the middle region the concept of "lead time" is critical because the greater the lead time in detecting incipient program change, the greater the opportunity to retool the evaluation and maintain the use of powerful methodologies. As surprise becomes inevitable, tactics for "agile methodology" come into play. These are design elements that are built into an evaluation design at the beginning, and which have the potential to be applicable within a broad envelope of program performance. Of course Figure 1 is simplistic and incorrect because it is based on two related erroneous assumptions. First, that in any situation it is possible to identify most of the powerful forces acting on a program to influence its behavior. Second, that there are real life program scenarios in which it is reasonable to exclude the possibility of complex system behavior. Still, there are several reasons why the approach depicted in Figure 1 works as a useful guide to action. First, it does serve to focus people's attention on the nature of their programs and the capacity of their evaluation designs. Second, there really are sets of unanticipated program behaviors that could have been reasonably foreseen, had the right methods been used. Third, the method proposes the use of multiple tactics spread across the continuum.

Tactics

Figure 1 provides a general sense of tactics that are particularly useful along the continuum from foreseeable to unforeseeable change. A complete explanation of what these tactics are, their intellectual foundations, and why they are recommended for particular purposes, is laid out in detail by Morell (Morell 2005; Morell 2010), and is beyond the scope of this present document. A brief overview however, will help to provide a sense of how the overall approach is applied.

- "Theory" and "past experience" both refer to the general idea that a great deal of knowledge can be brought to bear on understanding how a planned program might behave. It seems almost silly to provide this kind of advice, as it seems so obvious. On the other hand, it is so frequently not followed that it is worth pointing out. A major question though, is "what sources of knowledge should be tapped?" After all, it is one thing to advise the use of diverse bodies of knowledge, it is quite something else to choose a few from the multitude available. The solution proposed is to apply a small number of the bodies of knowledge most closely allied to the program at hand, and to rotate through others.
- "Time frame" refers to the notion that the further the predictions in time, the greater the likelihood of unintended consequences. Here too, the statement seems obvious. What is not obvious is how to pick time frames that are programmatically and politically meaningful, and still within the bounds of some acceptable level of predictability.
- At the far right of the continuum (the region where complex system behavior makes prediction impossible), an array of tactics are deployed that go under the general rubric of "agile methodology". These tactics include systematic efforts to retool program theory as a guide to data collection and interpretation, redundant and overlapping data sources to compensate for the loss of any one, multiple methodologies, and a variety of related methods.

The most interesting part of the continuum is the vast middle. There, surprise becomes inevitable, but methods are available for increasing the lead time between the detection of incipient change, and its

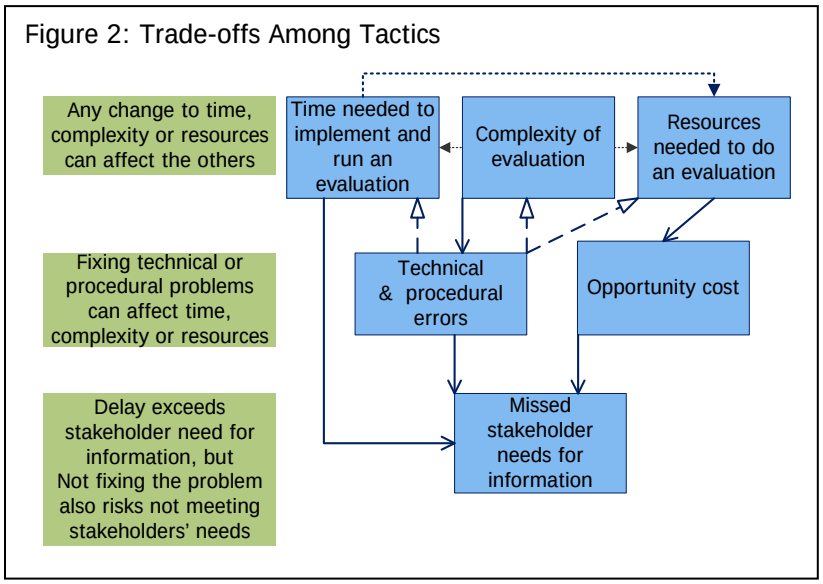
occurrence in a form that affects evaluation. Lead time is critical because the longer it is, the greater the possibilities for powerful changes to the evaluation design. Tactics that are useful in this part of the continuum are:

- Forecasting methods adapted from the field of Planning,
- Expanded forms of the "monitoring and evaluation" methods that are so popular in the evaluation of development and aid programs in developing nations, and
- Modeling approaches that are sensitive to shifts in the systems that constitute the fitness landscape of the program being evaluated.

Wise Choices

It is all well and good to suggest numerous ways to deal with evaluation surprise, but too much of a good thing is often a bad thing, and so it is here. The problem is that any technique implemented to deal with evaluation surprise might cause its own complications. For instance, it may be desirable to add records review to interviews with clients, thereby assuring data availability should the interviews fail to materialize. But adding the review adds time and cost to the evaluation, which may result in missing a window of opportunity to educate stakeholders, or in fewer resources for data analysis. Thus the evaluation approach to be tested includes a framework (Figure 2) to help identify why and how any proposed tactic may be counterproductive, and how wise judgments can be made.

Figure 2: Trade-offs Among Tactics



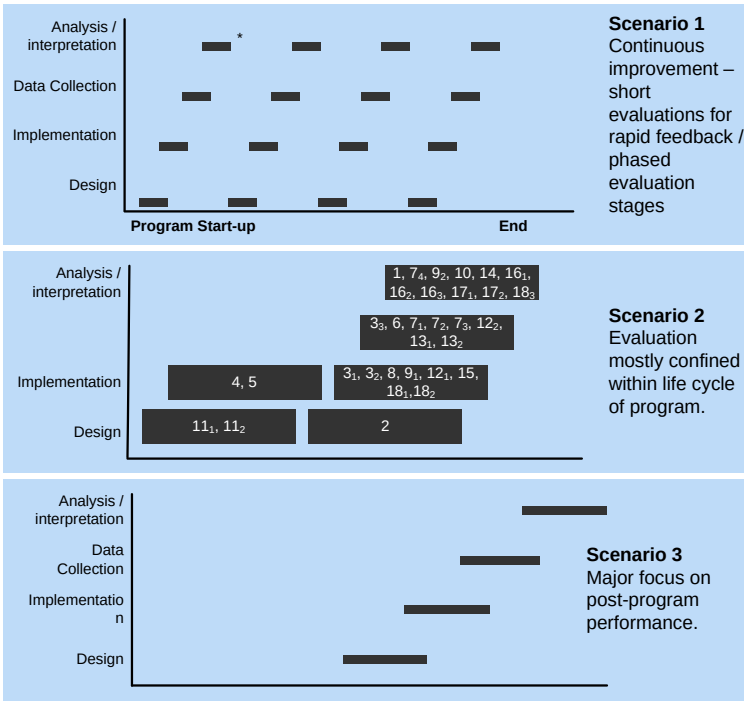
The Empirical Foundation

Underlying and supporting the proposed approach to evaluating unanticipated program behavior are a set of 18 case studies, comprising 33 unique instances of unexpected behavior. Among other findings from these case studies was that the same kinds of unexpected behaviors that affect programs also affect the evaluations of those programs. In retrospect of course, this finding is unsurprising. After all, as social entities, both programs and their evaluations are the same (i.e., collections of people and resources, organized for specific purposes, and set within an organizational / political / economic setting). As a result of this finding, the discourse around "unexpected program behavior" or "unintended outcomes" was shifted to the broader notion of "evaluation surprise".

Two a priori frameworks are used to organize the data. (See Figures 3 and 4 on the next page. Numbers in the figures represent placement of instances of surprise as determined by (Morell 2010)). As shown in Figure 3, a life cycle view was used to map the occurrence of surprise relative to the combined life cycle stages of programs and evaluations. The second framework (Figure 4) encompassed social/organizational factors, (e.g., whether an issue springs from internal program behavior or the program's environment) to place where surprise came from and what the evaluators in the cases did about them. In addition to these

frameworks, three other categorizations emerged from the data analysis: (1) pilot and feasibility tests; (2) resistance to evaluation; and (3) incorrect assumptions early in the evaluation life cycle. Taken together, these frameworks were used to shape the theory, to serve as guides as to where to look for sources of surprise, and to develop insight as to choose potential responses.

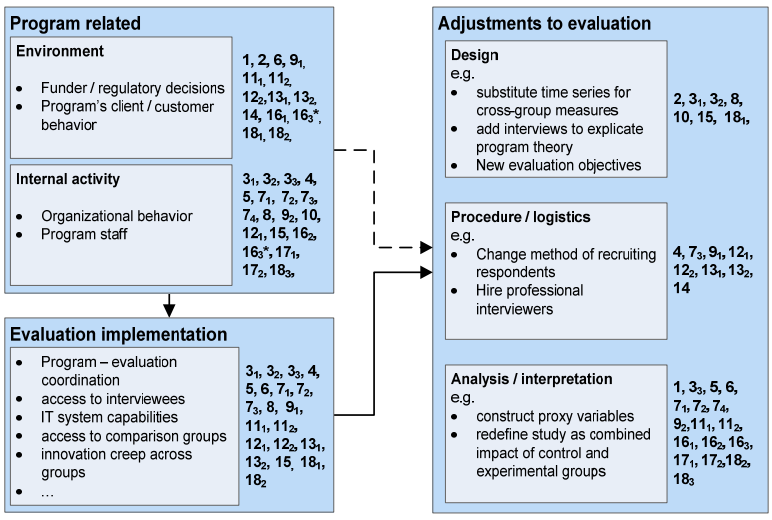
Figure 3: Program x Evaluation Life Cycle



The Innovation to be Studied

So far the theory and methods proposed above, and the supporting data, have been meeting with general acceptance in the evaluation community. However, none of these ideas have been tested prospectively; that is, applied to ongoing evaluations. We expect that were such a test to be conducted, both the theory and the tactics would be revised, and the overall goal of shifting the mode of dealing with evaluation surprise from firefighting to systematic action would be advanced. So too would the objective of developing a community of interest that would bring a greater variety of theoretical perspectives and empirical knowledge to bear on the subject. The design of the proposed research is to further these objectives by working within STEM settings. So doing would have direct impact on the evaluation of STEM programs, and indirect impact on the field of evaluation as a whole.

Figure 4:Social / Organizational View



The STEM Case Settings

The proposed research will be implemented in two settings in which active STEM activities are taking place. More than one site is critical because it is important to observe how the proposed evaluation methods work in diverse settings. More than two sites would be ideal, but budget constraints restrict us to only two. One site is the Center for Elementary Mathematics and Science Education (CEMSE) at the University of Chicago. The other is a set of five programs that are taking place within the Dayton, Ohio region.

Case 1. Building a Virtual Learning Community of Mathematics Educators: Center for Elementary Mathematics and Science Education (CEMSE) at the University of Chicago

The Center for Elementary Mathematics and Science Education (CEMSE) at the University of Chicago is currently working on a project to build a virtual learning community (VLC) for elementary mathematics educators. The project, entitled "Virtual Learning Communities: An Online Professional Development Resource for STEM Teachers," is in its first of three years of funding from the National Science Foundation's Discovery Research K-12 program. The aim of the project is to build an online space for teachers to watch lesson video, obtain and share resources, and communicate with each other and experts; the space will be built over the first two years of the project and launched in the third and final year.

The project directors are building the VLC to match what is known about effective professional development for teachers with the affordances of Web 2.0 platforms. Prior research, along with prior work by CEMSE staff, suggests that high-quality professional development draws on artifacts of teachers' practice, allows teachers to share ideas and support each others' improvement, and roots itself in specific instructional systems that teachers are using, such as curricula. Thus, the project team plans to design a VLC that intertwines (a) practical "learning objects" for teachers, including lesson video and instructional resources, (b) community-building tools that draw on the capabilities of Web 2.0 platforms, and (c) content that focuses on teacher improvement in the context of a specific elementary mathematics curriculum: *Everyday Mathematics*. The project directors plan to evaluate the success of this design in the third year of the project by measuring the sustained participation of VLC users and the growth in VLC users' "professional vision," a metric that is linked to positive changes in teachers' classroom practice.

The project team believes that the odds of success will be greatly improved if teachers are included as integral partners in the design and development of the VLC. Thus, the project team is currently working with 15 "teacher partners" to collect lesson video and other instructional resources and to develop the newly-constructed VLC interface. CEMSE's experienced staff of professional developers and mathematics education experts are editing collected resources for use on the VLC. The teacher partners will continue to provide resources and feedback for two years. In January, the project team will launch the site with a select group of some 100 teachers known as the "teacher panel," who will provide further feedback and analysis on the resources and interface until the VLC is launched in the third funding year.

The project team is interested in evaluating both the outcomes and the processes of this project. The outcomes it wishes to evaluate are (a) the sustained participation of VLC users, (b) the changes in professional vision of VLC users, and (c) the integrity of the VLC's three-part design. The processes it wishes to evaluate concern the ways that teachers were integrated into the project team and the way the project staff balanced teacher and expert feedback in the design of the VLC and its component resources.

Case 2. A constellation of STEM evaluations in Dayton, Ohio

In the Dayton, Ohio region there are five STEM activities of interest. The evaluation for the five efforts is headed by a Wright State University evaluation team. The evaluation for these efforts includes the impact on teachers' content knowledge, changes in attitudes towards teaching STEM fields, and changes in pedagogy and instructional strategies. For students, the evaluation includes performance on pre- and post- teacher developed content tests, standardized state and third party content tests, career interest development, post graduation plans, and performance post graduation. Details are below:

1 and 2) : Dayton STEM Center and NSF RET grant: K-12 STEM teachers will attend a 6 week summer institute that includes collaborative team work with participants, STEM professionals, pre-service teachers, undergraduate engineering students and engineering faculty. Regional STEM professionals will provide training and leadership for the 6 weeks immersion in problem solving. After the summer institute, participants will work with the Dayton STEM Center to develop curriculum based on their summer STEM experiences. The curriculum will be shared with regional schools via networking web sites; the participants will provide personal assistance to regions that request such assistance. In addition, participants will develop and administer pre- and post- student content tests for the curriculum developed. The goal is to measure student gains in the topics learned during the institute. Prior to the start of the summer experience and after the year at the Dayton STEM Center, participants will complete content knowledge tests reflecting attitudes towards teaching STEM fields, pedagogy and instructional strategies usedIn addition, participants will be observed during instruction to determine the level of reformed teaching present in their classrooms.

3) Dayton Regional School System (DRSS): There are a number of STEM evaluations within DRSS in progress and ongoing . Student performance indicators include the Ohio achievement grade level assessments, the NWEA MAPS, and the end of course ACT tests. In addition student displays of persistence, collaboration, inquiry and communication skills are being tracked ethnographically. Because research supports the importance of leadership for successful STEM schools, the Val-ED 360 leadership for Student Success protocol will be introduced in 2011. Student career interests and counseling efforts are tracked annually using the online KUDER career assessment tool. Also, in 2011, the means to track student successes and feedback after high schools will be implemented.

(4) State STEM Demonstration Grant. Recently state funding was awarded to support DRSS and Metro (STEM school in Columbus) to develop and implement common evaluation efforts to document the two schools' growth and progress. Even though both schools are STEM schools, their client base and missions are slightly different. The purpose of the state RFP was to encourage Ohio STEM school sites to share successful evaluation efforts between sites. The schools will apply the funding towards implementing common student performance assessments, common tracking of student career and post high school plans of graduates, common leadership studies, and common ethnographic studies for 21st century skills and school climate. The results will be compared for commonalities and uniqueness between the settings.

5) Improving Teacher Quality (ITQ) grants: Since 2006, WSU has provided summer science and math institutes for regional K-12 STEM teachers through Math Science Partnership (MSP) and ITQ

grants. The participants complete pre- and post- content and attitude surveys to track growth or change. Participants have been observed in the past using reformed teaching protocols to document the level of reformed teaching witnessed in the classrooms.

Collaboration and Dissemination Through Advisory Boards and Public Participation

Two dimensions of collaboration will characterize the execution of this project – executing organizations, and connection to the broader community by means of social networking.

Executing Organizations: Roles and Organizational Capabilities

Success will require an assembly of human capital and STEM activity that does not reside in any single organization. Hence we have assembled a team consisting of Western Michigan University, the Fulcrum Corporation, The University of Chicago, and the evaluation team responsible for the Dayton area STEM projects. Western Michigan University's involvement will include both its *Evaluation Center* and its *Interdisciplinary Doctoral Program in Evaluation*. Taken together, these entities represent a longstanding involvement in STEM evaluation, rich ties to the Educational Evaluation Standards, contributions to evaluation theory, doctoral student supervision (some of whom will be involved with the proposed work), and knowledge dissemination. (The Center is the home of the *Journal of MultiDisciplinary Evaluation*, and Center staff are very active in the American Evaluation Association.) Finally, the Center staff are well versed in working with the NSF and in running NSF-funded work.

Staff at the Fulcrum Corporation (in the person of Dr. Morell) are the driving force behind the evaluation theory that forms the core of this proposal. As editor of *Evaluation and Program Planning*, Dr. Morell is also in a good position to design outreach to the evaluation community. Over and above expertise in evaluation, Fulcrum has long experience providing STEM services to the Office of Naval Research and to the Federal Railroad Administration. Fulcrum is driving force behind ONR's SeaPerch program, and for an FRA Web platform to support workforce outreach. Both programs are targeted at parents, teachers, and students.

STEM and Evaluation Advisory Groups: Continuous Feedback Through Social Networking

The goals of both technical success and knowledge transfer will require guidance from a diverse group of people with expertise in evaluation and in STEM. To get that feedback, two advisory boards will be constituted, one of evaluators and one of STEM experts. (The notion of "two boards" is used for administrative and project management purposes. In reality all the experts will interact as a single group over the course of the entire project.) Because commitment is so important, members of these boards will be paid for their services. For maximum value, input from the boards should be continual and incremental. A social networking site will be used to assure the required pace of feedback.

No matter how committed and expert the group, however, a limitation is that they are few in number, and therefore, bound to have only limited intellectual perspectives and reach into the STEM community. Ideally this narrow deep expertise would be balanced with a wider range of diverse knowledge. To achieve this balance a social networking site will be established to connect the study team, site personnel, and advisors together, and all of them as a group into the STEM and evaluation communities at large.

By paying a group of acknowledged experts we are assured of useful but narrow feedback. By extending social networking to the greater community, we are opening ourselves to broader advice from a national group of interested volunteers, and in so doing, also promoting the cause of knowledge transfer.

Several important principles of successful social networking will guide the development of the collaboration site (Noveck 2009). Those efforts will fall into two categories – site design and outreach.

Site design: The key to successful site design is to construct the visuals and the user interface in a manner that entices engagement. Four tactics are useful to achieve that engagement. 1) Provide intriguing material to welcome potential new users. That material is assured because the advisory board, test site personnel, and the research team all intellectually committed to the project, and paid for their participation. 2) Channel user input to avoid ranting and unfocused discussion. Focus will be achieved by asking for specific answers to specific questions. Not: "Are randomized designs good to use in STEM evaluation?", but: Are there particular conditions under which STEM interventions are well enough defined, the target population is homogeneous enough, and the setting is stable enough, for a randomized design to be advantageous?". 3) Ask users to self-identify as experts in particular topics, thus providing a data base of expertise and a focus for new users. Not: "expert in classroom observation", but "expert in classroom observation of middle school students in science labs". 4) Allow sub-communities of interest to coalesce.

Outreach: Building it does not mean they will come. Without encouragement and publicity we may assume that public participation will be minimal. Thus from the beginning determined efforts will be made to publicize the site via tactics such as listserv postings, notices at conferences, and contact with opinion leaders. In particular, nodes of communication will be addressed for NSF's STEM programs such as REESE, DRK-12, TUES, ISE, and ATE.

The Advisors

The project team has made a concerted effort to assure expertise in STEM application and evaluation that is both deep and broad. To that end, we have assembled an advisory board as follows:

STEM

David Beer	PI for Chicago case, cultural anthropologist, Experienced at working with teachers and educational evaluation.
Daryl E. Chubin	Founding Director, Center for Advancing Science & Engineering Capacity, at the American Association for the Advancement of Science (www.aaascapacity.org).
Beth Greene Costner	Chair of Mathematics and Assistant Dean for Teacher Education Programs in the College of Arts and Sciences at Winthrop University. Co-principle investigator for the Winthrop Initiative for STEM Educators funded through Robert Noyce Scholarship Program.
Suzanne Franco	Evaluation point of contact, Dayton case, Associate Professor, College of Education, Wright State University. Conducted many evaluations of math and science STEM innovations in grades K 12.
Abby Ilumoka-Nwabuzor	Professor of electrical and computer engineering at the University of Hartford. Established a STEM program to determine best practices for increasing the participation of women in STEM disciplines.
Webe Kadima	Associate Professor of Chemistry, SUNY Oswego. Principal Investigator on the NSF Advance Institutional Transformation (IT) Catalyst grant "Recruiting, Retaining, and Promoting Women in STEM Fields: Preparing for an Institutional

	Transformation Grant".
Margaret Pinnell	Assistant Dean in the School of Engineering University of Dayton in charge of outreach and recruitment. PI on NSF RET working with the STEM community. Member of Dayton Regional STEM Center advisory board.
Margy Stevens	Assistant Superintendent of Curriculum, Professional Development, and School Improvement for the Montgomery County Educational Service Center. Executive Director of Ohio's STEM center – Dayton Regional.

Evaluation

Stewart Donaldson	Professor and Chair of Psychology, Director of the Institute of Organizational and Program Evaluation Research, Claremont Graduate University. On the board of the American Evaluation Association (Donaldson 2007).
Frances Lawrenz	Extensive experience in evaluating STEM programs, and has deep roots in STEM education (Huffman and Lawrenz 2006).
Michael Patton	Well known for his work in utilization focused evaluation, qualitative research methods, and developmental evaluation (Patton 2008; Patton 2010).

"Deliverables" – Publications, Presentations and Web-based Information

Our belief is that to have maximum impact, no single document should stand alone outside of an active community of interest. Rather, social networking (with the emphasis on the *social* to leverage the networking technology) is required. All the deliberations on the project's social networking site constitute public information that has been disseminated to the evaluation and STEM communities. Also, drafts of presentations and articles developed by the research team will be posted to the site for comment and discussion prior to their official submissions to conferences or journals.

While specific audiences for presentations and articles cannot be identified in advance, it is possible to identify the population of organizations whose members would be interested in the results of the proposed work. These include (in alphabetical order): American Educational Research Association, American Evaluation Association, Association of Mathematics Teacher Educators, Mathematical Association of America, National Association for Research in Science Teaching, National Council of Teachers of Mathematics, National Science Teachers Association, and NSF meetings of STEM principal investigators. This is a large list from which to choose. As of this writing we are inclined to focus on the evaluation related associations because this project is designed to improve evaluation methodologies. However, it is also designed to support STEM evaluation in particular, which argues for a different set of priorities for presentation. From a practical point of view, the final determination will depend on the project's findings and conversation with the NSF.

Methodology

The research question to be investigated is: How can STEM evaluation be improved by using a particular set of tactics that are embedded in a theory of evaluation, all designed to work together to produce powerful evaluations under conditions of unexpected program change? The methodology to be employed to answer this question can be summarized as a multiple case approach with inter and intra-case comparisons over a period of three years, and interventions entering the cases at different points in the program/evaluation life cycle.

Consulting Protocol for the Intervention

Consulting between the project team and case study representatives will be based on the approach to evaluation outlined in (Morell 2010), and set out in training form in workshops presented at the Centers for Disease Control and the European Evaluation Society. (Explanation of the approach appears in the first section of this proposal. Slides from the workshop are downloadable from www.jamorell.com.) In-person training will be followed up by telephone and Web-based support.

Multiple Case Study Design

The proposed research encompasses six different STEM implementations, comprising two of the multiple case study designs identified by (Yin 2009). Comparison of the University of Chicago and Dayton sites will be treated as "separate cases in separate contexts". Because the Dayton site has five STEM activities however, that site can be treated as a scenario of multiple cases embedded in the same context. Because of these logical differences, a wide variety of system and contextual factors can be factored into the analysis. (Using five settings in Dayton is practical because the same evaluation team is involved in each of the five.) In these cases the research team will provide evaluation assistance to the settings (in the form of consultation and graduate assistant support). As the support unfolds, another level of interaction will observe the activity and collect interview data from all the parties. Which parts of the methodology for dealing with surprise were adopted? Which were not? Which were adopted partially? Why were these decisions made? Were the techniques useful for detecting looming change and/or for building agile methodologies? Did the information received support the needs of the stakeholders? What techniques emerged that were not anticipated by the research team? As the data are collected, it will be analyzed within the context of the frameworks presented earlier: Tactics along the foreseeable ←→ unforeseeable continuum (Figure 1), project x evaluation life cycle relationships (Figure 3), social/organizational sources of surprise (Figure 4), and trade-offs among tactics (Figure 2).

Ideally, at least one case would serve as a "no treatment" condition in the sense that STEM activities were evaluated, but without the assistance of the research team. Unfortunately this tactic is not practical for budgetary reasons. However, the nature of the cases will make it possible to approximate a "no treatment" condition in two different ways. First, the STEM programs in Dayton and Chicago differ in how advanced they are with respect to both program implementation and evaluation implementation. Thus the research team will be entering at different points in the "evaluation x program" life cycle depicted in Figure 3.

The second way that "no treatment" can be approximated derives from the fact that some members of the advisory board are involved in STEM activities that are not part of the interventions proposed in this project. Thus these advisors can report on how evaluation activities are unfolding in their settings. Of course data of this kind is limited and murky. It is limited because advisory board members are

constrained to three days per year each. It is murky because advisory board members will have close knowledge of what is being learned by the research team. Still, obtaining "quasi no treatment" information from the advisory board may prove insightful in combination with the in depth information that will be collected from the test sites.

Inter and Intra-case Cross Sectional and Longitudinal Comparisons

Because six cases will be followed over three years, the resulting data set will provide a rich source of comparisons within sites and across sites over a protracted period of time. The longitudinal aspect of the research is particularly important because the nature of unexpected program behavior is that surprise can evolve over time as developmental trajectories work themselves out. Unexpected behavior may affect early-stage process variables in a program, all the way through long term outcomes. Sometimes there is dependence, i.e. unexpected developments at one point affect what happens in the future. Sometimes the unexpected changes are independent of each other, but still worthy topics for evaluation. It is because these developmental trajectories are so important, that the research needs the three year funding time frame that will support a longitudinal design. For maximum value, the analysis must be both cross sectional across cases, and longitudinal over time.

Data Collection

Two Sources, Two Distances from Implementation

Data collection will take place in two ways. First, part of the research team will continue the consultation described above throughout the project, and while consulting, will record their observations. The advantage of this process is that a continual stream of detailed information will be collected. The disadvantage is that those collecting and analyzing the data can become too close to the process and lose perspective. Thus a second layer of data collection will be carried out as well. Here, members of the research team not directly involved with ongoing consultation will conduct periodic interviews with all involved – both the STEM case actors and the researchers/consultants. These second level interviews will be done on a rotating basis throughout the project, although contact with each actor will be staggered at two month intervals. This plan will assure that evolving situations are carefully monitored, but that no single actor is the subject of too many (or too frequent) interviews. While the primary source of data will these interviews, a great many artifacts will undoubtedly prove useful, e.g. data collection instruments developed by the study team, implementation and evaluation schedules, and project planning documents. These too will be collected and used.

Data Organization and Templates

Data collection templates will be based on Figures 1, 2, 3 and 4 – the placement of tactics along the foreseeable ←→ unforeseeable continuum, the evaluation x program life cycle map, the social/organizational map of evaluation surprise, and the guide to trade-offs among evaluation tactics. These templates will be used because they constitute the theory of evaluation surprise that is being tested here as a way to improve STEM evaluation. We expect, of course that data will dictate modifications in these perspectives on evaluation surprise and that other data interpretation structures will arise.

Analysis and Interpretation

Data interpretation will be layered with respect to the groups who are in a position to analyze the data. The core of the interpretation process will be the research team involved in the case consulting, project

oversight, interviewing, and data recording. These are the people who are intellectually closest to the work and who have the primary responsibility for developing new STEM evaluation methods. The second layer will be participants in the case studies at the University of Chicago and the Dayton Region. This group has the most direct interest in good evaluation of the programs for which they are responsible. They have the motivation to assist and the closeness to the process needed to inform on what works, what does not work, and why. The third layer involves the members of the STEM and evaluation advisory boards. Given that each board member is only available for three days per year, reliance on these groups will have to be done judiciously. Still, we believe that because they will not have to travel, and because the project's social networking site will keep them continually aware of unfolding events, that their collective input will provide useful guidance. The final layer of data interpretation will come from the STEM public at large, through participation in the project's social networking site. We will conduct the project in such a way that outside input is unnecessary because it is by no means clear that no matter how carefully done, there will be much input from outside the project's immediate boundaries. Still, important expertise does reside out there, public involvement will serve the causes of both analysis and dissemination, good site design can facilitate participation, and aggressive outreach can make the site known.

Social Networking to Support Analysis and Dissemination

Because social networking is so important to the collaboration functions of the proposed work, considerable effort will be put into setting up a system that will work for all participants. Details of how the social networking site will be built and run are contained earlier, in the section: "STEM and Evaluation Advisory Groups: Continuous Feedback Through Social Networking". We will employ a rapid prototyping methodology that will include representatives of all stakeholder groups to determine the final design of the site.

Logistics and Timing

The first six months of the project will consist of intense interaction between the research team and representatives of the two in-depth cases. This time will be needed for the case representatives to get acquainted with the proposed approach, for our study team to understand the specifics of the cases and their evaluation requirements, for whatever logic models that may be needed to get developed, for reviews of evaluation findings to date, and for the necessary levels of trust and social relationships to be solidified. We expect data collection and analysis to continue up through the last four months of the project, at which point most project resources will have to go toward developing final reports. It is important to extend data collection as long as possible because of our strong belief that program impact can morph and expand over time in unpredictable ways.

References

Behrens, T. R. and D. O. Gray (2001). "Unintended consequences of cooperative research: impact of industry sponsorship on climate for academic freedom & other graduate student outcome." Research Policy **30**(2): 179-199

Courtney, M. E., B. Needell, et al. (2004). "Unintended consequences of the push for accountability: the case of national child welfare performance standards " Children & Youth Services Review **26**(12): 1141-1154

DiNardo, J. and T. Lemieux (2001). "Alcohol, marijuana, & American youth: the unintended consequences of government regulation " Journal of Health Economics **20**(6): 991-1010

Donaldson, S. I. (2007). Program Theory-Driven Evaluation Science: Strategies and Applications

Hillsdale NJ, Lawrence Erlbaum.

Dorner, D. (1996). The Logic of Failure. NY, Henry Holt & Co.

Fry, M. and J. Polonsky (2004). "Examining the unintended consequences of marketing " Journal of Business Research **57**(11): 1303-1306

Hoek, J. (2004). "Tobacco promotion restrictions: ironies & unintended consequences " Journal of Business Research **57**(11): 1250-1257

Huffman, D. and F. Lawrenz, Eds. (2006). Evaluation of STEM Educational Initiatives. New Directions for Evaluation. New York, Wiley.

Kaminski, M. (2001). "Unintended Consequences: Organizational Practices & Their Impact on Workplace Safety & Productivity." Journal of Occupational Health Psychology **6**(2): 127-138

Kildegaard, A. (2001). "Fiscal reform, bank solvency, & the law of unintended consequences: a CGE analysis of Mexico " The North American Journal of Economics & Finance

12(1): 55-77

Lukas, B. A. and A. Menon (2004). "New product quality: intended & unintended consequences of new product development speed." Journal of Business Research **57**(11): 1258-1264

Morell, J. A. (2005). "Why are there unintended consequences of program action, and What Are the Implications for Doing Evaluation?" American Journal of Evaluation **26**(4): 444 - 463.

Morell, J. A. (2010). Evaluation in the Face of Uncertainty

Anticipating Surprise and Responding to the Inevitable. New York, Guilford.

Noveck, B. S. (2009). Wiki Government: How Technology Can Make Government Better, Democracy Stronger, and Citizens More Powerful. Washington DC, Brookings Institution Press.

Patton, M. Q. (2008). Utilization-Focused Evaluation 4th ed. Thousand Oaks, CA, Sage.

Patton, M. Q. (2010). Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use. New York, Guilford

Pugh, K. J. and Y. Zhao (2003). "Stories of teacher alienation: a look at the unintended consequences of efforts to empower teachers." Teaching & Teacher Education **19**(2): 187-201

Stiles, K. (2002). "International Support for NGOs in Bangladesh: Some Unintended Consequences." World Development **30**(5): 835-846

Tenner, E. (1996). Why Things Bite Back. NY, Knopf.

Yin, R. K. (2009). Case Study Research Design and Methods, 4th ed. Thousand Oaks CA, Sage.