STAGED EVALUATION: AN APPROACH TO MORE COST-EFFICIENT AND USEFUL EVALUATION OF LARGE AND COMPLEX INITIATIVES

Laura Stokes

With assistance from
Mark St. John
Gerald Accurso
Elizabeth Horsch
Dawn Robles
Pamela Tambe

Final report for a Small Grant for Experimental Research (SGER), funded by the National Science Foundation. Award # 0909720

APRIL 2011
INVERNESS RESEARCH

TABLE OF CONTENTS

ABSTRACT

Inverness Research investigated the concept and practice of *staged evaluation*, using a study of the National Science Foundation-funded Undergraduate Research Collaborative (URC) in Chemistry as the core case. The investigation examined the proposition that before investing in a full-scale evaluation of a large and complex initiative, it would be wise to conduct a *Stage One* study first. A Stage One study is a brief and exploratory effort of "ground-truthing"; its purpose is to clarify the need, purpose, and design of a fuller and more rigorous study. Results of the Stage One study of the URC included observations of the ways in which the five funded URC sites had interpreted the multiple broad and ambitious goals of the URC initiative in their program designs, and included reflections on the import of similarities and differences in those designs. Stage One results also included the framing of multiple options for Stage Two (fuller, more rigorous) evaluation, each with specific purposes and audiences in mind. A panel consisting of current and former NSF program officers and evaluators in Science and Education directorates, as well as an independent evaluation expert, reflected critically on the process and results of the URC study for the purpose of assessing the value of Stage One studies as a cost-efficient and effective way to evaluate large-scale and complex initiatives. These reflections produced the following key findings: 1) A Stage One study can provide funders with early "reality checks" on the progress of an initiative as it is enacted in the field. This alone is a significant advantage of staged evaluation, leading to the conclusion that Stage One could well be termed "consultative evaluation." 2) A Stage One study can assist greatly in framing fuller evaluations, including the possibility that further evaluation is not needed. This is a cost-efficient alternative to designing full evaluations *a priori* and then discovering their foci or data definitions are not well aligned with program actualities. Stage One studies can also provide reluctant funders with an evidence-based rationale for investing in further evaluation. 3) Staged evaluation is not without disadvantages. Policy-makers or funders could over-rely on the quick results of a Stage One study rather than waiting for the results of the more rigorous but also more time-consuming Stage Two study. Also, a staged approach is less appropriate and advantageous for projects that require randomized controlled trials as the only legitimate evaluation design. The major conclusion of this study is that staged evaluation appears to be very useful and cost-efficient for the evaluation of large, complex initiatives, especially those where the design of programs is left at least partially to the discretion of grantees within funder parameters and goals. Additional examples of Stage One studies and their results would help build shared understanding of their value as well as help specify a model of Stage One evaluation for the field.

I would like to see a broader repertoire of tools for evaluation. NSF has the reverse site visit, and the committee of visitors, and some project evaluation, and then there are huge program evaluations. I am feeling like as a complementary tool, there are times that this "stage one" approach might be a very useful thing to have… some kind of small, intense, short-term look at a program. It could be very useful for different purposes.

—Mark St. John,
Inverness Research

I. INTRODUCTION

Inverness Research was awarded a NSF Small Grant for Experimental Research (SGER) to explore and test the concept of *staged evaluation.* To test this idea, Inverness conducted a pilot Stage One evaluation, using the NSF-funded Undergraduate Research Collaborative (URC) in the Division of Chemistry as the core case.

Staged evaluation is proposed as an approach to evaluating large-scale, multi-faceted, multi-year initiatives such as those often sponsored by NSF. The core proposition is that before investing in a full-scale evaluation of such an initiative, it would be wise to conduct a *Stage One* evaluation of the effort.

This report presents a comprehensive account of the study. Section I reviews the general concept of staged evaluation and provides the background and rationale for the URC study. Section II describes the design, activities, and results of the Stage One study of the URC initiative. Section III offers reflections on the staged evaluation approach and examines implications of the study, including prospects for promulgation of staged evaluation as a cost efficient and effective approach for large, multi-faceted initiatives. This section draws from a live discussion of the URC results in particular and the staged approach in general, at a February 4, 2010, conference in Washington D.C., as well as from written comments on a draft of this report that was prepared following the conference. Discussants and draft report respondents included NSF staff from the Science and Education Divisions, an independent evaluation expert serving as a consultant, and members of the research team conducting the Stage One evaluation.

II. BACKGROUND: A STAGED APPROACH TO EVALUATION

The core proposition underlying a staged approach is that before investing in a full-scale evaluation of a large initiative, it would be wise to conduct a *Stage One* evaluation first. A Stage One study is exploratory in nature, aiming to develop a grounded portrait of the initiative's activities and a preliminary understanding of its likely outcomes and benefits. The results of this first stage are then be used to clarify the need, purpose, and design of a fuller and more rigorous study—the Stage Two evaluation. The presumed benefits of staged evaluation are that this approach will lead to more cost efficient use of evaluation resources and useful evaluation results.

The rationale for a staged approach to evaluation

Just as improvement initiatives vary in their size and complexity, so do evaluations. With evaluation dollars often running at 8-12% of the cost of an initiative, there is little room for cost inefficiency, even as the stakes are high for evaluation results.

Complex initiatives imply complex evaluation designs and choices

The National Science Foundation—as well as other federal and non-federal funders of educational improvement efforts—increasingly supports complex educational initiatives that involve multiple partners and collaborations, and that often lead to the creation of new infrastructures such as centers and networks. Very often, these initiatives involve work in multiple domains of the formal and informal education system, have multi-layered "logic models" underlying their designs, and have potential to generate multiple kinds of proximal and distal outcomes. The functioning of these initiatives is often not straightforward or stable, but rather emerges and takes shape over time in a range of contexts. At the same time, evaluation studies can have multiple purposes and focuses. Hence it is often not possible to know ahead of time what will be salient. In short, the more complex the initiative, the more complex a task it is to design an evaluation that is both cost-efficient and useful.

The practice of investing in planning

Staging the funding of complex projects has become commonly accepted practice in grant-making. To optimize the impact of their investments, the NSF and other foundations use planning grants or other exploratory work that involves preliminary development of ideas or testing of feasibility. These low-level grants for planning allow for the careful, staged development of larger, much more expensive initiatives. Often, the planning grants lead to commitment to larger investment. Even if the initial outlays for planning do not result in such a commitment, it can be argued that planning grants and pilot projects are nonetheless a cost-effective investment by helping the NSF to see that extended funding is not warranted. In reality, many large highly-funded initiatives have to spend the first year or more in planning and early development, and they do so at a much higher cost than would have been incurred with a planning grant. Thus, there is a strong argument to be made for the funding of early exploratory work.

Investing in pilot, or staged, evaluation studies of complex initiatives

This Staged Evaluation SGER project is based on the idea that an analogous approach to designing evaluations of complex initiatives may be as useful and as cost-effective as staging the funding of the project itself. Just as a pilot project can produce results that lead to a more soundly designed initiative, a Stage One study can lead to a full evaluation that is focused on outcomes that the initial study suggests are high-potential. And just as it is sometimes unwise to fully fund a large-scale project, it may be equally unwise to fully fund a large-scale evaluation without evidence from a pilot evaluation approach.

There is little history of funding pilot evaluation studies in the NSF. Inverness Research's experience studying many complex initiatives is that large-scale evaluations, *de facto*, frequently devote the first year to exploratory study that results in a new or re-shaped evaluation design. These experiences have led IR to believe that it would be wise both to honor and to formalize the need for this initial planning stage for evaluation.

The concept of staged evaluation

To understand what IR means by a "Stage One" evaluation, one might think of a spectrum of evaluation approaches, varying across the dimensions of cost and depth. At one end of the spectrum would be an in-depth, thorough, and comprehensive evaluation with pre-defined outcome measures and methods. At the other end would be a brief program review such as those conducted by an NSF Committee of Visitors. Stage-one evaluations, as IR conceptualizes them for this project, fall somewhere in the middle of that spectrum—more comprehensive than a review and more exploratory than a full-scale evaluation.

The idea IR is pursuing is close to the idea of "evaluability" studies (Trevisan and Huang, 2003; Trevisan, 2007). Michael Trevisan and Min Huang (2003) describe evaluability assessment studies in the following way:

> *A strategy that can be used to determine the extent to which a program is ready for full evaluation, is known as evaluability assessment. Initially developed by Wholey (1979), evaluability assessment (EA) seeks to gain information from important documents and input from stakeholders concerning the content and objectives of the program. Outcomes from EA include clear objectives, performance indicators, and options for program improvement. Wholey (1979) recommended EA as an initial step to evaluating programs, increasing the likelihood that evaluations will provide timely, relevant, and responsive evaluation findings for decision makers.*

Trevisan studied 23 different EA studies conducted from 1986 to 2006 (Trevisan, 2007), finding that:

> *Most studies employed document reviews, site visits, and interviews, common methodologies previously recommended in the literature on EA… The most common rationale for conducting EA mentioned in these studies was determining program readiness for impact assessment, program development, and formative evaluation. Outcomes found in these studies include the construction of a program logic model, development of goals and objectives, and modification of program components. The findings suggest that EA is practiced and published more widely than previously known.*

The EA approach of early small-scale evaluations thus not only tests the readiness of a program for a full evaluation, but also has potential to provide useful formative feedback to project leaders.

Stage One evaluation, as IR set out to conceptualize and test it, has many of the features of an evaluability assessment. However, IR suggests that Stage One evaluation goes beyond assessment of "evaluability." More than providing formative feedback to a project, a Stage One evaluation can provide preliminary findings to the funder, and, in addition, it has the purpose of establishing guidelines and a framework for the design of a second, more in-depth stage of evaluation.

III. TESTING STAGE ONE EVALUATION: THE CASE OF THE URC INITIATIVE

A. THE CONTEXT

IR tested the concept of staged evaluation in two steps: first, by designing and conducting a Stage One evaluation of a large and complex NSF science initiative, the Division of Chemistry's Undergraduate Research Collaboratives (URC); and second, by examining that evaluation as a case of a staged evaluation.

The case: Undergraduate Research Collaboratives in Chemistry

The Undergraduate Research Collaboratives (URC) program is funded by the Division of Chemistry within the NSF to promote the involvement of undergraduates in research. The program sought to create "new models and partnerships with the potential (1) to expand the reach of undergraduate research to include first- and second-year college students; (2) to broaden participation and increase diversity in the student talent pool from which the nation's future technical workforce will be drawn; and (3) to enhance the research capacity, infrastructure, and culture of participating institutions." The URC initiative funded projects at five sites: University of Chicago; Purdue

University; University of South Dakota; Ohio State University, Columbus; and the University of Texas, Austin. Grants were staggered over a period of three years; thus, sites were in different stages of development during the study.

The URC is a complex initiative with goals that are multi-dimensional, multi-level, and ambitious, as evidenced by the criteria used to evaluate the proposals that were submitted to create the Collaboratives:

> *1. The extent to which the URC creates and tests a new model for building a research community and performing undergraduate research.*
>
> *2. The extent to which the URC model is scalable, sustainable, able to be replicated or adapted, and integrated into the curriculum.*
>
> *3. The quality of the research experience that URC-supported students will have, including the extent to which students will create new knowledge that is potentially publishable.*
>
> *4. The extent to which the URC will increase the number and diversity of students participating in undergraduate research, including students who might not otherwise be exposed to chemical research.*
>
> *5. The extent to which the URC builds research capacity, infrastructure and culture that is sustainable beyond the URC award at partnering institutions.*
>
> *6. The extent to which the URC partnership and management promotes inclusive and effective mentoring and enhances the professional development of mentors.*

The URC program provided an authentic example of the kind of initiative and funding context where a staged evaluation approach could be most productive. First, the URC program—which seeks to generate a symbiotic connection between the research and education functions that are part of the mission of the NSF—is a good example of a complex initiative housed within a science division at NSF. Given the size, complexity, and ambitious goals of the URC, it is not immediately clear how to most effectively or cost-efficiently design a comprehensive and in-depth evaluation. Additionally, the science divisions of NSF have little experience in designing educational evaluations, and hence design assistance is essential to framing and implementing a successful evaluation of the URC program. Finally the URC program provided a case where the evaluation was wanted by NSF—where the Stage One evaluation would be put to practical use, and where program staff were willing to collaborate in learning about the strengths and weaknesses of a staged approach.

The purposes of the Stage One study

IR framed several specific purposes for the Stage One study, having in mind that the results would directly serve the NSF Chemistry Division as well as this inquiry into staged evaluation:

- To document the purpose and theory of action of the URC
- To ground-truth the URC theory of action[1]
- To describe the landscape and context of undergraduate chemistry education served by the URC grants
- To arrive at preliminary assessments of contributions of the initiative at multiple levels
- To identify critical issues in the URC program and Chemistry education
- To determine the desirability and feasibility of a Stage Two evaluation
- To provide recommendations and a framework for the design of a Stage Two evaluation

The research team

The research team carrying out the Stage One study was composed of Inverness Research staff as well as independent experts who served as consultants providing a range of perspectives. The IR staff included a former IHE administrator, a former chemistry educator, as well as the IR president, noted for his background in both science education and evaluation. Additionally, Ron Christensen of Bowdoin College, a chemist and former NSF program officer, accompanied the team on site visits and conducted an independent review of student work products. Sarah Beth Woodruff, Director of Ohio's Evaluation and Assessment Center for Mathematics and Science Education and the evaluator for two of the funded URC collaboratives, and Mary Berry, of the Department of Chemistry at University of South Dakota (who served as evaluator for one URC site), also accompanied the IR team on site visits and assisted in document reviews. Expanding the team to include these three consultants—with their backgrounds as scientist, science educator, funder, and evaluator—was a deliberate effort to bring multiple perspectives and relevant expertise to bear on the Stage One study.

The second part of the study—the examination of the Stage One study as a case—was carried out by a different Inverness researcher, the author of this report, working independently.

Evaluation activities

Data gathering about the URC was carried out over a 9-month period, and involved the following activities, all designed to generate data adequate to the Stage One purposes of developing a grounded portrait of the initiative's activities and a preliminary understanding of its likely outcomes and benefits:

[1] By ground-truthing, IR means gathering data "on the ground"—that is, from the work and participants of the funded Collaboratives—to assess the extent to which project actualities are congruent with the theory of action.

The study of the URC by the IR Stage One team

- **Interviews with NSF program officers**. These interviews served to document the history and evolution of the URC initiative, to document the intentions and theory of action of the program, and to identify funder perspectives on key outcomes and the assumptions underlying them.
- **Review of documents**, including the URC solicitation, URC site proposals and internal evaluations. These helped to document the intentions and assumptions of the specific Collaboratives that were funded, to begin assessing the congruence of project designs with initiative intentions and outcomes, and ultimately, to provide input into initiative theory of action.
- **Site visits to URC projects**, involving interviews with PIs, faculty, evaluators, and students, as well as observing students presenting their projects. These visits enabled further ground-truthing of the theory of action. Additionally, site visits enabled the team to independently observe and assess the nature of the student experience and the quality of their work, as well as to begin to assess the most probable strengths and contributions of the initiative in light of its intended outcomes.
- **Independent expert review of student work products**. Ron Christensen conducted a focused review of the nature and quality of student and faculty research projects and products. This served the purpose of directly addressing a core value, expressed by members of the NSF Chemistry Division, that the URC program must generate benefits to scientific knowledge as well as to Chemistry education.
- **An in-person half-day presentation of the results of the Stage One study.** This meeting involved NSF staff from the Division of Chemistry and other NSF Divisions in Science and Education, as well as an independent expert on education research and evaluation, Daniel Humphrey of SRI. The meeting served a dual purpose. First, it aimed to serve the needs of the Chemistry Division to learn about the URC initiative and make decisions about future directions for the initiative, including Stage Two evaluation. Second, the presentation served as an example of what a Stage One evaluation can produce, and thus formed much of the substance for the second step, the examination of the study as a case.

Examination of the Stage One study as a case by the independent researcher

- **A half-day discussion among NSF staff and the independent evaluator of the Stage One study as a case**. This discussion served the purpose of reflecting on the key features, value, advantages, and disadvantages of a staged approach to evaluation, drawing from the URC study as an example.
- **Development of this report in draft form, followed by review and comment** by members of the research team, NSF staff, and an independent evaluation expert. This serves the purpose of further reflection on and refinement of the concept and practice of staged evaluation.

- **Preparation of this report for NSF and the field,** as documentation of the approach to evaluation of large and complex initiatives.

B. Results of the Stage One study of the URC

The Stage One study yielded two kinds of results. First, it yielded observations that generated a *preliminary assessment of the URC initiative*: that is, the key features of site programs and their rationales, the range and quality of the sites' efforts to strengthen undergraduate chemistry, and the types of outcomes the initiative appeared to be on the way to producing. Second, and following from the assessment of the URC, the Stage One study produced *recommendations about future evaluation* (i.e., the possible purposes for further evaluation), and recommendations to NSF about how to invest in evaluation under different funding conditions.

NSF officers and the independent evaluation consultant used these results to reflect on the value of a staged approach to evaluation.

1) Preliminary assessment of the URC initiative

The IR Stage One team's observations and reflections are organized according to the six criteria that the NSF established for the URC program review, plus a seventh added by the research team.

Criterion 1	*The extent to which the URC creates and tests a new model for building a research community and performing undergraduate research.*
IR Team Observations	The URC projects varied substantially, a result of their operating in different institutional contexts and their being shaped by PIs with different motivations for developing projects. Across these varied projects, however, there were some common features. All sites succeeded in fostering undergraduate research experiences, with students having opportunities to present their findings. These experiences were added via development of curriculum-based research modules that were then integrated into existing introductory courses. Sites varied greatly in the degree to which they built what could be deemed "research communities" among faculty and/or students.
Criterion 2	*The extent to which the URC model is scalable, sustainable, able to be replicated or adapted, and integrated into the curriculum.*
IR Team Observations	Sites varied both in their approach to each goal and the degree to which they achieved them. Modules were the primary vehicle for introducing research into the traditional chemistry

	laboratory curriculum and this process appears to be replicable. However, there are some limits to exportability. Integration between the research experience and the lecture content improved when sites designed and implemented content-specific modules at the site. When sites simply inserted "off the shelf" modules, integration into the existing chemistry curriculum was less apparent. Second, when modules were both designed and implemented by an author or a team of authors, there appeared to be more emphasis on revision in response to student input. Finally, institutions varied significantly in the amount of access students have to the modern implementation required by research. URC projects have been limited in their achievement of scale and sustainability. Thus far, there is little apparent faculty and administrative support beyond that provided to the grant participants. Also, facilitating student research is inherently more costly—requiring more space, more faculty time, more complex scheduling and logistics, more funds for mentoring of faculty and students—which has so far set limits on expansion.
Criterion 3	*The quality of the student research experience that URC-supported students will have, including the extent to which students will create new knowledge that is potentially publishable.*
IR Team Observations	The scale and nature of student research experiences varied greatly across sites, ranging from 10-20 hours in a course, up to 3 semesters and a summer experience. The IR team saw multiple examples and compelling evidence of authentic and quality research experiences at an appropriate level of rigor for undergraduates. There were some examples of publishable and published research, though it was difficult to discern level of undergraduate involvement in co-authorship. Findings derived from on-site observations of student presentations were confirmed by the independent review of the quality of student research; this review was conducted by the chemistry expert consulting with the research team.
Criterion 4	*The extent to which the URC will increase the number and diversity of students participating in undergraduate research, including students who might not otherwise be exposed to chemical research.*
IR Team Observations	All sites succeeded in increasing the number and diversity of students with access to science research experiences. Total annual student engagement was approximately 2500 students in 2008-09, Year 4 of the initiative. Numbers per site ranged from 56 to 973. Sites varied greatly in their recruitment

practices and degrees of effort; proactive strategies increased student access more than less proactive efforts.

Through student self-reports, the IR team found multiple instances of student success in chemistry and improved self-confidence in science that would not have happened without the URC. Students who had the most positive experiences and had changed their attitudes about their potential for succeeding in science tended to be at sites where the research experiences were more extensive and more integrated into the curriculum.

Criterion 5

The extent to which the URC builds research capacity, infrastructure and culture that is sustainable beyond the URC award at partnering institutions.

IR team Observations

The impact of the URC on institutional capacity, infrastructure, and culture varied widely across sites, depending on the size and existing capacities of the institutions. In smaller universities, the URC project helped build foundational research capacities, including greater access to instrumentation and to other faculty expertise, increased administrative support for student research, and greater ability to attract other outside grant funding. In the larger universities, the project provided much-needed empowerment and support for those faculty who were champions of curriculum improvement; in some cases the URC project also led to re-allocation of institutional funds.

Criterion 6

The extent to which the URC partnership and management promotes inclusive and effective mentoring and enhances the professional development of mentors.

IR Team Observations

Sites varied in how they defined and approached mentoring. Professional development for faculty focused on their mentoring role but also went beyond mentoring. The student peer mentoring models that arose appeared to be successful. At some sites post-docs were also an important mentor group.

Criterion 7 (added by IR)

The quality of URC sites' evaluation and dissemination plan.

IR Team Observations

The scope of evaluation varied according to the size of the institution and project. Evaluation played an important formative role at all sites. In some larger sites with greater evaluation capacity, studies focused in depth on the impact of the initiative at that site. Evaluation information and results were distributed within sites but dissemination beyond sites—including to other participating institutions—was minimal.

Summary of observations

Changes in the student research experience took different forms at the five sites. Nonetheless, there is ample evidence to suggest that students had engaging research experiences and produced worthy research results and products. Importantly, the five sites increased the number and diversity of students with access to research experience. Additionally, the project provided a base of support to champions of improvement in chemistry education, and enabled them to form new professional relationships through collaboration. At the smaller sites, in particular, faculty members developed a stronger research culture and community through the project.

On other hand, the research team did not find evidence that the model is sustainable at the institutions nor replicable at new sites. The team did not hear about cross-site networking or dissemination beyond the sites, even from the more mature sites.

IR team reflections on the goals and criteria of the initiative

The Stage One research process aimed to show not only how the accomplishments and work of sites can be examined through the lens of initiative goals and criteria, but also the reverse: how study of multiple sites can be used to reflect back on, and enrich understanding of, the goals and criteria of the initiative. The IR team posited that this can be very helpful to initiative planning and development.

The results of the Stage One study showed that adding authentic research to undergraduates' introductory experience of chemistry is a viable idea. The study also showed that the initiative generated a great deal of individual variation across sites. Variation was probably inevitable because sites had different motivations for participating, faced different contexts and constraints, and had different capacities. Additionally, the initiative held sites to multiple ambitious criteria, and as a result of their contexts and internal priorities, sites assigned different weights to the criteria. For example, in comparing findings related to criteria 2, 3, and 4 above, the team observed that, generally speaking, the short "replacement" modules were logistically easier to use and perhaps more exportable; however, those modules also appeared to be less likely to generate powerful alternative learning experiences for students. Conversely, re-designed courses that included well-integrated and in-depth research experiences appeared more likely to alter students' experiences of chemistry and produce high-quality scientific work, and also to provide faculty with a greater level of professional growth opportunity. However, changing courses to this greater degree was more taxing on institutional resources and infrastructure and therefore more difficult to sustain or replicate. Thus, the sites that put more effort into the quality and depth of student experience were more successful against some criteria, while sites that developed more efficient, more scalable curriculum change processes were more successful against different criteria.

This outcome suggests that the multiple goals of the URC initiative stand in some degree of competition with one another. The question then might be which criteria and goals are most important from the point of view of the initiative as an investment (for example, quality of student products *vs.* scale, or local faculty capacity-building *vs.* replicability). Ultimately, these observations led the evaluation team to raise the question about whether the range of site-level responses to the URC opportunity reflected five related manifestations of the same theory of action or five truly different "models" of undergraduate chemistry. One outcome of the Stage One study, then, was to give the NSF program officers a means to reflect on the range and causes of variation and similarity, which in turn could inform them about their theory of action and the goals of their initiative.

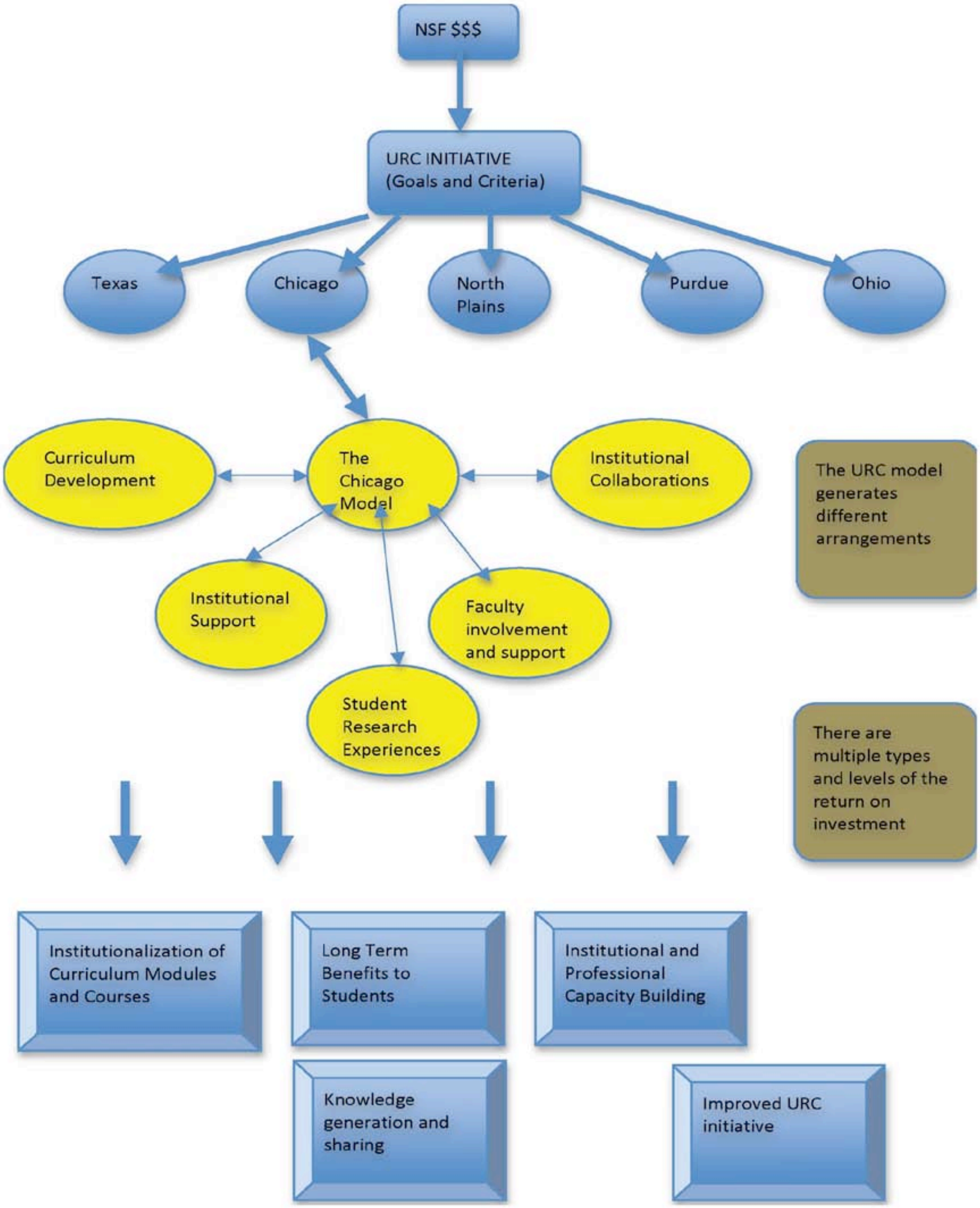
2) Recommendations for Stage Two evaluation

One reason the URC in the Division of Chemistry was selected as the case for this investigation of a staged evaluation approach is that the Division program officers were truly in need of assistance in framing an evaluation for the initiative. Although the URC program required each site to have an evaluator for the local project, the Division felt unprepared to craft a solicitation for the initiative-wide evaluation. This section summarizes the options and recommendations the IR team provided to URC staff as a result of the Stage One study.

A key assumption underlying Stage One evaluation is that resources for evaluation are quite limited, especially compared to the scale of resources invested in the projects themselves. And when there are multiple possible outcomes, then deliberate decision-making is required to use evaluation resources optimally. One consideration to take into account in framing future evaluation is the assumption that initiatives that have multiple and ambitious criteria for success are not limited to simple results or single outcomes; rather, such initiatives are intended to generate, and do generate, multiple types and levels of return on the investment. A critically important role of the Stage One evaluation is that it helps to illuminate the range of those potential returns and also helps assess the likelihood that various returns will be weaker or stronger, given the evidence to date.

The URC sites varied in their emphasis on curriculum development, on the nature and extent of student research experience, on the ways in which institutions built collaborations, and on the nature and extent of faculty development and support that were created. As suggested in the diagram on the following page, the specifics of each site-level model would differ from the others across these several dimensions. And through these theme-and-variation arrangements, the URC investment could be seen to produce a range of potential returns on the investment. These include direct benefits to participating students, institutionalization of curriculum modules, professional and institutional capacity-building, the production and distribution of new knowledge, and an improved or newly designed URC initiative. Evaluation dollars could be invested to document and assess any one or several of these contributions of the initiative.

Figure 1. Variation in site design and multiple outcomes to assess


NSF $$$
URC INITIATIVE (Goals and Criteria)
Texas
Chicago
North Plains
Purdue
Ohio
Curriculum Development
The Chicago Model
Institutional Collaborations
The URC model generates different arrangements
Institutional Support
Faculty involvement and support
Student Research Experiences
There are multiple types and levels of the return on investment
Institutionalization of Curriculum Modules and Courses
Long Term Benefits to Students
Institutional and Professional Capacity Building
Knowledge generation and sharing
Improved URC initiative

In sum, framing a second stage of evaluation involves making choices about which potential returns merit an investment in evaluation resources that can document those returns.

Options for a Stage Two evaluation

The IR team posited a series of options and recommendations for additional evaluation. The options were generated through a process of considering a range of purposes for evaluation and potential audiences or users of evaluation. They also reflected a range of foci, based on consideration of various outcomes that might merit an investment in further evaluative study.

Any of the first three options outlined below could be useful for NSF and also for others in the field of chemistry education, whether NSF continued or did not continue the URC initiative. They range from quite narrowly focused to very broad in their scope and purpose.

Option 1. Assess contributions to students, with special emphasis on increasing access to underserved students. A more in-depth evaluation would yield greater insight into the assumption that integration of research experience into an early chemistry course can make a significant difference to students, including those typically underserved by college science programs. Such a study could focus on one or more of these impacts: student attitudes about science, scientific research, and their potential to become scientists; knowledge of science content and research; research skills; and future behaviors and choices related to science.

Choosing this option over those below assumes that additional data—beyond what was gleaned from the Stage One study—about the benefits to students are needed to persuade NSF to continue the initiative and/or to persuade the broader field that changing general chemistry is a worthwhile enterprise.

Option 2: Share practical lessons learned about the design of effective models. In-depth case studies of the principles, components, and key strategies of the five URC models would generate practical design knowledge that could be taken up by existing sites and potential new sites.

Choosing this option over the first assumes general agreement that adding research to the general chemistry experience is of benefit to students, and it assumes that the goals of the URC are feasible and the models replicable or adaptable. This option would be valuable whether NSF decided to continue the URC initiative or not.

Option 3. Assess the extent to which the model has potential to improve chemistry education more broadly. Such a study would focus on key elements of the model and development strategies; its feasibility for sustainability and scale; and its products in the form of faculty and administrator leadership, practical design knowledge, faculty and institutional collaboration and partnerships, and local research and evaluation results. This broader study would assess the extent to which and ways in which the URC initiative has potential to build field capacity.

Choosing this option would help NSF and the policy-makers who oversee NSF understand the extent to which and how the URC investment generated educational capital—e.g., knowledge, leadership, relationships, models—that could contribute in multiple and lasting ways to science education and potentially to society more broadly.

Option 4. To inform the design of a future initiative, i.e., "URC 2.0." A Stage Two study that focused on examining the URC goals and criteria, in light of the key features of the effective models built, would inform the revision and improvement of the next-generation URC initiative.

NSF would choose this option if they wanted to continue the URC initiative and if the Stage One results described above signaled a need to re-think and refine initiative parameters.

What not to study. The IR team also recommended to NSF what not to evaluate further, given the results of Stage One. They posited that further studies of the scientific quality and merit of the student research projects (including peer-reviewed research publications stemming from URC chemistry courses) would not be an efficient or productive use of evaluation resources.

IV. REFLECTIONS ON A STAGED EVALUATION PROCESS

This section examines the advantages and disadvantages of a staged approach, situates staged evaluation in a broad array of evaluation stances and approaches, and puts forth propositions about what a "model" Stage One study might include, based upon this experience and given proposed purposes for staging evaluation. Quotations are drawn from the live group discussion of the Stage One study of the URC and also from written responses to a draft of this report. Discussants/respondents included NSF program officers from several divisions (science and education), an independent evaluation expert serving as a consultant, and the IR evaluation team, including the consulting chemist and URC evaluators. The hope is that the example URC study and this discussion, together, can provoke further experimentation with and reflection on staged evaluation as a good option for studying large-scale, complex initiatives.

Advantages and benefits

Discussants identified two major benefits of a staged approach to evaluation: 1) Getting an early "reality check" on conceptualization and implementation of the initiative in the field, and 2) Designing full evaluations with greater cost efficiency and intentionality. A third benefit is that staged evaluations can inform staged funding of projects meant to go to scale. Below are discussants' perspectives on these benefits.

Early reality check on conceptualization and implementation in the field

Because a Stage One study is timed fairly early in the life of a large, complex initiative, it enables the funders to get an early view of grantees' interpretations of the solicitation. This preview can help funders see the congruence between their own vision of the initiative (as reflected in the solicitation) and the vision of the field in the form of new projects developed. One university chemist serving as a member of the evaluation team said the ground-truthing process was "eye-opening" because it revealed that:

> *…the perceptions of the aims and priorities of the URC initiative varied between the Chemistry Division program officers responsible for the program, the PIs in each of the collaboratives, and the wider education and research communities, including those trying to define the options for a staged evaluation.*

One researcher was "struck by the extent to which each project was driven by the values of the PIs, project teams, and their institutions. This was an unexpected and important insight for me, I suspect because we don't talk much about values in the evaluation of STEM initiatives." She added that variations arising from any sources could pose challenges for typical evaluations:

> *Logically, it might be assumed that projects designed and implemented to address the same programmatic goals would be similar and would be internally consistent. Yet, aspects of context, scale and the potential for variability can undermine even the best-intentioned efforts at uniformity… Thus, projects that appear to be similar, such as the URCs, may produce different results for subtle and unanticipated reasons. This situation poses a challenge, as evaluators typically seek to isolate key features that contribute to project outcomes and/or to determine elements of an effective project.*

A member of the research team implied that an evaluation defined before a Stage One study could miss important unintended outcomes:

> *This early look [the Stage One study] made it clear that the targeted outcomes for this initiative were not all equal and in some cases, what the project was delivering was more important than the targeted outcomes.*

Absent the reality check that a Stage One study provides, evaluators can overlook differences of values and context that lead to different models and often to more varied outcomes than funders anticipate. A Stage One study thus has greater benefit than enabling mid-course correction: it can also surface unexpected positive results such as new or unanticipated models, or insight into context conditions that optimize the investment. A long-time leader in NSF's Education directorate commented that this review of grantees' work could help funders decide on the important questions to ask:

> *I don't think there has ever been an evaluation of any one of those initiatives that I have been completely satisfied with. I think partly it is because at the beginning, you really don't know what questions to ask. It takes most of those projects a year to get up and running and so it seems to me that with a particularly large, complicated project, a Stage One evaluation just makes sense…to do this scan of things and see, okay…what does it really look like when you get out there? In a solicitation, you write it and sometimes we know what we mean, but those [ideas] get interpreted in lots of different kinds of ways.*

A former NSF officer offered a caveat related to the benefits of early ground-truthing, saying that Stage One studies should not occur "too early on the learning curve" because they can "disproportionately capture early start-up problems and focus on them to an outsized degree."

<u>Greater cost efficiency and greater intentionality in designing full evaluations</u>

One question related to evaluation resources is whether it might be more cost-efficient to delay the effort that goes into an *a priori* evaluation design (especially if it is assumed the design will change) and instead, to invest in the initial exploratory stage and then use that process to design an evaluation based on more grounded understanding of project theory and project reality. One NSF officer observed that evaluation funds have sometimes been used inefficiently:

> *A lot of evaluations with which I am familiar have required enormous amounts of data, some of which never gets used. So, what I am wondering is, if you did this kind of staged approach, would you be able to be more efficient… You are only asking questions that you really know you want answers to and you are not collecting information for the sake of collecting information.*

A number of discussants share the belief that a staged approach is advantageous to a funder because the results of Stage One studies make more cost-efficient and effective use of resources for evaluation. One said, "a staged evaluation…allows the funder to make grounded decisions about the usefulness of a more extensive evaluation." A NSF officer added "For large, complex initiatives, a staged initiative can be very constructive in identifying issues and foci, especially for units where evaluation expertise is not a core strength."

Below are two additional comments, each emphasizing the role of Stage One in the many decisions connected to evaluation of large, complex initiatives, decisions which have costs attached:

> *The benefits include the ability to make more informed choices between options for larger scale, more expensive evaluations, the potential of allowing mid-course corrections in subsequent solicitations, and the ability to focus on specific, more promising issues in a more in-depth, Stage 2 evaluation. Staged evaluation also seems like a cost-effective means of educating program officers in NSF research divisions about evaluation and perhaps would result in more thoughtful consideration of assessment/evaluation in the design of future programs and solicitations.*

> *Another important advantage of the staged approach is the ability to provide informed options regarding Phase 2. This "menu" approach to evaluation would allow a comprehensive look at certain aspects of a multi-dimensional program as an alternative to a more expensive, full-scale evaluation of the entire program. I see this as a great help to program officers and divisions of NSF in getting their arms around evaluation and better understanding the benefits of more rigorous approaches in making policy decisions.*

One person pointed out that a staged approach also adds more flexibility to the overall vision for evaluation:

> *What I find interesting about this, or at least I would like to explore it more, is that you may not ever go to the next stage, so it decouples this [initial investigation] from a big, long, expensive evaluation.*

One advantage to a "decoupling" of evaluation stages, according to the IR team, is that the research expertise can also be decoupled. A Stage One study carried out by one group could generate recommendations for a Stage Two study that would be better conducted by a research group with expertise or capacities specific to the selected focus of the Stage Two study.

A former NSF officer, while seeing real benefits to staged evaluation, cautioned that too little is known about the proportion of Stage One studies that would lead to fuller evaluations or would lead to the conclusion that further investment in evaluation is not needed:

> *[This is] a topic for further thinking about guidelines or parameters for when to recommend proceeding to Stage Two versus when to stop at the end of Stage One. It would be helpful to have a set of illustrative results of Stage One evaluations that indicate circumstances when you should stop.*

<u>Informing a staged approach to the funding of projects</u>

One participant noted that Stage One evaluation can help the funder not only assess the need for and frame a more effective full evaluation, but perhaps as importantly, help the funder assess whether the initiative itself is well enough designed to expand to a large scale. This participant, citing the 50-state State Systemic Initiative project, suggested that a Stage One study could have been used to inform NSF about SSIs in some states so as to make more informed investments in SSIs to additional states:

> *It is not just the evaluation part. Maybe we ought to go slower with funding. Maybe before funding 49 states, we have an opportunity to look at the program and we learn something before we fully go on out.*

Disadvantages and concerns

Discussants raised a number of significant cautions about a staged approach. For the most part these concern the fact that Stage One studies are inherently less rigorous than full evaluations, and that Stage One studies could feed the policy system's appetite for evaluation in a way that ultimately undermines fuller evaluation and thus policy decision making.

Potential overuse of Stage One studies for policy decisions

One evaluator with many years' experience conducting evaluations for state and federal policy makers cautioned that turnover in agency leadership and ever-shifting policy priorities can thwart a staged approach. Fuller, more rigorous, Stage Two evaluations could fail to materialize because of agency turnover or a premature demand for results. Stage One studies could thus stand in, by default, for the full evaluation:

> *If you are going to do a Stage One study for an organization or a set of policy makers who are about to change, you rarely will get to a Stage Two evaluation because the agenda is going to change. I know firsthand what happens—the lead foundation loses interest, the new leadership doesn't care, and it is a different set of questions. The other caveat in all of this is that policy makers, particularly, are worrisome for me to try and do Stage One and Stage Two evaluations, because policy makers are always ahead of the evaluators and they always want to make a decision and they can't wait for you to do your rigorous study. And so, what they tell us is 'we are going to make a decision whether you have any input or not, and so you might as well tell us what you have got.' I think that there is a danger there of never getting to the more rigorous part of it.*

He adds, "some funders can have rapidly changing strategies that prevent initiatives from maturing and evaluations from measuring mature initiatives." Another participant noted that NSF is typical of such funders, insofar as its funding schedules tend to give more clout to short-term, rather than longer term and more rigorous, evaluations:

> *One potential drawback of the staged approach relates to the relatively short lifetimes of many NSF initiatives, the turnover in the personnel responsible for the creation and direction of programs, and the time-dependent diminution of interest in detailed evaluation of specific programs. At some level, "seat of the pants" qualitative evaluations are always going on in setting priorities and making decisions about budgets.*

While Stage One evaluation has the legitimate purposes of questioning the need for investment in a full study and of illustrating where better and more data may be desired, certainly it is important to acknowledge the danger of mistaking Stage One for a full study when it is not designed for that purpose. These disadvantages call into question the overall purpose that evaluation serves and the relationship that it has to organizational and institutional habits and rhythms. There is often a fundamental mismatch between the time span of full and rigorous study and the need to make policy decisions at any level and launch programs.[2] One implication may be that Stage One studies, if formalized and conducted in disciplined ways, could provide more timely and also more valuable information than "seat-of-the-pants" studies done with less intention.

One participant believes that the advantages of a staged approach for NSF initiatives clearly outweigh the potential disadvantages:

> *The prospect that evaluation might never move beyond Phase 1 should be balanced with the current reality that many programs do not receive any formal evaluation, due both to the lack of evaluation expertise in the research divisions of NSF and the large budgets associated with full-scale evaluations.*

In contrast, one participant believes that even when staged evaluation is appropriate, evaluation should not cease after Stage One:

> *When an initiative has a weak research base and funders are relying on their hunches, a staged evaluation seems most appropriate. However, sometimes a funder may have already committed to a large long-term investment even without a solid research base. In this case, a staged evaluation makes sense as long there is a commitment to maintain some level of evaluation throughout the life of the initiative.*

Incompatibility with requirements to use controlled comparisons for evaluation

Another disadvantage of a staged approach is that it may simply not be in alignment with the types of investment a funding agency is making and with their demands for evaluation. For example, one discussant pointed out that the U.S. Department of Education, unlike NSF, often requires controlled comparative studies:

> *A staged evaluation may be problematic when a randomized controlled trial (RCT) is the desirable form of evaluation…different government agencies are likely to have varying receptiveness to a staged evaluation. The current environment at the U.S. Department of Education seems particularly unfriendly to the concept of a staged evaluation…the over-emphasis on RCTs has made efforts to base decisions on less rigorous methods largely ineligible for funding. While there are early indications that the practice of only funding RCTs may be changing, the political appeal of "scientific research" and the views of the entrenched bureaucracy at the DOE's Institute of Education Sciences will make it hard to find room for staged evaluation.*

It is important to recognize that a staged approach may well be inappropriate for projects funded as research-based experimental tests. In fact, it is important for funders to be clear about whether they are funding a true experiment based

[2] One sign that this is becoming an increasingly broad concern is that the Brookings Institution sponsored an event in December 2010 that focused on "deep-dive, quick-turnaround" education research. See http://www.edweek.org/ew/articles/2010/12/30/15brookings.h30.html

upon prior research (what some call a "medical model"), or are funding a project that is meant both to deliver services and to serve as a more exploratory design research effort. The staged approach may well be advantageous in the latter case. Proposals for staged evaluation in this context would likely need to combine staged study of some facets of the program with RCTs and to provide a very strong rationale for that approach, citing the multiple benefits of staged study identified in this paper.

Even when a staged approach is acceptable and underway in this context, a disadvantage can be that evaluators would miss the opportunity to gather needed data or create experimental designs if the results of the Stage One study indicated such needs for the Stage Two work. Comments from discussants, one about comparison groups, the other about baseline data:

> *If you don't have the system set up at the very beginning, you may not be able to answer the question that you decide later [as a result of Stage One] that you want to answer…My point is, the problem of comparison groups.*
>
> *For some NSF-funded projects, it is very important to establish a baseline against which progress can be measured over the duration of the award. I am not sure how this commonly-critical need for early evaluation data is integrated with a staged evaluation effort. My concern is that [a Stage One study] might introduce one more "moving part," with negative consequences. My sense is that [Stage One study] will need to be an adjunct to some other efforts that capture these baseline data.*

Discussants varied in their belief about these disadvantages. For example, one NSF officer person pointed out that investing in an experimental evaluation design, without exploring the implementation of the project on the ground, could result in evaluation results that are not useful from a management perspective. She suggested that, for NSF, the medical model is often *not* the preferred approach, especially with a new initiative where the expectation is for "big effects" that are not easily captured in experimental designs:

> *If you think of a paradigm like drug trials, then you need a comparison group and even small effects are very important and important to study, but in most of the studies that we are talking about, we are looking for big effects. If there is a tiny, tiny difference, we are not very interested. So if you are looking for big effects, then you don't need a drug trial. It seems to me that qualitative evaluation is very effective for noticing big effects.*

The distinction between initiative and project may be valuable in addressing this concern. An initiative tends to be large and complex, serving to advance a field broadly and even to stimulate a wide range of activities and models; there may be no relevant comparison group or clear baseline. But within a funded project that has taken a certain approach to interpreting the initiative's opportunity, a comparison study may effectively test part of that model and what baseline data are needed may be quite clear. For example, one URC site emphasized faculty development as a major objective. That site's project evaluator used comparison groups of non-participating faculty to assess the benefits of URC participation for faculty.

Further, for projects where there are rolling cohorts of participants or sites over the life of a project, time is not always the enemy of comparison studies or the identification of appropriate baseline data. In the URC initiative, sites varied in the intensity and duration of the students' research experience. With cohorts coming into introductory chemistry every year or even every academic term, one possibility for a Stage Two study could be to compare chemistry students' outcomes across a range of models, and baseline data could comprise outcomes of pre-project cohorts.

A broader concern about ill-defined institutional purposes for evaluation

For one NSF officer, the discussion of Stage One prompted a concern about the lack of institutional clarity about the role of evaluation generally. When a member of the IR team suggested that staged evaluation could strengthen the logical connections between initiative-wide and project-specific evaluation, this participant said that that discussion should take place in a broader discussion of evaluation purposes:

> *I think that is true if we had a common epistemology or a common sense of what evaluation actually does. We [at NSF] have never, at least in my experience of looking at program evaluations, had a common understanding of the purpose of doing evaluations. Why do we do them, what do we want to get out of them? Is it a formative, is it a summative, who is it for? We often answer technical questions… I think what I would push back and say is, the staged idea needs to be inserted into some broader approach.*

Several NSF officers agreed that funders need to "do a better job of thinking through those big questions [about the purpose of evaluation] before we even get to the program." One former NSF officer, after citing several advantages of staged evaluation, concluded that "the potential for use of a staged evaluation effort is uncertain at this time…The limitation, as I see it, is that it does not mesh well with how NSF currently funds and manages many of its science education projects. I think that this is the key area for future thinking."

Defining Stage One as "consultative" evaluation

Discussion of the advantages and drawbacks of staged evaluation led the group to define it more precisely by identifying the major purpose and audience for Stage One studies. One NSF officer suggested that Stage One does not serve any of the usual evaluation roles of formative feedback to projects, summative assessment of accomplishments and lessons learned, or proof of producing a specific outcome. Rather, he suggested that Stage One studies serve primarily a "management consulting role" for project funders and leaders:

> *It seems to me that the real value of this Stage One is more of a kind of management consulting role, rather than a number collecting and crunching mode…Because the first couple of years [of a complex initiative] are often a struggle, there is a strong advantage in getting some outside evaluation and management/organizational guidance to the group/center/partnership during this time period. My view is that NSF often does not do this well, and maybe a Stage One evaluation process would be very useful under the right circumstances to help the awardees get through this start-up phase well-positioned to solve important issues.*

Several discussants shared the view that Stage One study can, in one person's words, "identify weaknesses, and strengths, early in the project's life and to allow the project to make important adjustments." This person added that, "In a sense, a Stage One evaluation is not really an evaluation. Rather it is an information gathering process that has potential to give direction to the project and to inform the funder."

One former NSF officer noted a caution, though: changing the rules of the evaluation game to serve "management" could chafe PIs:

> *The traditional roles of NSF, the community (reviewers/panelists/site visit members), and evaluators will get tweaked, perhaps substantially, if Stage One evaluations become commonplace. This will involve some higher level of intrusion into a PI's freedom to operate with minimal oversight. Not everyone will welcome this.*

Discussants noted that if Stage One studies become defined this way, the expertise of team members is vitally important: their knowledge base needs to include the content of the initiative (in this case chemistry and chemistry education) as well as the organizational context (higher education):

> *If an evaluation team that you bring in doesn't know the vernacular of that culture and the modes of operating, it makes it that much harder for them to get to [the real story]…I think the assembly of the team around organizational expertise as well as subject matter is crucial, particularly in this management consulting mode, or you are never going to get through to the real behind-the-scenes and deeper level contents, because people will know how to tell you what you want to hear if you are not careful.*

The IR evaluation team was composed of evaluation researchers with extensive knowledge of NSF culture, priorities, and programming; of chemistry education; and of the administrative culture of IHEs. The team bolstered their internal knowledge capacity with independent consultants with expertise in chemistry and IHE chemistry education, and in evaluation of federally funded initiatives in multiple content areas. The team also included two evaluators of URC sites. One discussant noted that this kind of "cross-fertilization" is of real benefit to NSF. The IR team suggested that the composition of the team for Stage One studies is vitally important because Stage One evaluation does not provide only data. Rather, a Stage One study requires a team that has sufficient background knowledge to adopt a consultative stance and provide *interpretive* perspective, along with some data. With an optimal combination of data and expertise brought to bear on it, the Stage One study can make greater use of limited time. The lead IR evaluator put it this way:

> *For the timescale that we are talking about and the depth of interaction that we are allowed to have, when you come back, as we did with the Stage One today, you are bringing back some data, but it isn't just bringing back data, it is bringing back your expertise and your perspective as an outside evaluator that presumably has some perspective on things and so it is part consulting in a way and part data gathering. I think it is a mixture of both. It isn't purely 'here is a bunch of data that is going to help you.' It is also outside perspective and judgment and ideas about your initiative.*

Discussants expressed some skepticism about the ability to make staged evaluation routine, because of the need for researchers with exceptional qualities. One participant cautioned that if Stage One evaluations become commonplace, "it is almost inevitable that this diligence in selecting a team will erode." Another participant suggested that having research team members/consultants who are part of the project or are former NSF staff are vitally important for Stage One, but having those team members for a fuller Stage Two evaluation could produce the appearance of bias. This is because the findings of Stage One are less about the project/initiative and more about future studies of potential benefits and lessons learned; thus, bias is less of a concern for this first stage.

Another participant expressed an even stronger concern that the enticement of cost savings could blind a funder to the level of sophistication needed to make Stage One studies of high quality:

> *While I agree that there are individuals who can navigate the tensions between [evaluation and consulting] roles (and I think the team on this evaluation was successful), my sense is that there are a very limited number of individuals with the skills and experience to pull it off. This is problematic because funders may be enticed by the benefits of a staged evaluation (especially its cost savings) without understanding the qualifications of the staff needed to conduct a staged evaluation. In other words, not just any evaluator can do this work.*

Finally, one person noted that the five funded URC sites had independent external evaluators, plus internal evaluators, in addition to this Stage One study of the fuller initiative. The meanings of these terms and the realities of relationships are nuanced and evolving. Several participants agreed that definitions of "internal," "external," and "independent" evaluators is "vexing" for NSF, with relationships forming and roles shifting as projects evolve over time. Perhaps greater clarity in framing the purpose and audience for evaluation can help clarify (e.g., consultative, for the funder; or formative, for the project) can help sort out relationships and roles.

Design principles toward a model of Stage One evaluation

One theme that arose in the discussion is that Stage One might be a more formal name for what evaluators often do as a matter of course even when they have defined an *a priori* evaluation design. Two comments illustrate this theme:

> *A good evaluation does a lot of what you [the IR team] are doing in the first part of the evaluation. You are blocking it out as a stage and asking for a developmental progression, but just because you have set it [an evaluation plan] up ahead of time, doesn't mean that this stage doesn't happen.*
>
> *One significant benefit is that [Stage One evaluation] validates a de facto evaluation practice in which we often engage during the first funded year of a complex, large-scale project or program. Even though we submit rather detailed evaluation plans with grant proposals, we generally modify those plans substantially during the first year of project planning and/or implementation.*

Using the language of Stage One and Two evaluations, then, may be a way to formalize or identify what "good" evaluation often already entails.

Here we draw from the URC study to posit design principles for a staged approach, that is, to define more clearly when staged studies are most appropriate and what features of Stage One studies are likely to enhance their advantages and diminish their disadvantages.

The purpose. Stage One studies may be best defined as consultative in their overall purpose. That is, they aim to provide funders and other institutional representatives with both data and perspective that can inform them about the realities of the initiative as it is interpreted and implemented in the field. The results of a Stage One study thus ground funders' and other leaders' decision-making about multiple aspects of the initiative—including, but not limited to, the need for and potential foci and designs of a Stage Two full evaluation.

The object of study. Stage One studies are initial documentations and examinations of large, complex initiatives about which little is yet known from direct observation. They are by nature exploratory. That is, they are not "tests" of an implemented innovation or program where the variables and expected outcomes can be clearly defined. Rather, they generate propositions and tentative findings to be probed further if deemed important.

The relationship of Stage One evaluation to other evaluations. Stage One studies have a well-defined purpose that is distinct from other evaluation purposes. While a funded site or project within a large and complex initiative or program may have an internal or external evaluator, or both, that project-specific evaluation does not serve the same purpose as the evaluation of the larger initiative. The project/site-specific study serves the project, primarily for a formative purpose and sometimes for a summative purpose within the scope of the site. The Stage One study, on the other hand, primarily serves the funder, providing data that serves management functions when the object of study is a large multi-site initiative or program. Management functions certainly include decision-making about full evaluation, and might also include decisions about funding additional sites, about changing initiative objectives, on expanding the initiative, and so on. Project-specific studies are important contributors to Stage One studies. And Stage One studies can lead to many forms of future evaluation, including rigorous tests of initiative designs and specific outcomes.

Composition of the research team. It is important to involve researchers who have enough familiarity with the content and objectives of the initiative to come up to speed very quickly and to make informed interpretations of observations. While the appearance of possible bias is less of a concern for Stage One than for Stage Two, it is important to have transparency in evaluation team membership and rationale for the selection and role of each member.

Timing. It is important that a Stage One study take place fairly early in a project so that findings can be of use and so that it is not too late to frame and conduct a fuller evaluation if indicated. And it is ideal that a Stage One study be conducted over a short time frame of 6-9 months, again so that the findings are timely. However, Stage One study should also not begin and end within the period of the onset of a large initiative, or all that would be documented is the early struggle of capacity building and starting up; rather, it is best done fairly early, but after some observable coalescence of project activity.

Methodology. Stage One study relies heavily on qualitative methods. Further, the methods include a strong reliance on interpretation that experts can bring to evidence. Stage One studies also rely on data gathered from internal project evaluations, which may be both qualitative and quantitative.

Sources of data. Stage One studies rely on interviews with project funders and designers, on reviews of key documents that reflect design intention and implementation, on interviews with project implementers, and on examination of artifacts that reflect implementation vision and reality. Unless the initiative itself relies only on virtual interactions, Stage One studies also probably rely on some in-person observation of events, meetings, and/or activities to gain a grounded sense of participant experience.

Sharing of results. Because staged evaluation serves primarily a consultative purpose, it involves an iterative mode of communication and reporting, in which the research team is well informed of management perspective, gathers data in the field, and engages management—ideally face-to-face—in joint examination and interpretation of field observations. This is in contrast to working at a distance from management and providing data and conclusions in a non-interactive mode.

Final reflection on implications for NSF

A number of discussants referred to features of the NSF organization that make a staged approach to evaluation—including refinement and systemization of Stage One studies—especially potentially valuable. These included:

- Quick turnover (2-4 years) of staff in the Science Directorates
- Lack of evaluation experience or expertise in these same directorates
- Insufficient "cross-pollination" between directorates and programs
- Insufficient clarity about the value and purpose of evaluation when initiatives are launched

Given these conditions, one person suggested that if NSF continues to explore staged evaluation as a model, that the Education and Human Resources directorate could play a central coordinating or clearinghouse role:

> *The Education and Human Resources (EHR) Directorate should be involved (and possibly have both a financial and program officer stake) in Stage 1 evaluations in order to serve as a central depository of experience and information for future evaluations, staged or otherwise. The research divisions of NSF typically have substantial turnover in program officers, staff, and division directors, and generally do not have the personnel or expertise, particularly regarding evaluation, to develop an institutional memory of these kinds of projects. In addition, there is little communication across NSF of initiatives such as the URC and/or their evaluation. EHR needs to function as a center of expertise to inform the design and evaluation of future projects that combine the two central themes of NSF: research and education.*

At the very least, it seems evident that a staged approach to evaluation has sufficient potential advantages, particularly for NSF, that further exploration is warranted. This is best accomplished through the development of additional examples of Stage One evaluations and examination of their results.

REFERENCES

Trevisan, M. S. & Huang, Y. M. (2003). Evaluability assessment: a Primer. *Practical Assessment, Research & Evaluation*, 8(20). Retrieved October 9, 2008 from http://PAREonline.net/getvn.asp?v=8&n=20

Trevisan, M. S. (2007) Evaluability Assessment From 1986 to 2006 (2007) *American Journal of Evaluation*, Vol. 28, No. 3, 290-303.