

EXPLORING THE CONCEPT AND PRACTICE OF STAGED EVALUATION AS A MORE VALUABLE APPROACH TO EVALUATING LARGE, COMPLEX EDUCATION INITIATIVES

Laura Stokes and Mark St. John
Inverness Research

Funders and evaluators alike face the challenge of designing cost efficient and appropriately focused evaluations of large, complex education improvement initiatives. Such initiatives typically involve multiple institutions and are framed around multi-layered logic models that include multiple goals and that aim at both proximal and distal outcomes, some of which cannot be pre-defined with confidence. Requirements to pre-specify evaluation designs early in the development phase may not make sense or be ultimately productive. In response to this challenge, Inverness Research (IR) investigated the concept and practice of *staged evaluation*, using a study of the National Science Foundation-funded Undergraduate Research Collaborative (URC) in Chemistry as a sample and illuminative case.¹ The study explored 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” evaluation to allow for a more thoughtful and productive final overall evaluation.

THE CONCEPT OF STAGED EVALUATION

Staged evaluation is reminiscent of “evaluability assessment,” developed some 30 years ago:

A strategy that can be used to determine the extent to which a program is ready for full evaluation, is known as evaluability assessment...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 and Huang, 2003)

A Stage One study is a brief and exploratory effort of “ground-truthing” an initiative. By ground-truthing, we mean gathering data from the work and participants “on the ground” to assess the extent to which project actualities are congruent with the initiative’s stated theory of action. Stage One study clarifies the need, purpose, and design of a fuller and more rigorous evaluation, a Stage Two study.

THE CORE CASE: NSF’S UNDERGRADUATE RESEARCH COLLABORATIVE IN CHEMISTRY

Despite numerous calls for reform, early chemistry experience for most college students has remained unchanged for decades. In 2004, NSF funded five multi-campus college/university Collaboratives in different regions of the U.S. to promote creation of new models of chemistry education that would infuse authentic research into freshman and sophomore courses. Proposed Collaborative projects were evaluated on six criteria:

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*

¹ The study was funded by the National Science Foundation, Award # 0909720: Small Grant for Experimental Research. Inverness Research is a national educational research and evaluation firm headquartered in Inverness, CA. See www.inverness-research.org.

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.*

Given the URC's size, complexity, and multiple and ambitious goals, it was not obvious to NSF how to effectively or cost-efficiently design a comprehensive and in-depth evaluation. The URC program thus served as an authentic example of an initiative where staged evaluation approach could be productive.

THE APPROACH TO EXPLORING AND TESTING STAGED EVALUATION

In its cost, time-frame, and depth, a Stage One evaluation falls between a Committee of Visitors (COV) review and a comprehensive study with pre-defined outcome measures. The following characterizes the Inverness Research Stage One study of the URC initiative:

Purposes

- To document the purpose and theory of action (logic model) of the URC
- To ground-truth the URC theory of action
- To describe the landscape and context of undergraduate chemistry education served by the URC grants
- To arrive at preliminary assessments of short-term outcomes of the initiative at multiple levels
- To identify critical issues in the URC program and in 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

Design and data sources

The study involved two levels: 1) study of the URC, and 2) study of the URC study as a case of Stage One evaluation.

1) URC study: The Stage One study of the URC was a short-term qualitative study, with data collection focusing on documenting the following: the initiative's theory of action from multiple perspectives, activities at multiple sites, contextual challenges and affordances, emerging strengths and weaknesses, and identification of potential immediate and long-term outcomes. The timeframe of the study included 8 months of research and 2 months of data analysis and presentation/review of findings by NSF staff and others. The research team included IR researchers with science education backgrounds as well as university chemists/chemistry educators as consultants. Key data sources were: Interviews of NSF staff who designed the initiative and with PIs and participating faculty at all five funded Collaboratives; observations of Collaborative activities designed to showcase student research; documents, including NSF documents describing the initiative, as well as internal evaluation documents from the five Collaboratives, including student participation data and some student work products. NSF URC and other staff participated in a review of findings related to the URC.

2) Study of the URC study as a case of Stage One evaluation. An IR researcher working independently from the URC study team led this portion of the study, which focused on interviews with the evaluation team, reviews of evaluation documents, and documentation of the NSF review of findings. NSF science, education, and evaluation staff, as well as an independent evaluation expert, participated in two stages of review and comment on the advantages and limitations of staged evaluation approach.

THE RESULTS OF THE STAGE ONE STUDY

Tentative findings about the URC Initiative

Core features of new models

- The five collaboratives produced widely varying models in response to the initiative.
- In all models, new curriculum modules served as the vehicle to promote authentic chemistry research. Modules varied in intensity and duration of research opportunity.
- Students had opportunities to share their research publicly. These varied in audience and formality.
- To varying degrees, modules promoted collaboration among faculty, students and faculty, and students

Preliminary outcomes of the URC projects

- Students found the new research opportunities to be more engaging than other lab formats, and they valued the opportunity to work with faculty on authentic research projects.
- Where research modules were more intensive and systematic, the new research experiences motivated students to pursue additional research and to become more interested in research-based careers.
- At smaller universities the project built research capacity of faculty.
- The project increased the number and ethnic diversity of students in research courses.
- Where research modules were more intensive and systematic, both students and faculty changed their perceptions of student roles, seeing students as scientists.

Challenges

- Engaging students in authentic research required greater faculty commitment, cost, and space than traditional courses.
- Models that showed greater potential for replicability or sustainability were those with least intensive new research modules.

“Implementing early student research faces significant institutional and discipline barriers, but our study offered evidence that... early participation in student research has strong potential for changing the teaching and learning of science at the college level.” (Horsch, et al., 2012)

Implications for Stage Two evaluation of the URC

Preliminary findings about the URC initiative showed that variation in models occurred because collaboratives chose to focus on different subsets of the many ambitious goals linked to the grant. Thus, Stage One study portrayed back to the funders the field’s perspective and capacity related to initiative goals.

Stage One study also generated an array of choices for fuller Stage Two study. They vary on the audience, purpose, and focus of inquiry:

Option 1. Assess contributions to students, with special emphasis on increasing access to underserved students.

This option assumes additional data are needed to persuade NSF and the broader field that changing general chemistry is worthwhile.

Option 2: Share practical lessons learned about the design of effective models. This option assumes that the new models benefit students and are feasible and replicable/adaptable.

Option 3. Assess the extent to which the model has potential to improve chemistry education more broadly.

This option would document 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 society.

Option 4. Inform the design of a future initiative, i.e., “URC 2.0.” 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. Further studies of the scientific quality and merit of the student research projects would not be an efficient or productive use of evaluation resources.

LESSONS LEARNED FROM THE CASE

Advantages

Stage One as “Consultative evaluation.” 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 called, as one reviewer coined the phrase, “consultative evaluation.” By serving a consultative purpose to the funder or designer of a large-scale initiative, Stage One study can be seen as a version of developmental evaluation:

The idea behind developmental evaluation is that the evaluator can assist in helping project leaders design, review and refine the design and implementation of the project work. Developmental evaluation involves processes and activities that support program, project, product, personnel and/or organizational development (especially the latter). The evaluator is part of a team whose members collaborate to conceptualize, design, and test new approaches in a long-term, on-going process of continuous improvement, adaptation, and intentional change. (Patton, 1994)

More cost-efficient and effective evaluation. A Stage One study can assist greatly in framing fuller (Stage Two) evaluations. First, the Stage One study produces a more grounded theory of action for the initiative, which provides a firmer basis for the design of a more in-depth subsequent evaluation. Fuller evaluations can focus on design features, implementation variables, and particular outcomes that emerge from Stage One study. Options can be presented that serve different evaluation audiences and purposes. Stage One studies may, in fact, lead to the conclusion that further evaluation may not be warranted. Either alternative—a better-designed full study or the decision not to do a full study—is more cost-efficient than designing full evaluations *a priori* and then discovering that the evaluation designs, foci or data definitions are not well aligned with program actualities or revised program goals.

Potential 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 may be less appropriate and advantageous for projects that require randomized controlled trials (RCT) as the only legitimate evaluation design. On the other hand, some reviewers suggested that a Stage One study could inform the logic model assumed by a RCT and thus improve its design and cost efficiency.

Conclusion

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.

REFERENCES

- Patton, Michael Q. (1994). Developmental Evaluation. *Evaluation Practice* 15(3): 311-319. JAI Press.
- Stokes, L. 2011. Staged evaluation: An approach to more cost-efficient and useful evaluation of large and complex initiatives. Final report for a Small Grant for Experimental Research (SGER), funded by the National Science Foundation, award # 0909720. Inverness Research. Available at http://www.inverness-research.org/abstracts/ab2011-04_Rpt_SGER-URC_REPORT_FINAL_041511.html
- Horsch, E., St. John, Mark, and Christensen, R. 2012. A case of reform: The Undergraduate Research Collaboratives. *Journal of College Science Teaching* (41) 5: 14-19
- 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>