Ensuring STEM Literacy: a National Conference on STEM Education and Public Outreach
ASP Conference Series, Vol. 483
James G. Manning, Mary Kay Hemenway, Joseph B. Jensen, and Michael G. Gibbs, eds.

Defining and Measuring Impact for Ourselves

Hilarie Davis[1] and Daniella Scalice[2]

[1]*Technology for Learning Consortium, Inc., 75 Sauga Avenue, North Kingstown, Rhode Island 02852, USA*

NASA Astrobiology Institute, NASA Ames Research Center, Mail Stop 247–6, Moffett Field, California 94035, USA

Abstract. This paper describes a practical method to flexibly and robustly measure the impact of EPO programs.

1. Introduction

1.1. What Is Impact?

“Impact” is a word that has taken on several meanings and is used today in various ways by different people to validate their efforts in EPO. Perceived impact can even be used to decide which projects should be cut. But we as an EPO community—with our numerous, diverse projects—and those who would judge our work do not share a common definition, and often nose counts or “successful” implementations (“I thought it went well!”) are mistaken for high impact.

Also, despite the fact that no two of our projects are alike, we are being asked to use standardized evaluation forms. The data we are asked to provide to senior-level managers focuses on congressional districts and participant demographics (outputs), and not on the outcomes we intended for our participants or the real meaning of our work and its value to our audiences.

How then should impact be defined and measured? And how can we as a community systematically do this in a way that incorporates scholarship, reflects grassroots experience, accounts for the unique nature of our projects, and maintains our autonomy to measure uniquely meaningful outcomes? How do we find a way that will ultimately make us better practitioners, stronger contributors to the field, and more frequent and more deserving grant recipients?

1.2. Origin of the Process: the NASA Astrobiology Institute

As a virtual, collaborative research organization, the model for EPO within the NASA Astrobiology Institute (NAI) has reflected the Institute’s distributed nature. Each NAI team has been responsible for proposing its own unique EPO projects, and each team does different activities with varying levels of evaluation. In this light, NAI can be thought of as a microcosm of NASA’s Science Mission Directorate (SMD); indeed, NAI is part of SMD. NAI even employs an EPO coordinator to maximize communication and collaboration, much as the SMD EPO Forums do for the SMD missions and projects.

Embedding EPO in SMD missions and programs like NAI has enabled innovative, meaningful EPO projects and deeply engaged scientists. SMD EPO projects are energized by being seated at the point of discovery and capitalize on direct scientist involvement. While the merits of this structure are obvious, obtaining a systematic measure of the impact of the NAI or SMD EPO Program as a whole has been problematic.

Inspired by the current dilemma surrounding the use of the word impact, and to address this issue of taking a systematic, standardized measurement of impact, NAI engaged a professional educational evaluator, Hilarie Davis (Ed.D., Technology for Learning Consortium, Inc.) in the summer of 2012 to undertake the creation of a grassroots, systematic, rigorous, and easy-to-use process for defining and measuring impact without sacrificing the individuality and autonomy of each project. We started by working with the NAI EPO Leads to develop the process and are now engaged in a pilot program with them.

2. Measuring Quality at Each Stage of a Project's Life Cycle

Initially looking at the NAI EPO portfolio and asking strategic questions of the NAI EPO Leads (~20 individuals), we investigated these questions: What activities are we currently doing? What are our objectives? How do we each define impact? How are we currently measuring impact? What impact are we having?

We organized the information about each of the NAI projects, including objectives, activities, and participants. We added data about level of effort, EPO activity type, and astrobiology content addressed. From these data, we were able to identify what we are doing and how we are doing it, but we still did not have a measure of impact that would work for different kinds of activities.

It became clear that looking at impact is hindered if one only looks at the "evaluation" stage of a project one typically associates with the "end" of an implementation. The concept of a project's life cycle emerged, and it was recognized that evaluation methods are embedded at every stage (Fig. 1).

Since optimum health or quality at each stage supports having good evidence of high impact on participants, the group set out to define quality at each stage, from Needs Assessment (NA), to setting Goals and Objectives (GO), to Design (D), Implementation (I), and Outcomes Assessment (OA). This eventually took the form of a single rubric with indicators of quality (fair, good, very good, excellent) for each stage (Fig. 2).

We went on to create the template in Figure 3 that tabulates ratings for the quality of the project throughout its life cycle, embeds the calculation for overall project quality, and documents both what has been done (reflection) and ways to increase quality (planning). The formula takes the scores at each stage, derived from the rubric above, and averages them into an overall score. Because our focus is on the impact of our projects on participants, Outcomes Assessment is counted twice and weighted in the formula:

$$OverallProjectQuality = \frac{NA + GO + D + I + 2(OA)}{5} \qquad (1)$$

High (Excellence) = 3.7–4.8

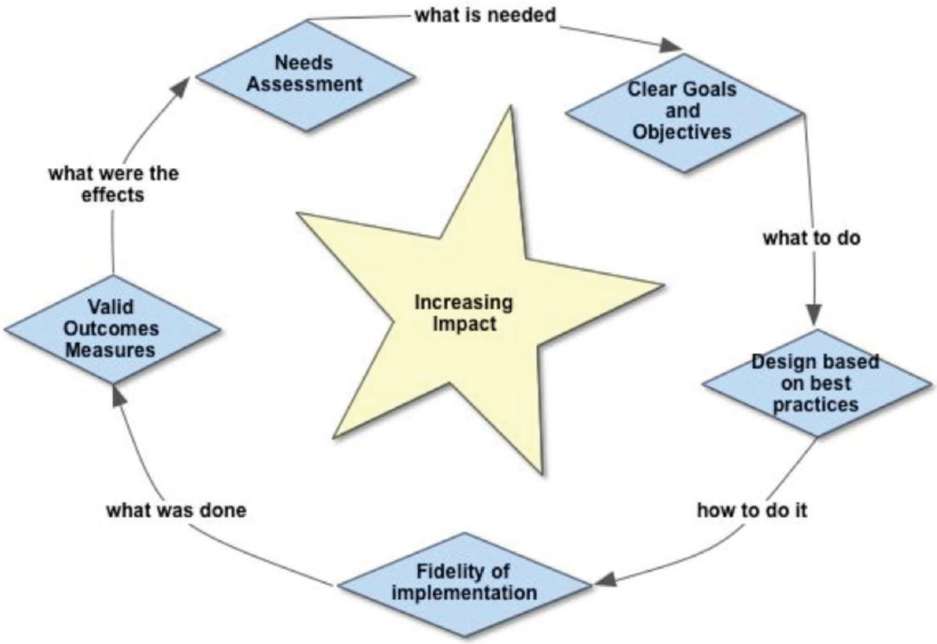


Figure 1. The five stages of a project's life cycle (H. Davis, May 2013).

Moderate (Could Improve) = 2.4–3.6
Low (Developing) = 1.2–2.3

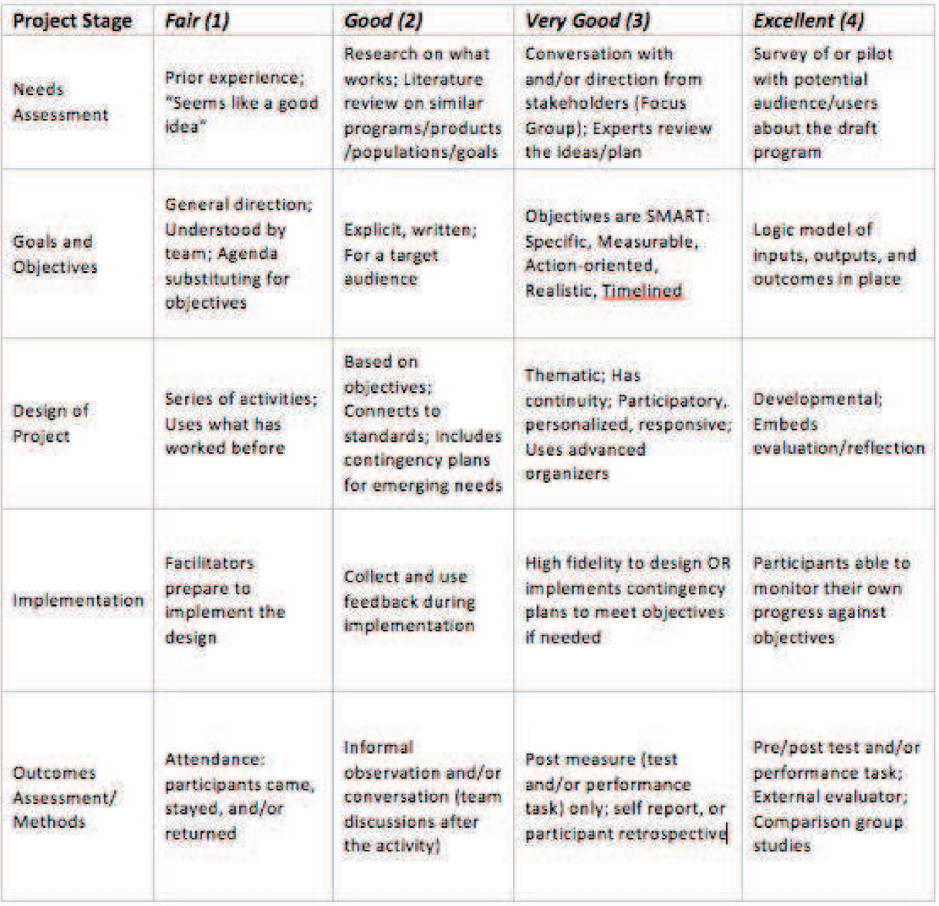
This felt like progress since we could now characterize how well we were doing at each stage of the cycle, and with further work, how we could use the information gathered about each stage and ideas for future planning to ultimately increase the impact of the activity. Now we could turn to looking more closely at outcomes and impact and how they have been measured.

3. Defining and Measuring Impact

What kinds and levels of impact were NAI EPO projects having? How was impact being measured? How confident were we in our findings about impact? As we explored the data on NAI EPO projects further, we discovered that in many cases, the only evaluation that was occurring was formative and informal—staff gathering data and observations to help improve a project's next implementation. We also found that some of the stated project objectives were describing the project activities (process objectives) rather than outlining the desired outcome for the participants. NSF's categories of impact [1] (Friedman 2008) proved useful to frame our thinking about the nature of the impact on a participant's "BASIK" [2] (Behavior, Attitude, Skills, Interest, and Knowl-

[1] `http://insci.prg/resources/Eval_Framework.pdf`

[2] NSF's original order of the impact categories was changed by Bradford T. Davey, TLC Inc., in February, 2013 to create the acronym "BASIK."



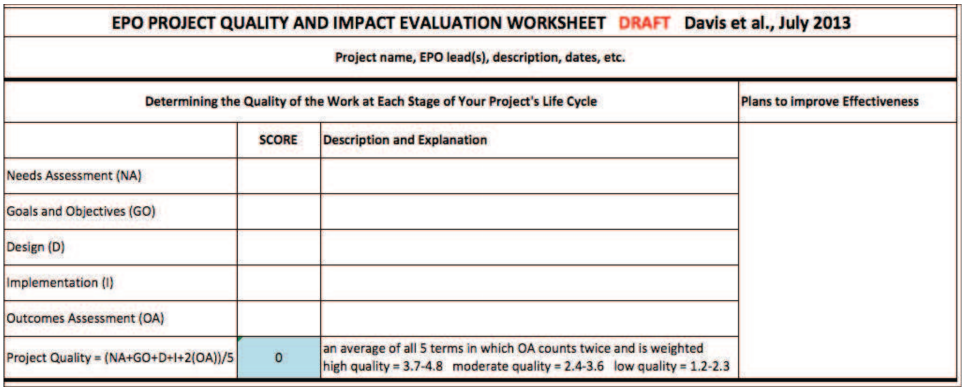
Project Stage	*Fair (1)*	*Good (2)*	*Very Good (3)*	*Excellent (4)*
Needs Assessment	Prior experience; "Seems like a good idea"	Research on what works; Literature review on similar programs/products /populations/goals	Conversation with and/or direction from stakeholders (Focus Group); Experts review the ideas/plan	Survey of or pilot with potential audience/users about the draft program
Goals and Objectives	General direction; Understood by team; Agenda substituting for objectives	Explicit, written; For a target audience	Objectives are SMART: Specific, Measurable, Action-oriented, Realistic, Timelined	Logic model of inputs, outputs, and outcomes in place
Design of Project	Series of activities; Uses what has worked before	Based on objectives; Connects to standards; includes contingency plans for emerging needs	Thematic; Has continuity; Participatory, personalized, responsive; Uses advanced organizers	Developmental; Embeds evaluation/reflection
Implementation	Facilitators prepare to implement the design	Collect and use feedback during implementation	High fidelity to design OR implements contingency plans to meet objectives if needed	Participants able to monitor their own progress against objectives
Outcomes Assessment/ Methods	Attendance: participants came, stayed, and/or returned	Informal observation and/or conversation (team discussions after the activity)	Post measure (test and/or performance task) only; self report, or participant retrospective	Pre/post test and/or performance task; External evaluator; Comparison group studies

Figure 2. Project Cycle Rubric (H. Davis and D. Scalice, July 2013).

edge), so we adopted them into the process. Overall we found that NAI EPO projects were having impacts on participants in many areas, but to a large degree, tools through which to measure it and language in which to express it were lacking.

We found that the NAI EPO leads were paying close attention to impact on participants, monitoring the effectiveness of what they were doing, and getting good results. In some cases, impacts on participants were being measured directly. For example, in a situation in which the objective is for participants to make gains in content knowledge ("K" from BASIK), a pre/post test or performance task was administered and the data from it show most or all of the participants making gains in their knowledge. Those are high-impact results collected through the use of a rigorous, high-confidence evaluation tool.

In other cases, impact was not being measured as formally or rigorously. For example, an all day workshop for educators is offered. They are engaged throughout the day, seem very interested, and at the end of the day they applaud the team and ask when the next workshop is. They talk about how they will implement what they learned in their own classrooms. The EPO lead rates the impact of this workshop as pretty high.



EPO PROJECT QUALITY AND IMPACT EVALUATION WORKSHEET DRAFT Davis et al., July 2013			
Project name, EPO lead(s), description, dates, etc.			
Determining the Quality of the Work at Each Stage of Your Project's Life Cycle			**Plans to improve Effectiveness**
	SCORE	**Description and Explanation**	
Needs Assessment (NA)			
Goals and Objectives (GO)			
Design (D)			
Implementation (I)			
Outcomes Assessment (OA)			
Project Quality = (NA+GO+D+I+2(OA))/5	0	an average of all 5 terms in which OA counts twice and is weighted high quality = 3.7-4.8 moderate quality = 2.4-3.6 low quality = 1.2-2.3	

Figure 3. Project Quality Worksheet Template (H. Davis and D. Scalice, July 2013).

But the rigor of the method in obtaining the data is low, and therefore our confidence in the rating of "high impact" is low.

It was analyses of the NAI EPO projects like these that finally led to a definition of impact and a way to measure it:

- **Impact** is defined as the intended and unintended effects on the Behavior, Attitudes, Skills, Interests, and/or Knowledge (BASIK) of participants. Impact is determined based on the *data you collect as evidence of impact (the results)* and the *rigor of the methods and measures* you use to collect those data.

- **Results** = *What is the data saying about how well the objective was met?*

- **Rigor** = *How confident can we be about the data based on how they were collected?*

To rate the impact of your activity, you must consider *both* the results/data and the tools used to collect them:

- **Results** × **Rigor = Level of Impact**

3.1. The Results: Data Collected as Evidence of Impact

The first part of the impact-measurement process focuses on the results: what do the data actually say about the impact you are having on participants' BASIK? For each objective, one looks at the data and rates them as high (3), medium (2), or low (1) impact. *This rating is independent of the evaluation methods used.* For example, for the objective of teachers creating lesson plans and using them in their classrooms: if the data show that most or all of the participants created and/or are implementing lesson plans, the rating is 3, high impact. For the objective of students learning more about extreme environments: if the data show that only 20% of participants could correctly define that term after the event, the rating is 1, low impact.

We purposely did not develop a formula, tool, or method to define how the results/data are rated. We felt keeping the interpretation of the data in the hands of either the project lead and/or the project's external evaluator, without constraints, was the best approach. Through the pilot test of this process we were able to assess 11 projects from around the SMD EPO community, and we found EPO leads were able to readily rate the impact as low, medium, or high based mostly on their own expectations for impact. Dr. Davis's independent assessment Soncurred with their self-ratings in 100% of the cases.

3.2. The Rigor of the Tools Used to Collect Data

After obtaining a results rating, the rigor of the tools used to generate those results is rated on a 1 to 3 scale. There are many types of tools available, and each one has a different level of rigor.

For example, a facilitator's observations of the level of engagement of participants does not provide high confidence in the claim that the participants have learned content or are feeling more confident (attitudes). Observations such as this, thought of as a data-collection tool, would be rated as a 1, for a low level of rigor. A post-only survey in which participants self-report their gains on BASIK rates as a 2, or moderate rigor since that knowledge may have pre-dated the experience. A pre/post test of content knowledge provides high confidence that attitude changes and/or participant knowledge gains after the experience were very likely the result of that experience, and rates as a 3, for high rigor. Verifying participants' lack of knowledge ahead of time gives you more confidence in high post-test scores, and at the same time alerts participants to what there is to know.

For each objective, one looks at all the tools used to collect data. The rigor rating is the most rigorous method you used to collect impact data (the tool with the highest score). In the case where many tools were used to collect data on a single objective, we consider summing the rigor ratings, but that could result in a high rigor rating for many low-rigor methods and measures.

Although the educational research and evaluation disciplines provide guidance about the rigor of various methods and measures, we could not point to a simple, easy-to-use resource for our EPO leads to use to identify the rigor of their methods, so we developed the following resource. It is designed to provide ratings of the rigor of different types of methods and measures commonly used to collect data on participants' BASIK. In the pilot study we found that even those EPO leads unfamiliar with evaluation were able to correctly rate the methods and measures they used. In addition, they could see from the rating the kinds of tools they could use to increase the rigor of their evaluation of impact.

3.3. How To Obtain Evidence of Impact (Methods and Tools)

The best evidence of effects on *knowledge, skills, and behaviors* comes from:

- Pre/post tests (3)
- Pre/post performance tasks such as lesson plans, models, concept maps, drawings, etc. (3)
- Evidence such as student work samples evaluated for knowledge; before and after (3)
- Post-test/task only plus follow-up after some interval to assess retention or implementation:

  - If the follow-up is observed directly or documented via evidence such as student work (3)
  - If the follow-up is self0-reported (2)

- External evaluator observation (2)
- Case study by external evaluator (2)

The best evidence for effects on *attitude and interest* comes from:

- Observation before and after of choices people make (3)
- Pre/post self-report (2)
- Pre/post reflections of status/condition noting changes with reasons and examples (2)
- Case study by external evaluator (2)

If you only get post data, you cannot be sure they did not already know the content or were able to do the skill before the EPO activity:

- Post-only test, performance task, or product of learning (2)
- Post-only retrospective (2)
- Post-only evidence such as student work samples evaluated for knowledge (1)
- Post-only survey (1)
- Follow-up surveys after some interval (1)
- Post-only reflection on what was learned (1)
- Post only interview (1)

Evidence of potential effect:

- Agenda shows what was intended (1)
- Attendance shows that people came (1)
- Web hits as a form of virtual attendance (1)
- Web spikes after a live event indicate follow-up interest by participants (1)
- Facilitator reports of engagement indicate awareness (1)
- Anecdotes about the nature of the impact (1)

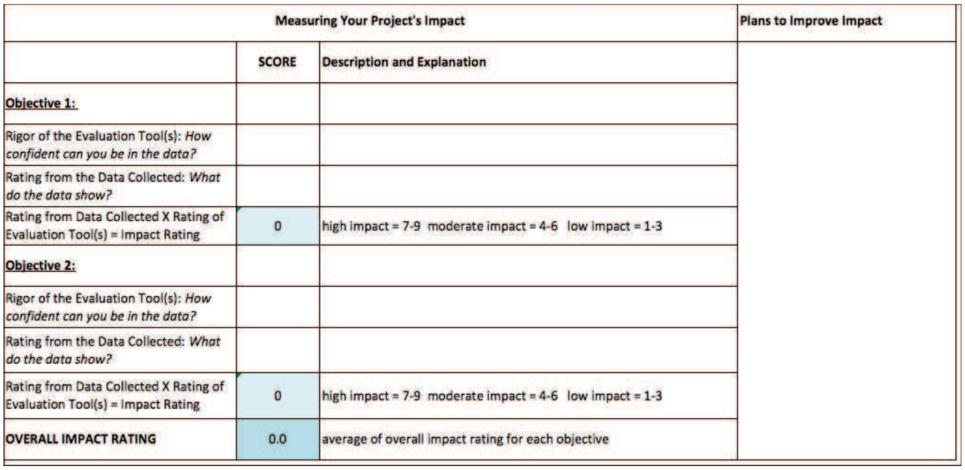
3.4. Taking a Measurement of Impact

Again, Results × Rigor = Impact, wherein, using a 1 to 3 rating system for both results and rigor of the tools, high impact is 7 to 9, moderate impact is 4 to 6, and low impact is 1 to 3.

For example, say there is a week-long "Astro Camp" for students to develop their understanding of key concepts around the origins of life, and the objective is for them to increase their content knowledge. When they arrive, they answer several multiple choice and open-ended questions about the origins of life to find out what they already understand, or think they know. At the end of the week, they are asked the same questions and the results are compared. The difference is startling: you see a change for all students, from being unable to answer most of the questions at the beginning of the week, to being able to answer them all correctly and articulate their ideas in open-ended responses. These results are rated a 3, or high impact. The rigor of the pre/post test is rated a 3, or high rigor. 3 × 3 = an overall rating of 9, which indicates high impact.

An example of low impact results collected by a moderately rigorous tool would be: for the objective of having undergraduate interns be able to prepare a research paper for publication (a performance task) at the end of a summer internship, you find that only 2 out of 10 were able to produce something of quality (as determined by multiple raters using a rubric). Your results/data show that 20% successfully completed the task (results rated a 1), and your tool was a post-only performance task, the rigor of which is a 2. 1 × 2 = 2, which indicates low impact.

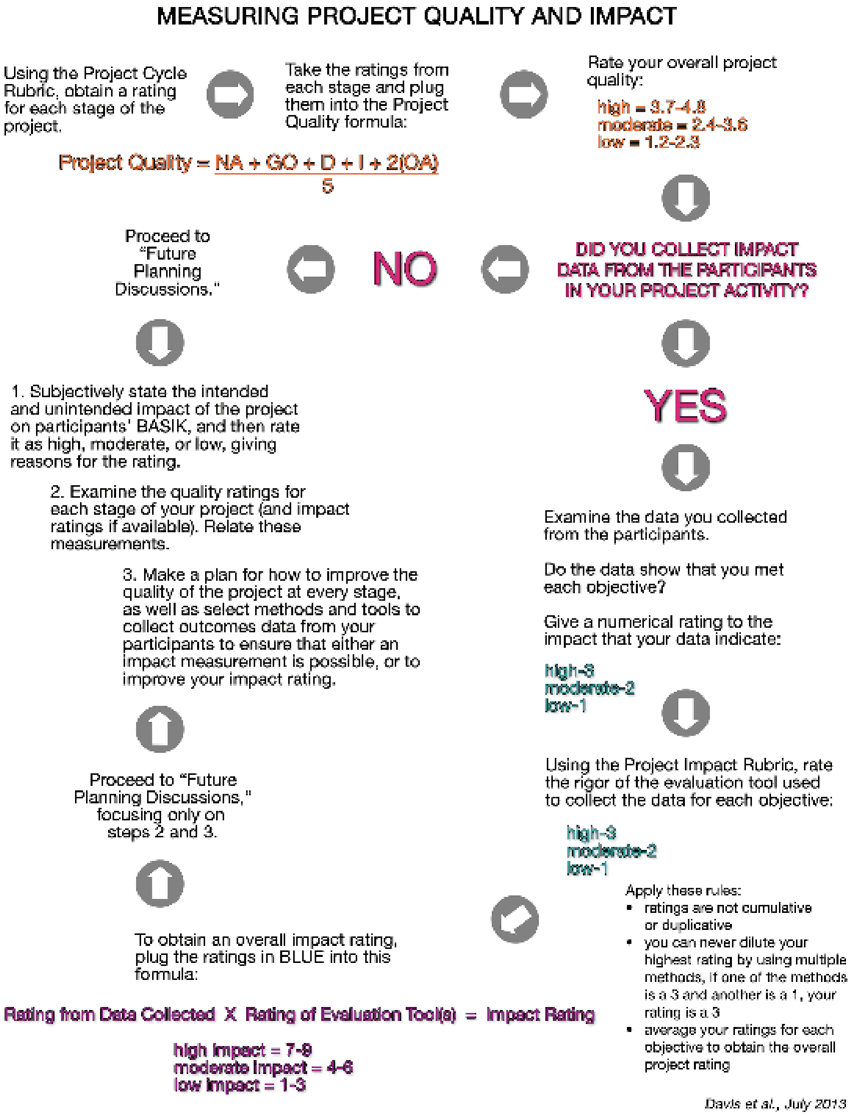
We expanded the working template to include the results and rigor ratings and embedded the formula to calculate both the impact rating for each objective and the overall impact rating for the project (which is presently an average of the ratings for each objective). Multiple objectives can be added to the template. A section entitled "Plans to Improve Impact" was added to collect ideas during the discussion of the ratings.



Measuring Your Project's Impact			**Plans to Improve Impact**
	SCORE	**Description and Explanation**	
Objective 1:			
Rigor of the Evaluation Tool(s): *How confident can you be in the data?*			
Rating from the Data Collected: *What do the data show?*			
Rating from Data Collected X Rating of Evaluation Tool(s) = Impact Rating	0	high impact = 7-9 moderate impact = 4-6 low impact = 1-3	
Objective 2:			
Rigor of the Evaluation Tool(s): *How confident can you be in the data?*			
Rating from the Data Collected: *What do the data show?*			
Rating from Data Collected X Rating of Evaluation Tool(s) = Impact Rating	0	high impact = 7-9 moderate impact = 4-6 low impact = 1-3	
OVERALL IMPACT RATING	0.0	average of overall impact rating for each objective	

Figure 4. Impact Measurement Worksheet Template (H. Davis and D. Scalice, July 2013).

To present a birds-eye view of the process, we developed a flow chart (Fig. 5) to guide a user through the process of assessing their project. As you can see, all pathways lead to Future Planning Discussions in order to ensure the process results in higher quality and increased impact for the project.



MEASURING PROJECT QUALITY AND IMPACT

Using the Project Cycle Rubric, obtain a rating for each stage of the project.

Take the ratings from each stage and plug them into the Project Quality formula:

$$\text{Project Quality} = \frac{NA + GO + D + I + 2(OA)}{5}$$

Rate your overall project quality:

high = 3.7-4.8
moderate = 2.4-3.6
low = 1.2-2.3

DID YOU COLLECT IMPACT DATA FROM THE PARTICIPANTS IN YOUR PROJECT ACTIVITY?

Proceed to "Future Planning Discussions."

1. Subjectively state the intended and unintended impact of the project on participants' BASIK, and then rate it as high, moderate, or low, giving reasons for the rating.

2. Examine the quality ratings for each stage of your project (and impact ratings if available). Relate these measurements.

3. Make a plan for how to improve the quality of the project at every stage, as well as select methods and tools to collect outcomes data from your participants to ensure that either an impact measurement is possible, or to improve your impact rating.

Examine the data you collected from the participants.

Do the data show that you met each objective?

Give a numerical rating to the impact that your data indicate:

high-3
moderate-2
low-1

Using the Project Impact Rubric, rate the rigor of the evaluation tool used to collect the data for each objective:

high-3
moderate-2
low-1

Apply these rules:

- ratings are not cumulative or duplicative
- you can never dilute your highest rating by using multiple methods, if one of the methods is a 3 and another is a 1, your rating is a 3
- average your ratings for each objective to obtain the overall project rating

To obtain an overall impact rating, plug the ratings in BLUE into this formula:

Rating from Data Collected X Rating of Evaluation Tool(s) = Impact Rating

high impact = 7-9
moderate impact = 4-6
low impact = 1-3

Proceed to "Future Planning Discussions," focusing only on steps 2 and 3.

Davis et al., July 2013

Figure 5. Flow Chart to Measure Project Quality and Impact (H. Davis, May, 2013)

4. Conclusion

We learned that project quality and impact can be measured with integrity and rigor. The process we have drafted both honors the informal feedback that well-intentioned and highly skilled EPO professionals naturally collect *and* provides a framework for them to incorporate more rigorous tools through which to measure impact. It acknowledges the best practices already in place and guides project leads to improve their practice at each stage of their project's life cycle.

By involving the NAI EPO Leads in defining and measuring impact, we ensured the process would be grassroots. Working with them revealed the true nature of this process to be less a means to measure how high or low an impact one's project is having, but more a professional development process through which one's project strategically grows and increases impact.

This process provides a standardized, systematic way to assess a project's quality and impact while neither sacrificing autonomy at any stage of a project's life cycle nor submitting to pre-scripted tools that do not measure outcomes that are meaningful to the project.

Involving the NAI EPO leads in the development of the process ensured their support and willingness to submit their projects to it. But we also included non-NAI projects in our pilot of the process and learned that even among those EPO leads with no hand in the development of the process there was full engagement and buy-in. It is our hope that the community will embrace this process and find it robust on the project level to improve practice and increase impact. We also envision and support its adoption on the program/division/directorate level, where we hope it can be effective in demonstrating program-level impact.

By rigorously refining the process, we can all be fluent in a language that discusses not only the level of impact, but also the reasons for it. Using this process, we as a community can maintain our "thousand points of light" landscape of unique, innovative projects while increasing project quality and more rigorously measuring and discussing our impact as a whole.

References

Friedman, A. 2008, "Framework for Evaluating Impacts of Informal Science Education Projects," http://insci.prg/resources/Eval_Framework.pdf