Factors to Consider In Making Evaluation Work for You

Lena Danaia[1]*, Joy Frechtling[2], Sanlyn Buxner[3]

Abstract

Many projects involving robotic telescope education programs do not have published peer-review evaluations (Gomez and Fitzgerald, 2017). Most evaluations in this discipline area tend to be unreliable with questionable methodology and published in non-peer-reviewed conference proceedings or in the grey literature. It is not the case that evaluations are required to be peer-reviewed to be useful to the people running the project, but it is likely that grant funding bodies and other researchers will largely only trust peer-reviewed studies. It is also not the case that evaluations have to be public, they can be done privately for the interest of project personnel and project development alone. Either way, there are many things that need to be considered when planning a project evaluation.

[1] *Faculty of Arts and Education, Charles Sturt University, Bathurst, Australia*
[2] *Westat, Maryland, USA*
[3] *Planetary Science Institute, University of Arizona, USA*
***Corresponding author**: ldanaia@csu.edu.au*

Why evaluate programs?

There are several reasons to evaluate your programs which may drive decisions about what to evaluate and how you plan your evaluation. Many programs include program evaluation because it is required by a funding agency, this may take the form of internal evaluation or external evaluation. You may use evaluation to make real time improvements in your program or collect data so that you can make changes to future versions of the program or future programs. Additionally, you may want to investigate the impact your program has had on participants or institutions, both in the short term or the long term, as well as investigate the impact of the program on learners that participants later engage with. The quality of your evaluation will depend on how rigorously you design and implement data collection and analysis. To be maximally useful an evaluation should include collecting evidence about both successes and places for improvement.

Evaluating your program: What do you want to know?

While some people may think that the only question an evaluation is "Did it work?", there are really a wide range of questions that might be addressed. Roughly speaking, questions can be of two kinds: ones that are for program improvement (formative) and ones that look at outcomes (summative).

Questions that address program *improvement* typically ask whether the program was implemented as planned and if not, why not, and what needs to be done? For a typical educational program this might include looking at:

- Recruitment, selection and retention procedures

- Professional development

- Instructional practices, and

- Assessment practices

Depending on the program you might also examine a range of other components, such as:

- Availability/use of technologies

- Establishment of required partnerships, or

- Engagement of parents and/or the community

Take a program, for example, that provides professional development for middle school teachers to help them engage their students in the use of telescopes to begin to track the path of stars and planets. A formative evaluation might ask:

> What were the recruitment and selection procedures expected to be? To what extent were they implemented as planned? Did selected teachers have these characteristics? If teachers dropped out, is there any relationship between differences in the extent to which teachers had the desired characteristics and whether or not they completed the program?

> What was the overall plan for training the teachers? How much time were they expected to be "in-class" or doing "hands-on" activities? Did they actually engage in the full set of activities? If not, why not? Were there logistical problems that created challenges? Was the instruction at the right level for participants? If problems were found, how might they be addressed?

The goal of this type of evaluation—a formative evaluation—is to look at your program design, see how it worked out in practice, and determine if changes are needed. Formative evaluation is typically done at at the beginning of the project and periodically throughout its early stages. A formative evaluation is used to ensure that you are doing what you said you would do, and doing it well, before you ask whether the program worked.

When a program is new or innovative program developers may adopt a procedure called "design-based implementation research" which may involve several cycles of design, implementation, examination, and redesign, drawing on the skills and perspectives of an interdisciplinary team[1].

Studies that look at *outcomes* examine the extent to which programs result in solving the problems they were designed to address or changing behaviors in expected ways. While with educational programs, gaining new skills and knowledge is typically the outcome first thought of, many evaluations also address changes in:

- Attitudes and beliefs

- Motivation

- Self-confidence or comfort level

- Expectations for the future

For example, in the professional development program described above an outcome of interest might be changes in teachers' comfort levels in using the equipment and making associated calculations, as well as understanding the basics of astronomy.

Further, while many evaluations address changes in individual participants, others look for broader changes such as changes in groups or even changes in systems in which the activity is embedded. Again, turning to the example above, an outcome of interest might be making basic astronomy an expected part of the curriculum in middle school science, as well impacting the students of the teachers attending the specific program.

Deciding What to Evaluate: Developing Your Logic Model

It can be a challenge to determine what questions you should ask about your program. As a first step, it is important to develop a logic model that

---

[1]see, for example, learndbir.org

visually depicts your "theory of change" or "theory of action". Logic Models:

- Organize/make visible your hypotheses about what leads to what

- Identify potential points for asking "how is it working?"—evaluation questions

- Check your logic—are there connections among parts that are missing? Are connections between activities and outcomes clear? Are there activities that don't seem connected to any outcomes?

Knowlton and Phillips (2012) offer the following description of logic models:

> Logic models are a graphic way to organize information and display thinking. They are a visual approach to the implicit maps we all carry in our minds about how the world does or *should* work...Logic models describe the planned action and its expected results.

While there are variations in how logic models are presented, at a minimum they include a set of components—inputs, activities, outputs, outcomes, impacts and context—and connections between the components (Frechtling 2007; Kellogg 2004). Specifically, a logic model contains the following components:

> **Inputs** are the resources brought to a project, typically funding sources or the experiences and knowledge of the individuals and institutions involved.

> **Activities** are the components of the treatments that a project is providing. They are the actions that will be taken to move toward the project's goals and outcomes.

> **Outputs** are the products of activities. Outputs document the implementation of an activity in simple, itemized ways.

They document the size and scope of the activity.

> **Outcomes** are results or changes in behaviors or learning. Outcomes identify the goals and objectives of the project. Outcomes are typically specified in the short and longer term. How to differentiate short- and long-term outcomes is project specific and determined in large part by the complexity of the changes that are sought.

> **Impacts** are broader changes in systems that can be linked to the interventions or activities depicted in the model. Typically such impacts include, but also extend beyond, the specific units that are the focus of study.

> **Contextual factors** are the special characteristics of a grantee setting that need to be considered in understanding how a set of interventions may play out. Context can include both current context and historical context, as previous experiences may color how a proposed change is perceived. Contextual factors help clarify the nature of the setting in which the project is being implemented and can play a role in identifying challenges that need to be addressed. Understanding of context also helps inform where approaches and findings might and might not be generalizable.

Connecting the components and showing relationships between and among them are a series of arrows and lines which show expected interdependences or consequences. The arrows are a critical part of the logic model, showing what is hypothesized to lead to what. The lines are also important and portray connections between potentially isolated pieces. Figure 1 shows a generic logic model.

Logic Model

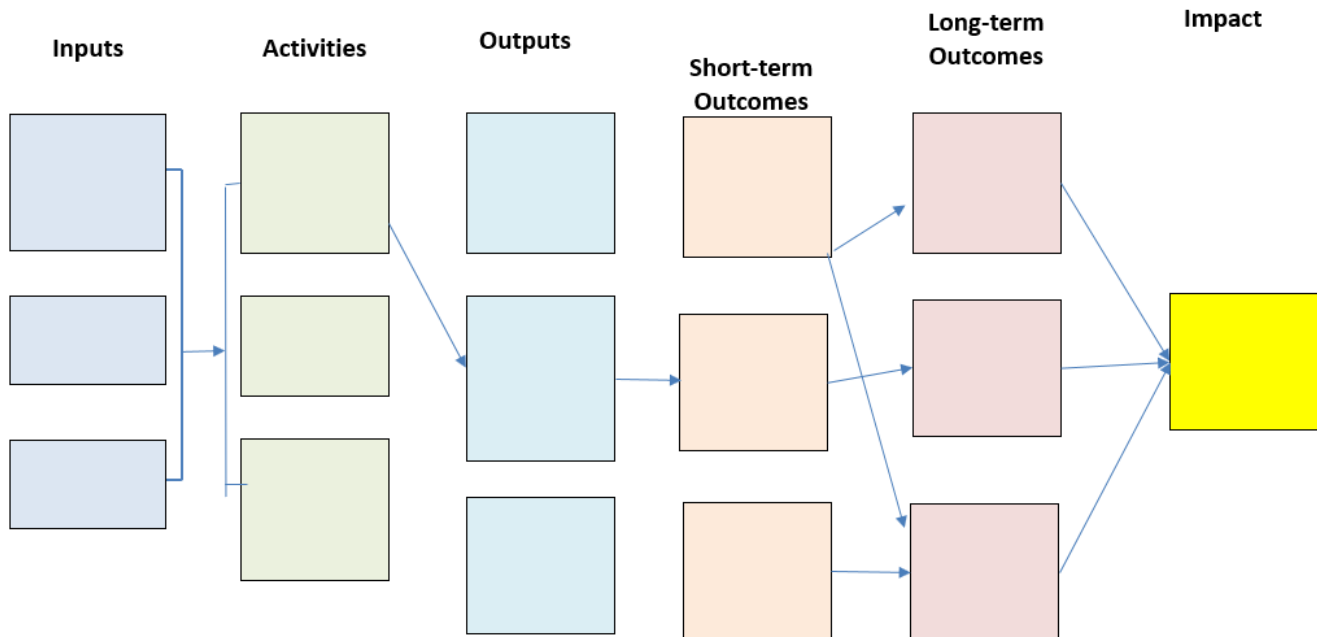| Inputs	Activities	Outputs	Short-term Outcomes	Long-term Outcomes	Impact

Figure 1.

Logic models can be tailored to a range of questions and areas of inquiry. Moving from the model above—the generic model—to one that is more aligned wth the topic of this conference, let's look at a potential logic model for the program described above (Figure 2).

This logic model provides a visual depiction of the activities and outcomes described earlier. It also shows expectations for different groups—teachers, students, and the broader educational system. It shows that outcomes can be considered in the short term (1-2 years) or the long term (greater than 2 years, in this case). It shows connections between components and suggests across and within them. Finally, it calls out contextual factors, suggesting that questions should keep these potential contextual differences in mind in examining both implementation and outcomes.

Developing Your Evaluation Design: What You Need to Consider

It is important to note that data collection can be done in a variety of ways. Often pre post program assessment strategies, involving collecting data at the start and end of an intervention, are a favorite to show changes in knowledge, but they are only one way to determine the outcomes of a program. If you employ pre post program designs you should aim to include comparison groups to strengthen the validity of your research. These comparison groups should consist of individuals with characteristics similar to those of the participants who differ primarily in the fact that they do not receive the services. You can also consider designs that involve a delayed post test or longitudinal designs where you make repeated observations of the same groups over a much longer period of time.

In designing your evaluation, you might consider employing descriptive designs that investigate descriptive questions such as *What is happening?*

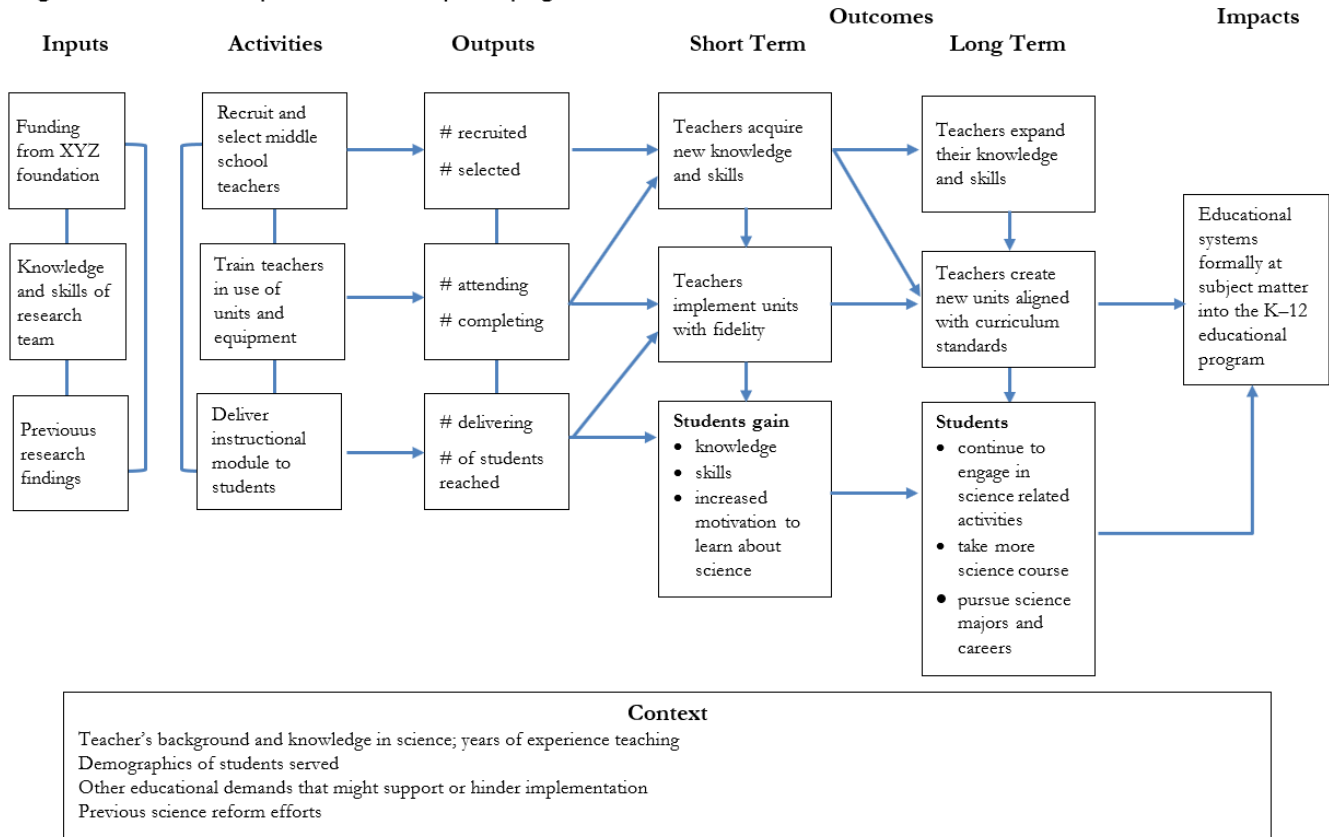Logic Model for a teacher professional development program

Figure 2.

What has happened? Or you may focus on explanatory questions *How or why did something happen?* These designs provide rich descriptive data that will allow you to look closely at program implementation. You can draw on a range of methods for collecting data. Some common sources of data include but are not limited to:

- direct observations (e.g., observations of instructional processes and/or students' engagement in activities- an observation protocol or checklist would be used)

- interviews (e.g., individual or in focus-groups- you may use a list of questions to inform the interview – semi-structured/structured OR interviews may be more open-ended)

- surveys (e.g., rating-scale, open-ended, drawing response)

- archival records (e.g., administrative data that tracks students' grades, GPA, attendance)

- Documents (e.g., project supporting materials, policies, programs)

- Artefacts (e.g., work samples, projects/products, teacher programs).

One caution to keep in mind. Many evaluations suffer from what we call 'confirmation bias' where results are cherry picked to 'confirm' what the project team want to see 'confirmed' i.e., that their project works and is great. Unfortunately, working well is not always the case. Often, rich data come from those evaluations that tell you something is not working. These data can than be used to help improve and refine your project. Evaluation provides a mechanism for you to critically reflect on what is and on what is not working in your

project. Consequently, evaluation can and ideally should, be ongoing.

In constructing your design and carrying out your evaluation, consider working in a cross-discipline research team that might include scientists, educators, and evaluators. This approach may help minimize the confirmation bias that is often inherent in evaluations. It will also ensure a diverse skill set and provide different perspectives when designing instruments and interpreting and integrating the data. It may also provide you with the opportunity to publish in multiple domains.

Overcoming Barriers to Evaluation

Scientists often complain that they do not have the funds, time, or knowledge to conduct evaluation. Additionally, when they do decide to embark on their own evaluation efforts, they may be unaware of issues that pose challenges and risks to conducting ethical and quality evaluations including understanding proper design of data collection instruments, participant fatigue, and government and local regulations related to collecting data from human subjects. Lastly, they may lack the partners to help them plan or execute appropriate evaluation activities. Despite these barriers, there are many resources to support the planning and implementation of rigorous evaluations that will not only benefit programs but also empower you to learn more about your programs, what is working well, and what can be improved.

One important set of issues to to be aware of are ethical considerations, federal laws, institutional rules, and regulations put in place by Institutional Review Boards (IRBs) that regulate data collection from your participants. Depending on who you want to collect data from, the kind of data you want to collect, and the setting you plan to collect data in, there are varying rules about what kind of information is permissible to collect and what you can do with that information. It is your responsibility to know these rules and to be sure that you are in compliance. Working with partners who regularly conduct assessments of programs

can lessen the burden. Additionally, if you are at a University or institution who gets federal research grants, there may be an institutional review board in place to help guide your work. For more information about United States Common Rule for Protecting Human Subjects, see https://www.hhs.gov/ohrp/regulations-and-policy/regulations/finalized-revisions-common-rule/index.html. You can learn more about specific IRB rules and sign up to take training through the Collaborative Institutional Training Initiative (CITI Program) at citiprogram.org/. There is also a series of Astronomy Education Research specific articles: (Brogt et al. 2007a; 2007b; 2008) that highlight several ethical issues to consider and provide advice on dealing with IRB requirements.

To be most effective, all assessment plans need to be planned ahead of program implementation, ideally in parallel with program planning. This includes articulating your program goals, developing your logic model, selecting measurement tools, and identifying how data will be collected and analyzed. Determining the appropriate data collection tools requires finding instruments, or developing ones, that align with your intended goals that are also appropriate to use with the intended audience. Important to selecting the right tools are also investigating and knowing the validity and reliability of the tools. Validity is a measure of how well the tool measures the construct that was intended. Reliability refers to the consistency of the tools in measurement. Whole books are devoted to these topics, the take home message here is that these are important aspects of instruments that need to be taken into account. A somewhat longer introduction to assessment tools and some of the more common science and astronomy education instruments are shown in Buxner et al. (2011).

Many important resources exist to get you started down a productive path evaluating your programs, here we highlight readily available ones. The American Evaluation Association[2] is a leader in setting standards for evaluation, provides a find an

---

[2]http://www.eval.org

evaluator tool, hosts a yearly conference and two journals, and provides networking and training in evaluation. The Center for Advancement of Informal Science Education[3] provides information on projects, education research, and evaluation resources that are applicable to informal STEM education settings. The 2010 User-Friendly Handbook for Project Evaluation[4] (Frechtling et al., 2010) is freely available and can be used as a basic guide for evaluating education projects for those unfamiliar with evaluations and includes information on different types of evaluation, developing evaluation questions and designs, selecting appropriate approaches to evaluation, collecting data, and reporting findings. In addition, many universities have central offices of assessment and learning as well as Colleges of Education or STEM learning centers to collaborate with.

Conclusion

Evaluation is a critical component of program planning and implementation. Evaluation can take many different forms and will be dependent on what you want to know about your program and the resources available. Evaluation can have many practical benefits for your program. It provides a means for you to critically examine your program and inform your program design. Sharing your methods as well as the results of your evaluation (both positive and negative) will contribute to the overall community of practitioners who are planning their own programs and program evaluations.

References

Brogt, E., Dokter, E., and Antonellis, J. (2007a). Regulations and Ethical Considerations for Astronomy Education Research. *Astronomy Education Review*, 6(1).

Brogt, E., Dokter, E., Antonellis, J., and Buxner, S. (2007b). Regulations and Ethical Considerations for Astronomy Education Research II: Resources and Worked Examples. *Astronomy Education Review*, 6(2).

Brogt, E., Foster, T., Dokter, E., Buxner, S., and Antonellis, J. (2008). Regulations and Ethical Considerations for Astronomy Education Research III: A Suggested Code of Ethics. *Astronomy Education Review*, 7(2):57–65.

Buxner, S., Wenger, M., and Dokter, E. (2011). Exploring Assessment Tools for Research and Evaluation in Astronomy Education and Outreach. In *Earth and Space Science: Making Connections in Education and Public Outreach*, volume 443, page 125.

Frechtling, J., Mark, M. M., Rog, D. J., Thomas, V., Frierson, H., Hood, S., Hughes, G., and Johnson, E. (2010). The 2010 User-Friendly Handbook for Project Evaluation. National Science Foundation.

Frechtling, J. A. (2007). *Logic modeling methods in program evaluation*, volume 5. Jossey-Bass San Francisco.

Gomez, E. L. and Fitzgerald, M. T. (2017). Robotic telescopes in education. *Astronomical Review*, 13(1):28–68.

Kellogg, W. (2004). Logic model development guide. *Michigan: WK Kellogg Foundation*.

Knowlton, L. W. and Phillips, C. C. (2012). *The logic model guidebook: Better strategies for great results*. Sage.

---

[3]informalscience.org

[4]Handbook available at http://www.evalu-ate.org/resources/doc-2010-nsfhandbook/