Framework for Evaluating Impacts of Informal Science Education Projects

Report from a National Science Foundation Workshop

The National Science Foundation

The Directorate for Education and Human Resources

The Division of Research on Learning in Formal and Informal Settings (DRL)

Framework for Evaluating Impacts of Informal Science Education Projects

Report from a National Science Foundation Workshop

Editor:

Alan J. Friedman

Authors:

Sue Allen
Patricia B. Campbell
Lynn D. Dierking
Barbara N. Flagg
Alan J. Friedman
Cecilia Garibay
Randi Korn
Gary Silverstein
David A. Ucko

To cite: Friedman, A. (Ed.). (March 12, 2008). Framework for Evaluating Impacts of Informal Science Education Projects [On-line]. (Available at: http://insci.org/resources/Eval_Framework.pdf)

Prepared under Contract Number GS-10F-0482P, NSF Order Number DACS06D1421, Evaluation Activities Related to the Academic Competitiveness Council's Examination of STEM Education Programs.

12 March 2008

The National Science Foundation

The Directorate for Education and Human Resources

The Division of Research on Learning in Formal and Informal Settings (DRL)

TABLE OF CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

ACKNOWLEDGMENTS

We are extremely grateful to the authors who shared their considerable expertise in writing the sections of this *Framework for Evaluating Impacts of Informal Science Education*: Sue Allen (Exploratorium), Patricia B. Campbell (Campbell-Kibler Associates), Lynn D. Dierking (Oregon State University), Barbara N. Flagg (Multimedia Research), Cecilia Garibay (Garibay Group), Randi Korn (Randi Korn & Associates, Inc.), and Gary Silverstein (Westat). Very special thanks go to Alan J. Friedman for his dedication to organizing and editing this publication, along with facilitating the NSF workshop on which it is based.

Thanks also to the external reviewers who provided many excellent suggestions for improvement and to those at NSF, especially Alphonse DeSena and Bernice Anderson, who made major contributions.

David A. Ucko
Deputy Division Director
Division of Research on Learning in Formal and Informal Settings

PART I GENERAL INFORMATION

Part I explains the origins of this book in the National Science Foundation's work to advance the informal science education field as a whole, followed by advice on how to use the book. Then two chapters provide overviews of impact evaluation and a look at some of the common issues, concerns, and opportunities in evaluation practice.

CHAPTER 1 INTRODUCTION TO EVALUATING IMPACTS OF NSF INFORMAL SCIENCE EDUCATION PROJECTS

David A. Ucko

The Informal Science Education (ISE) program at the National Science Foundation (NSF) invests in projects designed to increase interest in, engagement with, and understanding of science, technology, engineering, and mathematics (STEM) by individuals of all ages and backgrounds through self-directed learning experiences. In addition to these public audience impacts, projects must demonstrate how they seek to advance the knowledge and practice of informal science education. The ISE program also supports projects that directly target professional audiences to improve the infrastructure and capacity of the field. All projects are required to build on informal learning research, practice, and prior work and then add to this knowledge base through evaluation.

The ISE program has played a major role in promoting the use of project evaluation. As a result, attitudes and practices regarding evaluation have changed dramatically over the past several decades. Today, ISE professionals generally recognize the importance of front-end, formative, remedial, and summative evaluation in guiding projects, improving them, and ascertaining whether they achieve their intended outcomes.

Several years ago, the ISE program solicitation underwent a revision that furthered this trend by placing greater emphasis on identification of project impacts. Proposals now must start with intended impacts on both target audiences and the field, then present the innovative deliverables and strategies designed to achieve those impacts, and lastly, identify the project team and collaborators who have the expertise necessary to develop them. A summative evaluation is required to assess the target audience impacts, going beyond simple numerical outcomes such as numbers reached.

Because they offer potentially valuable information about the impacts of a project along with related findings, summative evaluations provide a way to advance knowledge and practice, a primary goal of the ISE program. Assuming these evaluations are accessible, they can enable others to build upon the results of prior work and further the state-of-the-art. They provide a mechanism by which the field overall can enhance its efforts to provide the most effective informal learning programs and resources.

Summative evaluations included in ISE proposals vary widely in what is evaluated. That observation should not be surprising given the very different forms of deliverable, which for public audiences include: exhibitions; community and youth programs; television and radio series; giant-screen films; and technology-based and cyber-enabled learning projects. In addition, there are relatively few standardized instruments and approaches. The resulting diversity in summative evaluations makes it difficult to conduct cross-project portfolio reviews.

To help NSF better understand the impacts of its investments in informal science education, the ISE program has developed an online Project Monitoring System that will facilitate the capture, synthesis, and analysis of project outcomes and impacts. Unlike FastLane, an NSF system that stores information in largely narrative form that cannot be easily sorted or aggregated, the online survey creates a relational database with searchable fields containing ISE-specific data. It enhances the program's capacity to monitor the progress of individual projects, describes its portfolio of awards, examines trends within and across project types, and more readily responds to external and internal inquiries. Equally important, the project database enables program staff to generate *ad hoc* reports on topics of interest to a wide range of stakeholders. Examples are the number and characteristics of projects that have been funded in a particular Congressional District or those that serve a specific target audience, the types of challenges encountered by a specific project type (and steps taken to overcome them), and the range of audience impacts delineated by a specific project type.

While this online system was being developed, the Deficit Reduction Act of 2005 created the Academic Competitiveness Council (ACC), which initiated a cross-agency review of all federally funded STEM education programs. NSF took the lead, with ISE support, on the Informal Education and Outreach working group, which included representatives from Defense, NASA, NIH, NOAA, and other agencies. After formulating a set of definitions and common goals, the group discussed ways to evaluate more effectively the impact of their informal STEM education and outreach programs. Emphasis was placed by the ACC on experimental and quasi-experimental designs as the most rigorous means to determine whether interventions were effective. The ACC process also recognized that the evaluation design must be appropriate to the project, and that informal learning is most challenging to assess. Here is what the ACC report concluded about informal science education programs:

First, the variety in types of programs is expansive. Informal education and outreach activities can take place in schools, museums, the community, the media, and various other locations where people gather information and experience the world. Almost all ACC agencies had some type of program that was designed to generate awareness and engage the public in the agency's work. Further, the types of activities varied considerably across programs.

Second, the nature of these programs makes it difficult to conduct rigorous evaluation because, among other reasons: (1) the audience for these programs is diffuse and difficult to identify; (2) the multiple factors affecting and affected by these activities cannot be isolated for assessment; and (3) the modest scale of these efforts does not warrant a costly assessment approach. There are examples of pre- and post- quasi-experimental evaluations of these programs, but it is extremely challenging to carry out rigorous studies to identify causality in these programs.

Third, despite all of these complexities many programs share the same or similar goals. Though the programs are varied, the Informal Education and Outreach group agreed on the importance of better interagency coordination and information sharing. Sharing best practices across agencies could offer significant benefits as these programs often bear a closer relationship to outreach and informal education programs in other agencies than they do to K–12 or postsecondary education programs in the same agency (U.S. Department of Education. [DoEd], 2007, p. 26.).

The ACC report recognizes the difficulty of conducting the most rigorous forms of summative evaluation in the informal learning realm. At the same time, it recommends "federal STEM education programs designed to improve STEM education outcomes should not increase unless a plan for rigorous, independent evaluation is in place, appropriate to the types of activities funded" (DoEd, 2007, p. 26). The evaluation requirements in the NSF ISE solicitation, accompanied by the online survey, will help the program to address this recommendation by gathering basic data needed to evaluate the impact of projects in the portfolio.

The Informal Education and Outreach working group adopted a framework consistent with the online survey. The following table identifies those broad categories of potential project impact. They can be applied both to projects that target *public audiences* by means of an informal STEM education or outreach deliverable, as well as, projects that target *professional audiences*, those who work in the field or directly influence that work:

Table 1-1 The Informal Education and Outreach Framework

Impact Category	**Public Audiences**	**Professional Audiences**
Awareness, knowledge or understanding (of)	STEM concepts, processes, or careers	Informal STEM education/ outreach research or practice.
Engagement or interest (in)	STEM concepts, processes, or careers	Advancing informal STEM education/outreach field
Attitude (towards)	STEM-related topic or capabilities	Informal STEM education/ outreach research or practice
Behavior (related to)	STEM concepts, processes, or careers	Informal STEM education/ outreach research or practice
Skills (based on)	STEM concepts, processes, or careers	Informal STEM education/ outreach research or practice
Other	Project specific	Project specific

These impact categories make it possible to communicate a range of project impacts to stakeholders. They enable the program to disaggregate, sort, and analyze the wealth of data collected from individual projects in its portfolio with an emphasis on outcomes, rather than descriptive categories (such as project type or target audience). Identification of these categories was based on analysis of project impacts from a comprehensive review carried out on a representative sample of ISE proposals, final reports, and summative evaluations. Addition of an "Other" category to this table in the examples in this book allows for impacts unique to a project that may not otherwise be captured, as well as, impacts in areas like "creativity" which may be defined and categorized in multiple ways.

In planning ISE projects and their assessment, proposers will be asked to use these categories to identify each of their intended audience impacts. It is up to the proposers to decide which categories fit their goals for the project, since it is highly unlikely that all or even most of these impacts will apply to any particular project. Any of the categories may be applied to a specified target audience, such as an underserved group. Proposers also will be asked to indicate the means by which they will seek to demonstrate each impact, the form of evidence that will be used to assess whether the impact was attained, and, in the final submission, the extent to which the impact was achieved. Of course, the evaluation design and methodology should be based on the nature of the project and the questions being asked.

It is important to recognize that neither the evidence associated with these impact categories nor the online survey is intended to capture fully all the outcomes of a project. Rather, they were designed to allow a program such as ISE to collect project-level impacts in a systematic way. This collection is important not only to assist in reviewing and analyzing a portfolio, but also to demonstrate the impact of the ISE program investment on the field and the impact of informal STEM learning more broadly.

In addition, each project's evaluation design may collect a variety of other important information in the form of naturalistic evaluation studies, case studies, lessons learned, and innovative contributions to the field, which will continue to be documented as well. Some will be captured in narrative form within the survey; other information will be provided through FastLane as part of annual and final project reports, which may include copies of front-end, formative, and summative evaluations (the summative evaluations are required to be posted at the web site www.informalscience.org.). So the evidence associated with these impact categories and the online survey are designed to record only a portion of each project's evaluation information, which will be complementary to the rich range of information captured by other means. Overall, the intent is to create a flexible project planning and reporting framework that does not constrain project design or evaluation any more than is necessary to collect this valuable data.

We recognize that applying the impact categories to projects will be challenging. As the ACC report recognizes, informal learning is individualized, complex and multifaceted. Impacts depend on personal, physical, social, and cultural contexts and may not become evident until sometime after the experience. In addition to these types of factors, audiences may be highly heterogeneous. For these reasons, impacts are typically more complicated to assess than in the classroom, especially since the most important learning outcomes may be non-cognitive and more difficult to measure. To assist proposal developers and evaluators, the ISE program has been supporting several evaluation-related initiatives through recent awards. The informalscience.org web site (University of Pittsburgh Center for Learning in Out-of-School Environments), which contains evaluation resources in addition to project evaluations, is being enhanced. A grant to the Visitor Studies Association is making possible mid-career professional development for evaluators. A forthcoming National Academies synthesis study on research on learning in informal settings will establish a foundation that will inform assessment. In addition, the new Center for the Advancement of Informal Science Education (formerly known as the ISE Resource Center) will provide another mechanism for sharing information about project findings, along with other resources and activities designed to advance the field and foster a community of practice (see www.insci.org).

The March 12-13, 2007 workshop at NSF on informal science education evaluation brought together a distinguished group of experts to discuss how impact categories might be best applied to various types of informal learning projects. This publication is an outcome of that meeting. The authors have strived to make the sections as helpful as possible given the primary focus of this workshop on project impacts. It should be viewed as part of an ongoing process to improve the ways in which evaluation can most benefit ISE projects, NSF, and the field. The publication is intended to help those developing projects to think about and better articulate the intended public or professional audience impacts. Since the design of deliverables and strategies should be based on achieving these desired outcomes, this publication should also encourage project leaders to work more closely with evaluators at very early stages in the conceptual development of their projects. More broadly, this effort should advance understanding of summative evaluation by the field. Although this guide was written for the ISE program at NSF, we hope that many aspects will be relevant to other agencies, foundations, or organizations that wish to evaluate aspects of informal learning. Because this endeavor is, and likely will remain, a work in progress, the NSF staff welcomes your feedback.

REFERENCES

U.S. Department of Education. (2007). *Report of the Academic Competitiveness Council.* Washington, D.C. Available for download at:http://www.ed.gov/about/inits/ed/competitiveness/acc-mathscience/index.html

CHAPTER 2 USER'S GUIDE TO THIS BOOK

Alan J. Friedman

This book introduces a framework for evaluating the impacts of informal science education projects. The authors have extensive experience evaluating projects across the NSF portfolio. We are writing for Principal Investigators and Project Directors, who may not do the evaluation themselves but certainly need to know if their initial designs are reasonable, how much they will cost, and who can perform them. Evaluators, both in-house and external, will find advice about what techniques are currently in use for different types of informal science education, and references to case studies and tools. And proposal developers will find a wealth of information on good practices and standards of the field in order to write competitive proposals.

But we also hope this book will be useful in general for people looking to find appropriate evaluation frameworks for their STEM education projects, whether supported by NSF or not. If you are developing a project designed to learn more about STEM education or improve practice in STEM education, you will want (and review panels will want) to have a plan to evaluate the impact of your work. Your project evaluation will not only tell you what's happened as a result of your work, but it will help others who find your ideas interesting and are considering applying what you did for their own work.

THE FOCUS OF THIS BOOK

As Chapter 1 of this book explained, NSF has established a set of "impact categories" and an on-line survey tool to help gather information about the impacts of the projects it supports. The kind of evaluation we will be describing, and that will generate the information you will need to complete those surveys, is called *summative evaluation* (see the Glossary in Appendix A for definitions of terms used in this book). Selecting appropriate methods for performing summative evaluation is what this book is all about.

Good summative evaluations are not universal, and funders such as the NSF need summative results to help them assess their portfolios and justify their enterprises in the never-ending challenge of maximizing results for the taxpayers' investments. So while you are evaluating the impacts of your individual *project*, NSF will be using what you report to learn about the impacts of its entire *program*.

The evaluations you will be providing can also help to advance the entire field of informal science education. Several decades of support for informal science education by NSF and others should have advanced the field to a point where we are actually discovering more generalizable insights into learning. To capture that advance, a consistent knowledge base of summative findings is needed. The framework described in this book is intended to help generate those findings. That goal of developing more generalizable results takes us beyond evaluation, which is the focus of this book, and into the realm of research, where the deliverables are indeed findings about learning, rather than specific evaluation studies of any particular learning strategy.

Improving summative evaluation will provide fertile ground for doing research, and some individual evaluations may suggest insights which can be tested in a research agenda. But again, our focus here will be on individual project summative evaluations, the basic findings from which broader discoveries can emerge.

APPROPRIATE EVALUATION PLANS

As you will see in this book, there are many possible ways to do a summative evaluation. The authors will look at evaluation in different NSF ISE program areas including exhibitions, mass media, and community group projects. What constitutes an appropriate evaluation strategy depends, of course, on what your particular project involves. You want to use the most rigorous methodology possible, and you want to be sure that you squeeze out the most information you can about the impact of your project.

Sometimes the most appropriate evaluation will involve a methodology known as randomized controlled trials (RCT), which is often used in medicine and pharmacology to evaluate the impact of a new procedure or a new drug. In testing your learning innovation with RCT methodology, you would create at least two groups of participants and audiences, with individuals randomly selected for the groups. One will be a control group, participants of which do not use the particular learning innovation you have developed. Other individuals will be in the experimental group, which do use that innovation. You would then collect quantitative data on the impacts - such as knowledge, attitude, and skills - for both groups and compare the results.

That's an excellent methodology for some projects, where you are seeking to establish if there is a causal relationship between a specific innovation (treatment) and a set of specific effects or side effects of that treatment. This methodology may be expensive and time consuming, but it can provide hard, numerical evidence for the causal effects of one or more variables. Nevertheless, RCTs may be neither practical nor the most appropriate for the task at hand. Suppose your project is designed to learn how teenagers are using the Internet today to pursue science related hobbies, like building model airplanes. There is no "treatment" being tested, no causal links to be established, so the RCT methodology, powerful though it is in other situations, would be inappropriate for your project.

Many years ago I worked on a team which created a new science museum in a capital city. It occurred to me while we were designing the museum that it would be great to evaluate the impact of the new institution on the science knowledge of the population of that city. But what would the appropriate methodology for such a study be? Was a randomized controlled trial possible? The first challenge would be to determine the experimental and control groups. Perhaps we could compare the science knowledge of the population in a city which did not open new science museums with the population's science knowledge in this city which did. But different cities are very different in economic, cultural, and many other potentially important characteristics, and the people who live in those cities are certainly not randomized across the universe of cities. Also, within any one city and any one period of time, there are a vast number of other variables which might affect science knowledge, such as what was happening in the schools, on television, and in other informal science organizations serving the population of the city. The only way to perform an RCT would be a fantasy: we'd have to select many admittedly

non-identical cities and build science museums in some of those cities, selected at random. It would take a large number of cities, and a large number of new museums, to begin to smooth out the effects of the uncontrolled variables which could be affecting science knowledge in the populations of each city. I had to admit that as much as I liked RCT methodology for evaluating the impact of magnetic fields on properties of an antiferromagnetic crystal (my dissertation project), it was not appropriate for learning about the impacts of a new science museum in this case. Fortunately, there are many other techniques discussed throughout this book, which are appropriate.

The challenge is then to select *the most rigorous design appropriate for the work at hand.* Every project needs to find an evaluation design which gives the most reliable analysis practical while making the most efficient use of finite time and money. That's what this book is designed to help you do.

Every choice of evaluation methodology is going to be a compromise, but it should be a compromise that maximizes the final impact of a project on learning and on the capabilities of the field as a whole. So this book is filled with examples of appropriate evaluation strategies for a range of different projects in informal science education.

HOW TO READ THIS BOOK

Few readers will want to study this publication cover to cover. Part I, including the introduction in Chapter 1, this chapter, Chapter 3 on evidence and categories of impacts, and Chapter 4 on common issues evaluation planners will encounter, will probably be valuable for everyone. Part II, the largest section of the book, assumes readers will have read the general information presented in Part I. Each chapter in Part II then deals with applying the general ideas of summative evaluation for the various program areas of NSF's ISE program, including exhibitions, mass media, and youth and community projects. The authors have assumed that most readers will probably want to examine only those chapters in Part II that are relevant to their specific projects.

But you might enjoy reading chapters in Part II about evaluating impacts in program areas that are different from your own. The examples given are interesting, and you might find something applicable to your own area. Because there are so many possible strategies, no one chapter tries to cover all possibilities. Each author selected techniques to describe those which are most often used in that chapter's program area. So you can find more techniques, which might still be useful to you, by reading other chapters in addition to the ones that most directly relate to your program area of ISE.

Should you decide to read more than one chapter in Part II, you will find some repetition. That's because we did want each chapter to be able to stand on its own. Many basic ideas are covered in each chapter of Part II, with examples and terminology specifically chosen for a particular program area.

WHAT THIS BOOK IS NOT

There are many things this publication is not, of course. It is not a step-by-step instruction manual to performing evaluation, although there will be pointers to such manuals, such as *The 2002 User-Friendly Handbook for Project Evaluation* (NSF 02-057).

Also, this publication is not about evaluation in general, but will deal only with summative evaluation. That is *not* intended to diminish the importance and usefulness of other forms of evaluation. It has been widely demonstrated that they can lead to better end results and save time in the process. We encourage all readers of this book to use one or more of these other forms of evaluation, and the bibliography at the end of this book includes guides to these powerful techniques:

- **Front-end evaluation.** This means asking questions to find out what your audience members already know, what they don't know, what they are interested in and what they are not interested in. Knowing these before you develop a project can save you from having to create unnecessary experiences, or from neglecting to treat subjects which are both interesting and needed.

- **Formative evaluation.** Decades of uses of formative evaluation, which is iterative testing of learning strategies to improve them as they are developed, have proven invaluable over and over. No matter how well we imagine a strategy will work, it takes exposure to real audience members to discover just what actually works, and for whom. With formative evaluation, we can improve our strategies before they are set in concrete and become too expensive to change.

- **Remedial evaluation.** Remedial evaluation requires discipline on our part: saving enough money and time so that we can look at our "finished" products, investigate how audience members experience them, and make hopefully minor adjustments to improve (remediate) the end results. Remedial evaluation is concentrated near the end of a project, like summative evaluation, and may use the same tools. But the purpose of remedial evaluation is different: it is performed to make one last round of improvements to the project's deliverables, rather than to evaluate the impact of the project. Remedial evaluation can take place before, during, or after summative evaluation, and may even use the same data.

Each evaluation strategy can be invaluable for getting the most out of an education endeavor, and this publication encourages your attention to these other forms of evaluation.

WHO WROTE THIS BOOK?

The authors and editor of this book are all experienced in informal science education and its evaluation (see Appendix C for author profiles). They have been leaders and senior staff of hundreds of projects in informal learning, and are reporting here what they have found out through their own work and what they have learned from others in the field. This team has been assembled by the ISE program of the NSF, but the findings and opinions expressed here are those of the writing team, and do not necessarily reflect official positions of the Foundation. We

hope, however, that our suggestions and information will be useful to everyone working with the NSF, will help the Foundation track the contributions its portfolio make to learning, and will be useful to anyone interested in assessing the impacts of their education endeavors.

CHAPTER 3 EVIDENCE AND CATEGORIES OF ISE IMPACTS

Lynn D. Dierking

If you don't know where you're going, any road will get you there.

Lewis Carroll

You have an idea that you think will make a great ISE project. As a potential PI, your most important consideration is how to develop and write a competitive proposal that will enable you to actualize this great idea. Now the challenge that you and your project team face is figuring out how to describe succinctly, yet concretely, the impacts intended for your project. In order to be funded, you will need to describe the intended impacts of the project for the proposed audience(s), demonstrate how your project activities will achieve these impacts, and describe a plan for how you will evaluate the intended impacts. The purpose of this chapter is to help you and your project team develop a set of appropriate, measurable and valid project impacts; and a plan for evaluation that will demonstrate whether, and if so how, you accomplished your project goals and objectives.

In the late 1980s, Mary Ellen Munley - a leader in museum education - wrote a paper in which she argued persuasively about asking the right question(s), suggesting that project developers and evaluators could benefit greatly from the insights of Lewis Carroll (Munley, 1986, 1987). Although tongue in cheek, Munley's point is well taken—evaluation, in and of itself, is merely a process and a set of tools to document the outcomes and accomplishments of any effort. An even more essential part of project development - from which any evaluation plan will flow - is the development of clear goals and objectives for what one plans to do, for whom, and why.

Before proceeding though, it is important to emphasize that evaluation is a process and tool for planning. It is not what happens 'behind the curtain' but is a deliberate system of monitoring and tracking that spawns from, and feeds back into, the planning process in a cyclical and iterative way. It is conducted to ensure that programs are on track and successful and is key to effective practice. In other words, evaluation is the flipside of good planning. This means before setting out and even beginning to design a project, let alone an evaluation plan, it is critical to be able to clearly describe what one is actually attempting to accomplish by using a *backward research design* approach (Wiggins and McTighe, 2001). Some of the questions a project team should be able to answer at the outset of initiating a project using a backward research design approach include:

(1) What audience impacts will this project facilitate?

(2) What approach/type of project will best enable us to accomplish these goals and why do we feel that this is the best approach to take?

(3) How will we know whether the activities of the project accomplished these intended goals and objectives and with what evidence will we support the assertion that they did?

(4) How will we ensure that unanticipated outcomes are also documented?

These are the essential questions of the field; some you and your team should be able to answer yourselves, and others you will want to involve an evaluator in helping you answer early on while developing and writing the proposal. Finding an evaluator from the outset whose philosophy and approach matches your ideas and the type of project you are thinking about undertaking, particularly to help you plan and design the summative evaluation, is critical to writing a successful proposal. Although you and your team will want to be involved in thinking about all phases of evaluation, it is important the summative evaluation be conducted by an independent and unbiased evaluator to ensure the integrity of the process. This is money well spent since in addition to research skills, good evaluators bring an understanding of the planning process as cyclical and iterative – a set of deliberate steps which outline and track progress toward a goal.

However, evaluation is not just for preparing good proposals. It is also an integral part of running good projects. During crucial stages of program development, evaluation documents or measures achievements or outcomes against intended goals and objectives (while also being open to unanticipated outcomes as well). All forms of evaluation play an important role in planning, enabling "reflective practice" and facilitating project team/institutional learning. Since evaluation is a process that contributes to decision-making at key points of project development and implementation, and evaluation can be used to ensure success throughout the process of project development, it is important to include a comprehensive plan for evaluation. At a minimum, that includes front-end formative and summative evaluation, and ideally also includes remedial efforts to tweak and improve projects as they are initially implemented. Utilizing all forms of evaluation helps to ensure the progress and success of your efforts.

WHAT IMPACTS DO YOUR PROJECT TEAM WANT TO FACILITATE?

As we suggested, the challenge that you face as a PI and project team is figuring out how to describe succinctly and concretely the impacts intended for the proposed project. This is challenging for many reasons. By their very nature, informal science education projects and experiences are varied and designed to serve different audiences. It is a field in which multiple outcomes are the norm, and where learning is often the result of combined, interwoven and overlapping experiences (informal, formal and everyday). Thus the focus needs to be about understanding how the experience of participating in/or engaging with your project has *contributed* to fostering, reinforcing and sustaining science interest and understanding. This is a tall order indeed.

As described in Chapter 1, with the new online project monitoring system, the ISE program is going to be able to better document and track the project outcomes of individual projects. In addition, the outcomes you document will be used by the program overall to assess outcomes across the various categories of projects that ISE funds (exhibitions, youth and community

programs, media and cyber-enabled learning projects, projects for either the public or professional audiences or sometimes both). A critical component of this monitoring system is the set of impact categories provided in Chapter 1, and defined in Table 3-1, below. As emphasized, your intended project impacts should fall within *some*, but *not all* of these categories (each project should target a few intended outcomes that fall within at least one of these categories of impacts).

Table 3-1 Impact categories as they relate to public audiences:

Impact category	Generic definition
Awareness, knowledge or understanding	Measurable demonstration of assessment of, change in, or exercise of awareness, knowledge, understanding of a particular scientific topic, concept, phenomena, theory, or careers central to the project
Engagement or interest	Measurable demonstration of assessment of, change in, or exercise of engagement/interest in a particular scientific topic, concept, phenomena, theory, or careers central to the project
Attitude	Measurable demonstration of assessment of, change in, or exercise of attitude toward a particular scientific topic, concept, phenomena, theory, or careers central to the project or one's capabilities relative to these areas. Although similar to awareness/interest/engagement, attitudes refer to changes in relatively stable, more intractable constructs such as empathy for animals and their habitats, appreciation for the role of scientists in society or attitudes toward stem cell research
Behavior	Measurable demonstration of assessment of, change in, or exercise of behavior related to a STEM topic. These types of impacts are particularly relevant to projects that are environmental in nature or have some kind of a health science focus since action is a desired outcome.
Skills	Measurable demonstration of the development and/or reinforcement of skills, either entirely new ones or the reinforcement, even practice, of developing skills. These tend to be procedural aspects of knowing, as opposed to the more declarative aspects of knowledge impacts. Although they can sometimes manifest as engagement, typically observed skills include a level of depth and skill such as engaging in scientific inquiry skills (observing, classifying, exploring, questioning, predicting, or experimenting), as well as developing/practicing very specific skills related to the use of scientific instruments and devices (e.g. using microscopes or telescopes successfully).
Other	Project specific.

What follows is a brief and general description of the framework of impacts. These impact categories are not arbitrary but theoretically grounded in the informal science education professional literature specifically, and educational research more generally. They represent valid and common impacts observed as a result of informal science education activities. These categories though are not as black and white as they appear. There are often relationships and intersections between different types of impacts you may want to emphasize in your project, for example, the relationship between attitudes and knowledge.

Table 3-2 provides a work sheet template that you and your team can use to think through the potential impacts of your project, the indicators you will use to assess this impact, and what your criteria for evidence will be. In the next chapter of this guide this work sheet will be used to develop impacts, indicators and evidence for a hypothetical project. In subsequent chapters, you will see how each of these categories can be used to specifically define your intended outcomes depending upon whether the proposed project is an exhibition, a youth and community program, or a media project; focused on a public or professional audience, or both.

1) Knowledge: This category of impact emphasizes what a participant, be s/he a youth in a community program, or a visitor to a museum, or a web site user, consciously knows. Impacts in this category include knowledge, awareness, or understanding that can be stated by participants in their own words, whether that is during, immediately after, or long after, the experience. The content of the impact depends upon the project topic and can include STEM-related concepts, principles, phenomena, or theories, the history or philosophy of science, careers, or science as a process. Evidence for this impact includes changes in participants' knowledge (directly assessed or self-reported), as well as observed cognitive activities such as reinforcing prior knowledge, making inferences, or building an experiential basis for future learning (though this is more difficult to assess). It also includes memory of an experience over time, especially aspects of the experience that relate to STEM concepts, processes, or activities. Participants' reflections and monitoring of their own learning also falls into this category.

2) Engagement: Impacts in this category capture the excitement and involvement of participants in a topic, area, or aspect of STEM. The category includes participation and engagement, prerequisites for other types of learning which are also linked to interest. It could be supported by evidence that a project deliverable has evoked short-term interest, or has strengthened prior longer-term interest, in a topic or area of STEM. This impact is often a focus of projects that aim to engage historically under-represented participants in STEM.

3) Attitude: Impacts in this category encompass changes in long-term perspectives toward a STEM-related topic, a group of people, species or ecosystem, activities, theories or careers. An ISE project may strive to influence attitudes where none existed before, or may change attitudes. Indicators for the "attitude" impact can be less reliable than indicators of knowledge or engagement, because they tend to rely on self-report by participants, who may not always be fully aware or entirely honest about their attitudes. For this reason it is desirable to assess for attitude in multiple contexts and over a range of time frames if possible.

Table 3-2. Work Sheet for Developing Intended Impacts, Indicators & Evidence

ISE Category of Impact	Potential indicators	Evidence that impact was attained
Awareness, knowledge or understanding of STEM concepts, processes or careers		
Engagement or interest in STEM concepts, processes, or careers		
Attitude towards STEM-related topics or capabilities		
Behavior resulting from experience		
Skills based on experience		
Other (describe)		

4) Behavior: Some ISE projects propose to change participants' long-term behavior after the experience, be it an exhibition, a youth program, or a giant screen film. This category of impact is particularly targeted in projects that are environmental in nature or have some connection to the health sciences since subsequent action is a desired outcome. Evidence of behavior change might include participants' self-reported intentions to change their behavior, and longitudinal follow-ups with them (or others) to determine whether such behavior change has occurred. Like attitude, evidence for behavior change can be influenced by people's bias to please (their tendency to say what they think the researcher wants to hear), so follow-up assessments of actual behavior change are particularly important. Clearly there is a potentially important relationship between the categories of attitude and behavior change. For projects where this relationship is a

central focus, PIs and evaluators may wish to refer to background theories such as: the Theory of Reasoned Action, Theory of Planned Behavior, Prochaska's Theory of Behavior Change, Elaboration Likelihood, and Social Marketing.

5) Skills: This impact category targets the procedural aspects of knowing. Indicators include evidence that participants have learned to do something STEM-related that they could not previously do, or that they used skills they already possessed to reinforce and enhance existing STEM-related capacities. Less experienced members of a visiting group often learn skills by watching, mimicking, and jointly participating with more experienced members. Typical STEM-related skills include scientific inquiry skills (such as observation, exploration, questioning, prediction, experimentation, argumentation, interpretation, and synthesis), as well as, specific skills related to using scientific technologies or representations. This category also includes skills related to learning in the particular informal environment, for example learning how to manipulate an interactive exhibition, navigate a web site, play a computer game or collaborate with a group of youth in an after-school program. In addition, impacts can include broader skills that are related to STEM themes or are linked to lifelong STEM learning. Evidence for this kind of impact includes self-reported reflections by people of the development of new skills, or the practice of developing skills, or direct observation by researchers/evaluators.

6. Other: This category is for impacts which cannot be fit into any of the above categories. An example might be a project which is designed to impact the creativity of its audience members. Creativity might be classified in one or in several of the five categories above, but depending on the definition a particular project uses, it might be necessary to give it is own category. There will undoubtedly be projects defining novel impacts which will simply not fit under any of the five categories given. So that's why the "Other" category exists. However, we encourage you to not to overuse the "other" category and put at least part of the impacts your project is intended to produce under knowledge, engagement, attitude, behavior, or skills, to support the NSF-wide reporting process. "Other" category impacts will mostly be *one-off* examples and will be difficult to aggregate to see multi-project trends or progress.

In the descriptions of impact categories above and throughout this book, we often refer to "changes" in describing the many different kinds of impacts a project may have. But as noted in the description of the "knowledge" category above, there are other forms of desirable impacts in addition to changes. Many projects are assessing baseline information of what people know or how they behave when using a tool such as the Internet, an exhibition, or a television program. Sometimes a project may be providing opportunities for reinforcing existing knowledge, for practicing skills, or for clarifying attitudes. Change in a quantitative measure is one of many valid, verifiable ways of demonstrating impact. For the sake of convenience, we will sometimes use *change* in describing outcomes, but our intention is to include all careful measures and evidence of impact.

Now that we have described this framework for impacts, let's look at a concrete example. Imagine that your team has decided to create a project for the public that will have the following impacts on girls: (1) create awareness for STEM careers; (2) strengthen girls' sense of competence in, and identity with, STEM; (3) enhance understanding of specific STEM processes and information in the social sciences; and (4) raise interest in STEM. Each of these outcomes can be categorized in some way using this framework. If you want to

compare your categorizations of these outcomes with the authors', our classification of these impacts is in the box at the end of this chapter.

You may have found it difficult to classify some impacts, and a couple of things are important to note in terms of this example. Two of the proposed impacts fit into the same category. This is fine, but what is essential for making the impact data serve their important function is that your outcomes are represented within the framework in one or more categories of impacts. One of the outcomes also fits into two possible categories. This is also possible, though your team, with advice from your evaluator, may decide that the best evidence for that impact will only be manifested in one of the categories. In other words, you and your team may choose to specifically define the indicator for this impact depending upon the approach and type of project you decide to pursue.

How one goes about selecting the type of project to create is the focus of the next section of this chapter. As this classification system begins to be used, the NSF will be monitoring its use and may find that it will need to provide more guidance about which kinds of impacts best fit where so that the reliability of the impact categories is maintained. After all, the database will only be useful to the degree to which PIs are consistently classifying similar impacts into the same categories.

WHAT APPROACH/TYPE OF PROJECT WILL BEST ENABLE YOUR TEAM TO ACCOMPLISH THESE GOALS AND WHY DO YOU FEEL THAT THIS IS THE BEST APPROACH TO TAKE?

Now that you have defined your intended goals it is important to think about the best way to go about accomplishing them. The truth is that many project teams begin with an idea of *what* they want to do (create an exhibition, design a community literacy project or produce a giant screen film) before they think about *why* and *for whom* they want to do it. Ideally however, this is not the case and even if it is, as was suggested at the beginning of this chapter, NSF guidelines now encourage, in fact require, the *backward research design* approach. You first think about what you want to accomplish with the target audience you feel you can best reach and then describe how the particular type of project will enable these outcomes to be accomplished. Such an approach requires starting with a clear formulation of the intended project impacts and the audience(s) which will be targeted. Then by working back in a disciplined and systematic way, you can define project goals, as well as develop the program elements and strategies that can effectively achieve them.

Ultimately, defining and prioritizing outcomes requires value judgments and decision-making based on research, a clear sense of your institutional/organizational assets, and the expertise of your team. In certain contexts, some outcomes will be more appropriate, valuable, or useful than others. Thus, it is important to understand the nature of different types of informal learning experiences and the typical outcomes that are best accomplished or afforded in different types of projects. Some of this can be found in the literature and some is alluded to in subsequent chapters as individual types of ISE projects are specifically discussed in terms of the framework of impacts. In selecting outcomes, remember that it should be as specific to the experience you

are designing as possible, and be something that is more optimally accomplished through that experience than another.

Let's go back to our earlier example. Your project team identified girls as a target audience and the following four outcome goals: (1) create awareness for STEM careers; (2) strengthen girls' sense of competence in, and identity with, STEM; (3) enhance understanding of specific STEM processes and information in the social sciences; and (4) raise interest in STEM. You have been reading the literature and talking to other professionals and based upon your research and knowledge of girls, you and your project team have decided that the best approach is to design a youth program. There are some additional things you need to think through though.

(1) What age and background of girls will you work with?

(2) Are there certain ages or backgrounds that could benefit most from this project or would enable you to have the greatest impact?

(3) What is a topic of STEM that might appeal to girls of the age and background you chose?

All of these are decisions that you and your team must make in the course of developing a project. Whatever kind of project you decide to pursue, you must decide who the target audience will be as specifically as possible.

Based on further research, you and your team decide to work with pre-adolescent and adolescent girls (ages 11-16) who have not traditionally been involved or interested in science. Given the social nature of girls at this age, you also decide to focus on an aspect of science not often selected: introducing them to various aspects of the social sciences including psychology, sociology and anthropology. You begin to develop your ideas for what this program will be, over how long a period and with how many young women. For example, girls will actively learn about and engage in the social sciences, designing experiments, learning interviewing, oral history, observational and ethnographic techniques and how to compile and make sense of data. It will be an intensive summer program with opportunities for extensions and special projects in the school year. Girls can participate over the course of 3 summers, gradually becoming mentors and facilitators to the younger girls entering the program and so on.

As you and your team make each of these decisions be clear about why you have made each one and build an internal case for why you feel this is the best approach to take. Rigor in your decision-making and thought process will be invaluable as you make the case for your proposed project. Your outcome goals should naturally flow into what you intend to do, for whom and why. By thinking through each of these aspects of your project carefully at this stage, you will be in a far better position to write a solid, compelling and exceedingly fundable proposal.

HOW WILL YOU KNOW WHETHER THE ACTIVITIES OF THE PROJECT ACCOMPLISHED THEIR INTENDED GOALS AND OBJECTIVES, AND WITH WHAT EVIDENCE WILL YOU SUPPORT THE ASSERTION THAT THEY DID?

An increasingly important component of ISE proposals is the summative evaluation plan. This is the part of your proposal that will enable you to demonstrate when the project is done; whether, and if so how, you accomplished your project goals and objectives; and with what evidence you support this assertion. This is critical information for the ISE field as it competes for resources, and needs to justify the importance and public value of such experiences. This is the importance and value of the NSF's effort to create this framework for impacts and the online project monitoring system.

However, it is one thing to have an idea for a project and a vision for the broad impacts that it will accomplish and quite another to be able to actually demonstrate in measurable ways that these impacts have been accomplished. Thus, a plan for how your project will evaluate its accomplishments is an essential part of your proposal.

There are many approaches to this step, but one that works well is to develop a Logic Model, ideally developed by working with your identified summative evaluator. A Logic Model is a process which helps you describe the focus/topic of an exhibition/program, present the planned activities, and detail the anticipated outcomes and measures you (and your evaluator, of course!) will use to assess whether those outcomes/impacts have been accomplished (called indicators in evaluation language). The steps taken to develop your outcome objectives/ indicators for a Logic Model are very similar to the process we encouraged you to use to develop the idea for your project. For a more complete discussion of logic models, with illustrations, see the section, *Using Logic Models to Identify Desired Impacts and Audience Objectives***,** beginning on page 34.

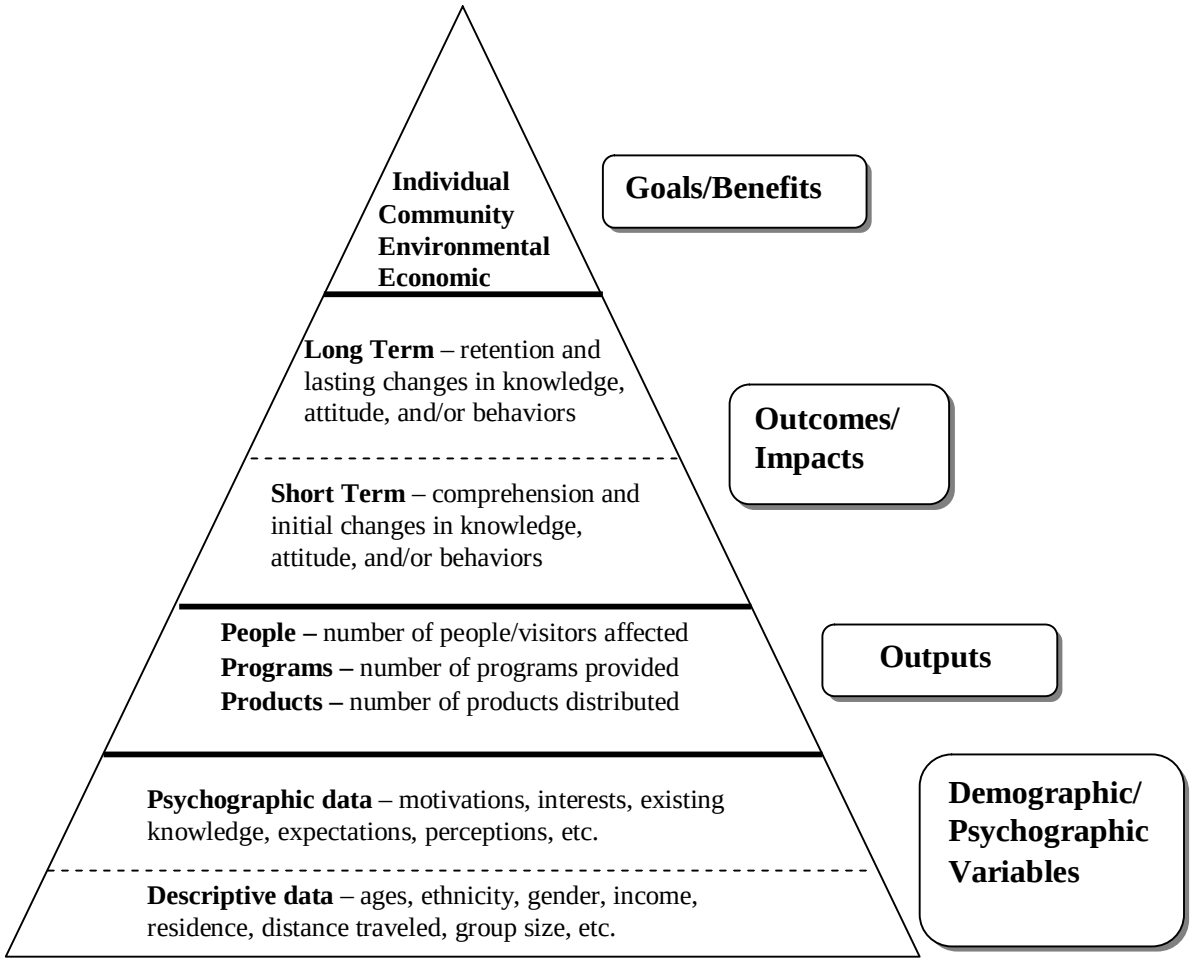
This approach distinguishes between (a) demographic/psychographic variables that describe participants (e.g. age, ethnicity, motivation for participating, prior knowledge, etc.); (b) "outputs" – for instance, the number of activities that are part of your program or how many people will participate in those activities; (c)"outcomes/impacts," – what participants will do, think or feel as a result of their experience; and (d) "goals/benefits. For your proposal, you will need to develop outcomes/impacts, both short term and long term, as shown in Figure 3-1.

Returning to the hypothetical youth program proposed above, outputs would include: "that x number of girls participate in the program each year, or that three sequential programs will be designed in the first year." This is what they will do in the program. Outcomes/ Impacts are the four broad goals that the project team hopes will be accomplished if girls participate: (1) create awareness for STEM careers; (2) strengthen girls' sense of competence in, and identity with, STEM; (3) enhance understanding of specific STEM processes and information in the social sciences; and (4) raise interest in STEM.

However, there is another important step. One must move from the broad impact of "strengthen girls' sense of competence in, and identity with, STEM," to a more precise, measurable

objective or indicator that will concretely demonstrate that the goal was accomplished. In research lingo this is referred to as "operationalizing the impacts." In the case of this particular goal, indicators might include: (1) In interviews, girls participating in the program are able to discuss their comfort and success conducting interviews as compared to how they felt prior to the program (pre-post); (2) Girls participating in the program score higher on a self-esteem measure (pre-post); (3) Girls participating in the program are able to talk about the role of science and its importance in their life as compared to a similar group of girls who did not participate. These are just examples but demonstrate how one can move from a broad impact to a more measurable or observable objective.

Figure 3-1. Hierarchy of Anticipated Outcomes (Adapted from Wells, M & Butler, B. (2004)



There are a few other important ideas imbedded within this example. First, it is always important when discussing evidence of impact to consider: Evidence for whom and compared to what? Depending upon the audience for the evaluation findings and the questions being raised, different levels of rigor are required. However, at a minimum if at all possible, baseline data should be gathered so that one understands where the particular audience was in terms of any indicators prior to the experience. In other cases as the example above shows, it may be important to demonstrate that your program had an impact by comparing observed impacts to a comparable audience that did not participate. Such approaches are more time-intensive and costly and may not always be appropriate but they should be considered when possible.

The challenge for choosing an appropriate design is to select approaches that are appropriate to specific types of ISE projects and contexts. In subsequent chapters there will be additional discussion about the "nuts and bolts" of evaluation including research designs, data collection approaches, sample sizes and so on. An understanding of these complexities is also an area of expertise that your selected external evaluator can contribute to the project.

HOW WILL YOU AND YOUR TEAM ENSURE THAT UNANTICIPATED OUTCOMES ARE ALSO DOCUMENTED?

A major discussion at the meeting leading up to the writing of this guide focused on the importance of being open to unanticipated outcomes, impacts that often are the most interesting of all. You may discover unintended outcomes during your evaluation, and it will help both your project and the field if you adapt your evaluation plan to capture those outcomes. For example, the girls participating in your program might show an unexpected interest in the biographies of women scientists, or in the technology used in the oral histories (video/audio recordings, digital photography) rather than in the social sciences themselves. These are positive outcomes, but not ones you expected. So even though it might be late in the project, it would important to pursue information about these outcomes and include them in your impact evaluation. And it is still important for evaluation approaches to not only focus on whether outcome goals and objectives are accomplished, but also to probe *how, why* and *for whom* impacts are observed.

Finally, we emphasize that this system is to be used to categorize some of your intended impacts. Not all impacts will, or even should, fit neatly into the categories of the framework. There is an "Other" category as noted above, and project teams and evaluators should feel comfortable using it when necessary and not feel constrained by the system. The reporting system and the impact categories are serving an important role but will not capture everything which happens in the NSF ISE portfolio. What has made this field unique is its creative spirit and innovation - and the desire is for that to continue - while at the same time the ISE program compiles some comparable data with which to understand the important impacts and roles that ISE experiences play in lifelong STEM learning.

REFERENCES

Munley, Mary Ellen. (1986). Asking the right questions. *Museum News, 64*(3), 18-23

Wiggins, Grant and McTighe, J. (2005*) Understanding by Design*. Upper Saddle River, NJ: Prentice Hall Inc.

Our suggested categories of outcomes for the sample project discussed in this chapter:

(1) Creating awareness for science careers (awareness / knowledge / understanding)

(2) Strengthening girls' sense of competence in, and identity with, science (attitude and development/reinforcement of skills)

(3) Enhancing understanding of specific science processes and information in the social sciences (awareness / knowledge / understanding)

(4) Raising interest in science (engagement/interest)

CHAPTER 4 TOOLS, TIPS, AND COMMON ISSUES IN EVALUATION EXPERIMENTAL DESIGN CHOICES

Sue Allen

This chapter deals with several aspects of evaluation that are common to most projects. First we look at some of the choices you can make in evaluation design to gather the impact data your project needs. Experimental designs are very powerful for some purposes, but often present difficulties in ISE settings. We'll discuss those methods first, and then summarize the broader array of strategies you can consider. In Part II, you will find examples of most of the choices described here, including both experimental designs and naturalistic methods.

EXPERIMENTAL DESIGN CHOICES

Sue Allen

As noted in Chapter 2, experimental study designs are not necessarily more appropriate than naturalistic ones; rather, you should use the most rigorous study designs that are best suited to the nature of the project and its intended outcomes. When a project can consider conducting experimental designs with representative sampling, the following are study designs worth exploring:

(a) Randomized controlled trial: One theoretically powerful experimental design is the randomized controlled trial (sometimes called "randomized clinical trial," "RCT," or "true experimental design"). It is a pre-post study with comparison group, in which participants are assessed before and after experiencing the project materials, and their learning is compared with that of a control group who were also assessed twice but without experiencing the materials. Ideally, audience members are randomly assigned to these two groups.

This design is often summarized as:

R OXO

R O O

where the "R" indicates random assignment, the "O's" indicate an assessment (often called a pre-test or a post-test), and the two rows indicate that one group experiences the project materials "X" between their pre and post-test, while the other does not.

If properly implemented, this design rules out many competing possible causes of audience members' learning (such as practice with the assessment), but it is expensive, and potentially taxing for the audience members – especially those who are told they cannot immediately see

the exhibition / film / other deliverable and are then assessed twice for no obvious reason. This study design is rarely used in practice to assess audience member's learning in informal environments.

(b) Randomized post-only design: An equally powerful design that is somewhat more feasible in many ISE projects is the randomized post-only design. As in the RCT design above, participants are randomly assigned either to a group that experiences the project deliverables or a group that does not experience them (at least, not until after the study). All participants are then assessed once, and any differences are attributed to the effect of the project materials.

This design is often summarized as:

R XO
R O

This kind of study is somewhat less expensive and taxing to participants than the RCT design, but requires assessing a relatively large number of participants, and it requires that they be randomly assigned to the control or treatment groups (which may be unrealistic). Random assignment can sometimes be achieved by recruiting audience members who are willing to experience two sets of materials (say, exhibitions) in any order; they can then be randomly assigned to see the target exhibition either first or second.

(c) Using comparisons where possible: It is often feasible to provide evidence of impact with at least some form of comparison. Some kinds of assessment, such as direct questions, card-sorting tasks, or concept maps, may be used before and after participants experience the project's materials (in a pre-post study design without a control group). Alternatively, participants' responses after their experience may be compared to a group of participants who have not yet had the experience, matched if possible by key demographic descriptors such as age and education level (if random assignment is not realistic).

Failing such direct comparisons, it may be possible to compare the measured indicators of participants' learning to rates reported in other literatures, such as front-end evaluation studies, summative evaluations of similar exhibitions, misconceptions literature in science education, etc. Such benchmarks can provide at least some sense of the degree to which a project has been effective as an aid to learning.

Comparisons also strengthen evidence of learning when process-based measures of learning are used. For example, if a particular exhibit engages museum visitors in asking their own questions, this becomes a stronger form of evidence if the frequency or quality of visitors' questioning is compared to that of visitors at "typical" exhibits or in other kinds of settings.

(d) Cases where comparisons are unnecessary: Some kinds of evidence do not require a comparison to be compelling, particularly when a plausible case can be made that

participants could not already have had the knowledge at the time they experienced the project's materials. Examples are:

1. visitors figuring out an exhibition's main idea(s);
2. viewers making connections between a TV program and their own lives;
3. professionals remembering their experiences in a workshop and their responses over time;
4. visitors sharing something they know that is inconceivable for them to have known previously (such as pieces of information uniquely displayed in an exhibition);
5. participants self-reporting that they had not previously realized something.

Finally, a few notes of caution when planning study designs:

• When planning experimental studies, there is often a trade-off between rigor of the design and the authenticity of the situation being studied. For example, it may be possible to design a fully randomized controlled trial, but the implementation may require that learners be constrained in ways that significantly undermine the informal, free-choice nature of their experience. Such design choices should be made thoughtfully, in consultation with an evaluator early in the project.

• While we value rigor, in experimental designs, case studies, or naturalistic observations, it is even more important that participants not be traumatized or alienated because of over-zealous assessment practices (which would also lessen validity of the results). Evaluators should therefore pilot-test their methods and be sensitive to participants' emotional responses.

• Irrespective of the rigor of their study designs, evaluators should be careful not to over-interpret their data or over-generalize their claims, lest they lead to misguided or simplistic policy decisions that may adversely affect learners in other settings.

The next section of this chapter summarizes a broad selection of possible evaluation design choices, including the ones discussed in some detail above.

AN ARRAY OF EVALUATION DESIGN CHOICES

Pat Campbell

There are a number of designs that can be used in the evaluation of your program or project. Table 4-1 provides an overview of many of the designs including some of their advantages and disadvantages.

Regardless of the evaluation design you choose, a "logic model" is a very useful tool to clarify the goals of the evaluation and the project as a whole. The following discussion provides an introduction to this increasingly popular tool.

Table 4-1: Sample Evaluation Designs

Study Type	**Design**	**Representation** (X= treatment; O=measures/evidence; R=random assignment)	**Advantages**	**Disadvantages**
Quantitative Case Study	One-shot Post-test only Design	X O	Takes fewer resources Can present a "snapshot" of a point in time	Doesn't look at change
Quasi-experimental Study	One-shot Pre-test- Post-test Design	O X O	Looks at change over time	Other things besides treatment could be causing change
Quasi-experimental Study	Post-test Only Intact Group Design	X O O	Compares to another group	Doesn't control for any initial differences in groups
Quasi-experimental Study	Pre-test- Post-test Intact Group Design	O X O O O	Allows statistical control for possible extraneous variables	Doesn't control for any effect of testing
Experimental Study	Post-test Only Design With Random Assignment	X O R O	Controls for pre test effects Random assignment reduces the chances of extraneous group differences	Random assignment is often not possible in evaluation Doesn't control for extraneous variables
Experimental Study	Pre-test- Post-test Design With Random Assignment	O X O R O O	Allows statistical control for possible extraneous variables	Random assignment is often not possible in evaluation. Doesn't control for any effect of testing
Experimental Study	Solomon Four Group Design	O X O R X O O_a O_b O_b	Strongest quantitative design controls for all possible extraneous variable	Random assignment is often not possible in evaluation Very resource intensive
Quasi-experimental Study	Time Series Design	O O X O O	Looks at longer term change	Doesn't control for extraneous variables
Ethnography	Participant observer examination of group behaviors and patterns	NA	Explores complex effects over time	Resource intensive Story telling approach may limit audience Potential observer bias
Case Study	Exploration of a case (or multiple cases) over time	NA	Provides an in-depth view Elaborates on quantitative data	Limited generalizability
Content Analysis	Systematic identification of properties of large amounts of textual information	NA	Looks directly at communication Allows for quantitative and qualitative analysis	Tends too often to simply consist of word counts Can disregard the context that produced the text
Mixed Methods Study	Use of more than one of the above designs	NA	Can counteract the disadvantages of any one design	Requires care in interpreting across method types.

Adapted from: Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Chicago: Rand McNally, 1963). Gary Ingersoll, *Experimental Methods (in Encyclopedia of Educational Research (Fifth Edition); Harold Mitzel ed. New York: The Free Press 1982. pp 624-631. Lydia's Tutorial Qualitative Research Methods* http://www.socialresearchmethods.net/tutorial/Mensah/default.htm Accessed April 15, 2007. Writing@CSU http://writing.colostate.edu/index.cfm Accessed April 15, 2007.

USING LOGIC MODELS TO IDENTIFY DESIRED IMPACTS AND AUDIENCE OBJECTIVES

Gary Silverstein

Program staff and evaluators frequently use a "logic model" (sometimes called "a theory of action") to think through how they intend to achieve and document their intended outcomes. By illustrating the underlying rationale of a program or activity, logic models can be used to show how different facets of an intervention are linked. As such, they provide funders and stakeholders with a visual representation of the resources available to operate a set of activities—as well as an overall framework for understanding the relationship between the program's inputs (i.e., resources and activities) and the changes or results those inputs are designed to achieve. They also provide evaluators with a useful roadmap for determining the range of questions that need to be addressed as part of an overall assessment of program implementation and impact.

Figure 4-1 presents a general logic model for the ISE program that includes the following components. These components include those discussed in Chapter 3, and additional elements useful for the format of a Logic Model.

- **Inputs—**the resources that are brought to a project. Typically, resources are defined in terms of funding sources or in-kind contributions.

- **Activities—**the actions that are undertaken by the project to bring about desired ends—e.g., the development of a museum exhibit or radio program.

- **Outputs—**the immediate results of an action (e.g., services, events, and products) that document the extent of implementation of a particular activity. They are typically expressed numerically—e.g., the number of persons who visit a museum exhibit or listen to a radio program.

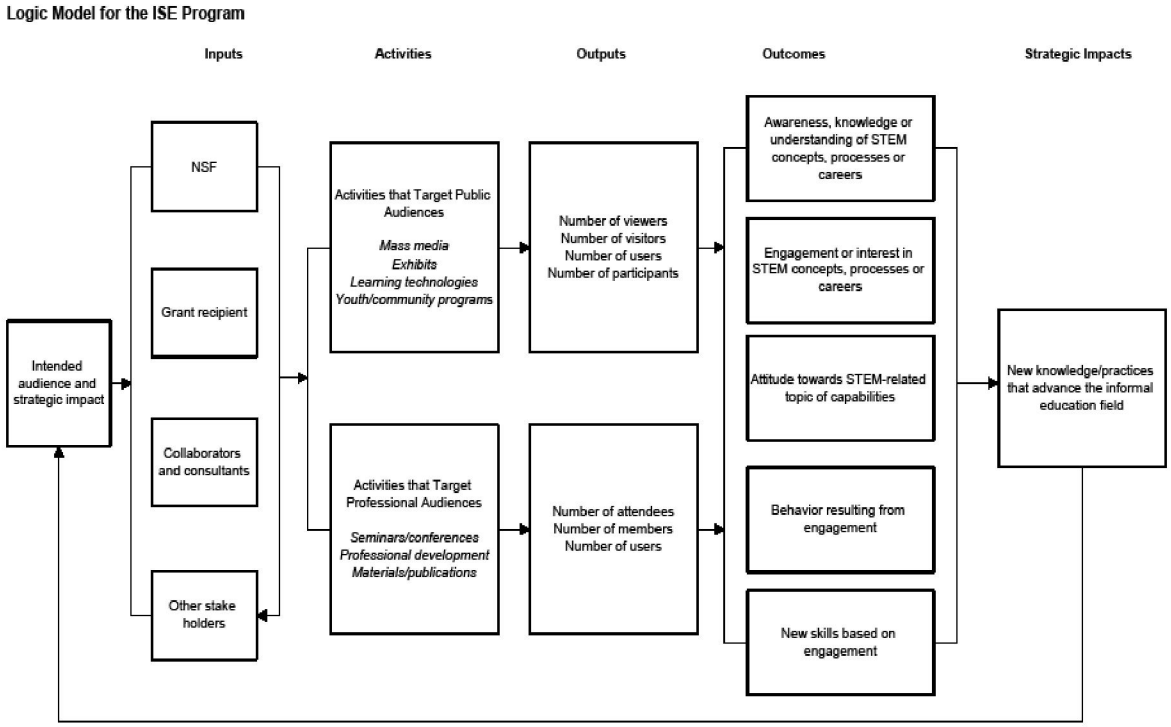
- **Outcomes—**the changes that show movement toward achieving ultimate goals and objectives—e.g., the number of persons who enhance their knowledge as a result of visiting a museum exhibit or listening to a radio program.

- **Strategic Impact—**a term from the ISE program solicitation that refers to steps taken by individual projects to "improve theory or practice through approaches, strategies, findings, or models having impact on the institutions or systems that promote informal learning." The purpose is for ISE projects to identify and influence a leverage point for advancing the informal education field in a meaningful way so that they can extend their impacts beyond those directly reached by the project deliverables.

Project-specific logic models can also depict contextual factors (i.e., conditions that facilitate and/or hinder the extent to which an ISE project is able to implement its proposed approach and attain its anticipated outcomes).

In reviewing this model for the overall ISE program, it is important to note that two interrelated types of program activities are being addressed: activities that target *public* audiences (e.g., mass media, exhibits, learning technologies, and youth/community programs) and activities that target *professional* audiences (e.g., seminars/conferences, professional development, materials/ publications). It is also important to note that this logic model illustrates the range of activities, outputs, outcomes, and strategic impacts for the overall ISE program. A model developed for a single ISE project would only include those activities, outputs, outcomes, and strategic impacts that the project is trying to put into place.

Figure 4-1. Logic Model for the ISE Program



Although they are useful for overall planning, logic models do not provide sufficient detail about the range of outcomes and evidence that will be used to demonstrate that a program's activities and intended impacts have been attained. Therefore, programs will often identify (1) specific and measurable objectives that can be used to assess progress toward each of their desired impacts, and (2) the overall methodology and individual data elements that will be used to ascertain whether each objective has been met. Although these objectives may be reflected on the logic model, the level of detail specified at this stage can provide stakeholders with significantly more information about what a project is designed to achieve.

The following example illustrates the uses of logic models and corresponding objectives. Example 1 provides background information about a hypothetical project designed to increase visitors' knowledge of—and interest in—cars that use hybrid engines. As shown in the accompanying logic model (Figure 4-2), the hypothetical project was also designed to increase the number of museums that make podcasts available to their visitors. Example 1 builds upon the logic model by delineating audience objectives for each of the anticipated outcomes. The evidence provides information about the data used by the project's evaluator to ascertain whether each of the public and professional audience objectives was met. The numbers used in this table, like the project itself, are purely illustrative, and are not intended to suggest any kind of typical or recommended findings.

Example 1: Information about the (imaginary) Hybrid Engine Exhibit

This Advanced Technology Project involved the creation and presentation of a museum exhibit on the topic of hybrid vehicles. The hybrid engine currently represents one of the biggest trends in automobiles, but few among the general public can articulate how hybrid vehicles provide the benefits claimed for them. The Museum of Advanced Technology received a grant from the ISE program to develop "How Science and Engineering Drive Hybrid Vehicles"—a 6-month exhibit on the practical and environmental benefits of driving cars that have a hybrid engine. The cost of viewing this particular exhibit was $6 above the admission fee normally assessed to the museum's visitors.

Findings from a feasibility study (front end evaluation) conducted for this exhibit revealed visitors would prefer to listen to pre-recorded information about a technical topic than read a series of technical labels. Therefore, a secondary purpose of the ISE project was to assess the feasibility and impact of using podcasts to impart information about a given topic. Specifically, the project examined whether the use of podcasts increased visitors' interest in and understanding of hybrid technology, a result which could be useful to professional audiences as well as to the public audience.

The evaluation plan for the exhibit included an experimental design that enabled the external evaluator to examine the added value of using podcasts to engage adult visitors and impart knowledge about the benefits of hybrid vehicles. Specifically, individuals participating in the museum's annual membership program were mailed a coupon that enabled them to view the hybrid vehicle exhibit at no additional cost. Half of these individuals also received a brochure about the podcasts (including information about how to access the podcasts before visiting the exhibit). The other half were in a control group that did not receive the brochure with their coupon—and information about the podcasts was not publicized to the general public until after a telephone survey (described below) was completed.

Three months after the exhibit closed, all of the individuals that received a coupon were asked to participate in a brief telephone survey. The purpose of the survey was to (1) ascertain whether these individuals actually visited the exhibit, (2) assess respondents' knowledge about the benefits of hybrid vehicles, and (3) examine

respondents' interest, attitudes, and behavior on the subject of hybrid vehicles. In addition, individuals who received the brochure were asked additional questions about whether they used the podcasts. Those who had used the podcasts were asked about their experiences; those who had not were asked about why they had not attempted to access the website that contained a link to the podcast.

Survey findings were used to compare the knowledge, attitudes, and behaviors of those individuals who did and did not attend the exhibit. Equally important, for those who attended the exhibit, findings were also used to compare the knowledge and interest of those who were and were not provided access to the podcast website—with additional analyses being used to isolate whether the self-reported knowledge, interest, and behaviors of those individuals who actually listened to the podcasts differed from those in the control group who did not have the opportunity to do so.

After the exhibit closed and survey data had been analyzed, lessons learned were shared with professional museum exhibit developers from across the country. During this seminar, they also received instruction on how to integrate podcasts into a science museum exhibit. Follow-up surveys were used to document knowledge gained as a result of the seminar, as well as whether participants have made use of podcasts in their own museums.

Figure 4-2. Logic Model Example

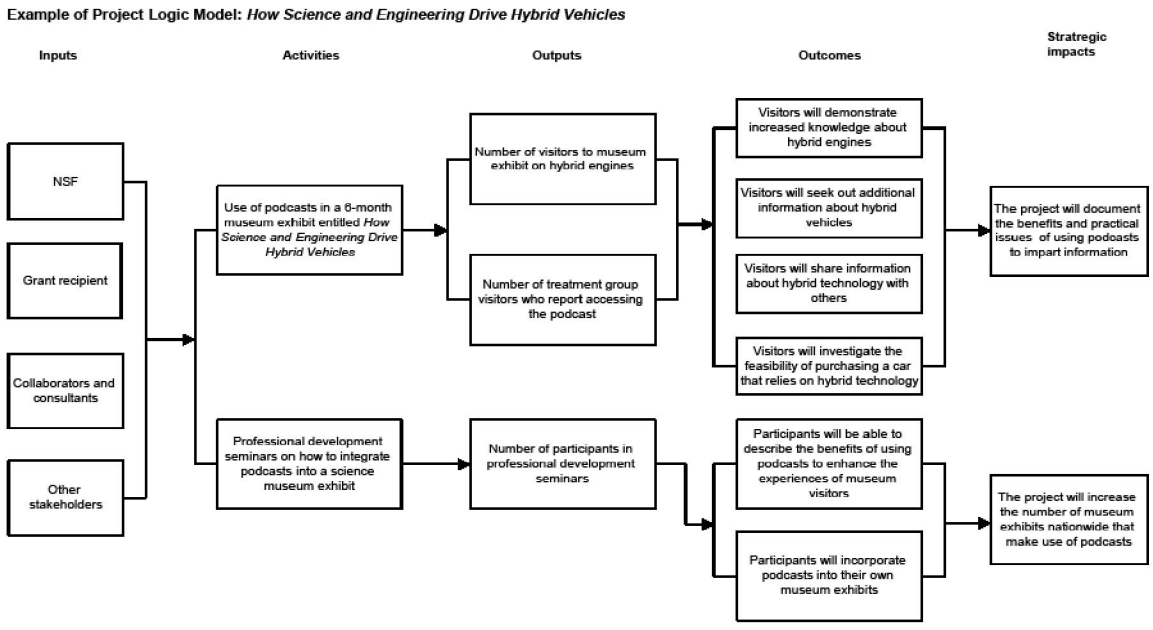


Table 4-2. Impact Worksheet: Example for Museum Exhibit on How Science and Engineering Drive Hybrid Vehicles

Impact	Impact Category	Audience Objective	Evidence
Public Audience			
The use of podcasts will result in an increase in knowledge about hybrid engines	Knowledge/ Comprehension	The use of podcasts will increase adult visitors' knowledge about the *practical* benefits of driving cars with a hybrid engine	55 percent of survey respondents who had access to the podcast described at least one practical benefit of driving cars with a hybrid engine (compared with 40 percent of control group respondents)
		The use of podcasts will increase adult visitors' knowledge about the *environmental* benefits of driving cars with a hybrid engine	80 percent of survey respondents who had access to the podcast described at least one environmental benefit of driving cars with a hybrid engine (compared with 55 percent of control group respondents)
The use of podcasts will result in an increase in interest about hybrid vehicles	Engagement/ Interest	Adult visitors who had access to the podcast will seek out additional information on the Internet about hybrid engines after visiting the exhibit	50 percent of survey respondents who had access to the podcast performed an Internet search related to hybrid technology since visiting the exhibit (compared with 34 percent of control group respondents)
		Adult visitors who had access to the podcast will purchase a book related to hybrid technology	22 percent of survey respondents who had access to the podcast purchased a book related to hybrid technology since visiting the exhibit (compared with 8 percent of control group respondents)

Table 4-2. Impact Worksheet: Example for Museum Exhibit on How Science and Engineering Drive Hybrid Vehicles

Impact	Impact Category	Audience Objective	Evidence
		Adult visitors who had access to the podcast will share information about the exhibit and/or hybrid engine technology with others (e.g., family, friends, colleagues)	45 percent of survey respondents who had access to the podcast shared information about the exhibit and/or hybrid engines with others (compared with 8 percent of control group respondents)
The use of podcasts will produce a change in behavior	Behavior	Adult visitors who had access to the podcast will consider purchasing a car that uses a hybrid engine	8 percent of survey respondents who had access to the podcast investigated the feasibility of purchasing a car with a hybrid engine (compared with 7 percent of control group respondents)
Professional Audience			
The seminar will result in an increase in awareness about podcasts	Knowledge/ comprehension	Exhibit developers who attend a seminar will be able to describe the benefits of using podcasts to enhance the experiences of museum visitors	100 percent of survey respondents described at least one benefit of using podcasts to enhance the experiences of museum visitors
		Exhibit developers who attend a seminar will be able to describe an approach they would use to incorporate podcasts into their own museum	85 percent of survey respondents described an approach they would use to incorporate podcasts into their own museum
The seminar will result in an increase in podcast use	Behavior	Exhibit developers who attend a seminar will incorporate podcasts into their own museum exhibits	40 percent of survey respondents described how they were able to incorporate podcasts into one of their own museum exhibits

This chapter closes with a look at a number of issues and opportunities that may arise in ISE evaluations, regardless of the particular study design or tools used.

SPECIFIC ISSUES IN EVALUATION

Sue Allen

There are some common issues that may arise when conducting summative evaluations and reporting impacts from any informal education project. Here are some of these common issues, and how we think they may be approached.

- Diversity of audiences, experiences, & impacts: Compared with school-based science education, ISE project materials often attract heterogeneous public audiences. Learners range in age from newborns to centenarians, in formal science background from novices to professional experts, and in language, culture, and motivation across many dimensions. Not only is the range of learners large, but each person has a unique experience because they move individually through a set of choices, interactions, and interpretations. All of this variation makes ISE deliverables particularly challenging to evaluate. Some implications include:

> *People learn different things, not just different amounts.*
> Because learners build on their existing knowledge and experience, the diversity of audience backgrounds means that project deliverables (exhibitions, TV programs, websites, youth programs, etc.) will offer a large range of possible learning outcomes. Assessments are most likely to show learning impacts if they are open-ended enough to capture this range, including unintended impacts.
>
> *Assessments should be inclusive.*
> With such diverse audiences, ISE project deliverables provide opportunities to engage and study a range of learners who might be excluded from typical school-based studies: young children, adults, seniors, people with disabilities, speakers of languages other than English, and people who would never choose to take a formal course in science. It is important to design assessments that include such groups as much as possible, using methods that are respectful, appropriate, and comfortable. For example, it may be useful to supplement visitor interviews with alternative forms of assessment, such as drawings, taking photographs, sorting tasks, narratives, or think-alouds, (Appendix B lists some resources in this area.)

- Avoiding pre-test trauma: Conducting pre-post comparisons is often problematic when assessing ISE materials, particularly in the case of assessing skills. This is because learners may be traumatized by an evaluator's efforts to prove that they are truly unable to do something, rather than simply not choosing to do it. In such situations, it may be more appropriate to use a process-based measure: assessing the skills that learners use while engaged with the project materials, and comparing this result with any benchmarks of "typical" skill use, if possible. Similarly, pre-test trauma may be an issue when assessing visitors' knowledge or understanding about a specific scientific area. If so, it may be more appropriate to assess visitors' knowledge-building processes while using the materials, or to rely on visitors' self-reports of what was new

to them, after they finish. Sometimes people can reflect on each other's learning and even give evidence for it (especially parents describing what they think their children learned or practiced).

- Doing versus knowing: Visitors may not be consciously aware or able to name their skills, although if they can, this is worthy of additional report.

- Assessing exhibition impacts with special and underserved audiences: Projects may target special audiences, including different cultural groups, ages, genders, ethnicities, abilities, and languages. Such target audiences may require custom-designed evaluation approaches in order for the evaluation to be valid (which includes being non-threatening). Appendix B includes a list of studies and methods used in evaluation with some special audiences. There is ongoing interest in tracking evidence of impacts for special and underserved audiences. In cases where such information is available, it should be reported in addition to findings for the broader audience, for the benefit of the field.

- Reporting negative findings: Any evaluation may yield negative findings, for example, that a particular aspect of the project had no impact on the audience, or even that it created misunderstandings. Such negative findings constitute an important learning opportunity for the project team and the broader field. These are important to report, especially if the project team has a plausible interpretation of why a particular impact was not observed. The goal of NSF's impact reporting system is to give a realistic account of what was found, for purposes of generalizing and accumulating evidence. We would all love to have only positive findings, and it takes some courage to report when something didn't work. But negative findings demonstrate that the project team members were open to maximizing their learning, and recognized the value to the field of demonstrating both what worked and what did not. Complete reporting, while sometimes uncomfortable, can establish the credibility and reliability of a project team in a most convincing manner.

- Case studies and findings from naturalistic methods: Methodologies should depend on the nature of the project, question, and assessment being undertaken. Case studies and qualitative analyses are no less valuable than experimental data as forms of evidence, and all studies done should be included in the project's final report. Because the focus of this book is an impact-category-based reporting system for data that can be accumulated and generalized, it is more likely that experimental and quantitative studies conducted will be selected for reporting in this form. However, case studies and qualitative studies can be included in summary form where appropriate, especially in cases where they present strong evidence of learning.

- Sampling: There is no single answer as to how visitors or professional audiences should be sampled. When project teams decide which data to include in the impacts report, they can use randomly sampled data (i.e., representative of the larger audience served), which is likely to be amenable to generalization and accumulation. But they could also use purposive sampling (i.e., case studies selected for in-depth study). Random sample data should generally be included in the impact indicators, while purposive samples can be reported in the final project report and elsewhere.

- Making recordings of participants: Audio and video recordings of visitors using project materials (such as exhibitions, professional meetings, or media materials) can provide a highly fruitful form of data for analysis of impacts, especially in projects that focus on skills. However, such recordings and other aspects of specific audience studies may require approval of an Institutional Review Board (IRB). "Human subject" protocols and Institutional Review Boards are beyond the scope of this book, but may (or may not) be applicable to your evaluation plan. Because requirements and good practices in this arena are in flux, we suggest that you consult the NSF web site, which is regularly updated with rules and references: http://www.nsf.gov/bfa/dias/policy/human.jsp.

PART II IMPACT EVALUATION FOR VARIOUS PROGRAM AREAS OF INFORMAL SCIENCE EDUCATION

Every chapter of Part II covers essentially the same ground, but with examples, language, and emphases selected for particular program areas of Informal Science Education. Sometimes these differences are as small as the word used to describe the beneficiaries of a project: "visitors," "audiences," "viewers," or "participants," each of which is slightly jarring to professionals in other areas of ISE. At other times there are great differences in how program areas can reach the people they serve, and the relationships they have with those individuals. These chapters do assume you have read Part I, but each chapter in Part II then stands on its own.

CHAPTER 5 EVALUATING EXHIBITIONS

Sue Allen

DEFINITION

In this section, we use the term "exhibition" to refer to a group of individual elements that are related to each other in some way, and funded as a collection by the ISE grant. They may be thematically related, based on a single principle, a big idea, or a number of content-based connections that make them coherent from a content perspective. Alternatively, they may have a process connection, such as being an innovative genre of exhibit that offers visitors some particular kind of activity or experience. This section gives guidelines for applying the framework of ISE impacts to evaluation of exhibitions, recognizing that an exhibition may be only one of the deliverables of an ISE grant.

BROAD IMPACTS AS APPLIED TO EXHIBITIONS

1) Knowledge: This category of impact emphasizes things that visitors consciously know. It captures knowledge, awareness, or understanding that can be expressed by learners in their own words or images, whether that be during, immediately after, or long after, their experience. The content of this impact is dictated by the project topic: it could be to help visitors understand a STEM-related topic, concept, principle, or theory, the history or philosophy of science, or science as a process.

Evidence for this impact includes changes in visitors' knowledge (directly assessed or self-reported). It also includes cognitive activity, so one indicator might be visitors' success in verbally synthesizing their experiences to identify the intended "big idea' of the exhibition, or their generation of appropriate connections between the STEM aspects of the exhibition and their own lives. It can also include evidence of visitors noticing relevant features of the exhibition or the natural world, understanding concepts embedded in interactive experiences, reinforcing their prior knowledge, making inferences, or building an experiential basis for future abstractions to refer to (though this is more difficult to assess). Another form of evidence in this category is memory of an experience over time, especially aspects of the experience that relate to STEM concepts, processes, or activities. Lastly, visitors' reflections on, and monitoring of, their own learning falls into this category, including insights they have either during or after their exhibition experience.

2) Engagement: This category of impact captures the excitement and involvement of learners in a topic, area, or aspect of science. One key assumption of the ISE field is that exhibits can provide visitors with an accessible, enjoyable, compelling introduction to an area of science that they

may not yet know much about. Visitors may self-report a range of possible emotional responses, such as joy, delight, awe, wonder, appreciation, surprise, caring, inspiration, intrigue, satisfaction, and meaningfulness, as well as negative emotions such as horror, anger, or sadness, which may be appropriate for the subject matter in some types of exhibition.

This impact also includes participation and engagement, a prerequisite for other types of learning but also linked to interest. It could be supported either by evidence that an exhibition has evoked short-lived interest (e.g., is highly attractive and/or sustains visitors for unusually long periods of time, or is used by an audience not usually interested in the topic), or by evidence that it furthered longer-term interest, supporting already-interested visitors in their increasing involvement in a topic or area of science. This impact is particularly likely to be a focus for projects that aim to involve new or under-represented audiences.

3) Attitude: While somewhat similar to the previous impact, attitude goes beyond engagement in particular activities (such as using the exhibition), to encompass the longer-term stances that visitors take toward groups or issues. For example, visitors may change in their degree of respect, empathy, support, allegiance, or appreciation. An ISE project may generate attitudes where none existed before (e.g. visitors who hadn't heard of a novel field such as nanotechnology may support its development after using an exhibition on that topic), or may change attitudes (e.g., visitors who felt uncomfortable or even alienated by science may feel more comfortable with some aspects of it).

The targeted attitude may be toward a STEM-related topic (statistics), a group of people (ancient Mayan astronomers), species or ecosystems (alligators, swamps), activities (building of a new particle accelerator), theories (global climate change) or careers (forensic science). Attitudes and attitude changes don't necessarily have to be positive; a successful project might have as its goal an increase in the skepticism of visitors, for example towards commercial advertising or pseudoscientific claims.

Indicators for the "attitudes" impact tend to be less reliable than indicators of knowledge or engagement, because they rely exclusively on self-report by visitors, and visitors may not be entirely honest, even with themselves, about their attitudes to sensitive topics. For this reason it is desirable to assess for attitude in multiple contexts if possible: e.g., self-report immediately after seeing the exhibition, interview weeks or months later, behavioral evidence such as actions to find out more about a previously vilified group or practice, reports by any others who might have observed a change, etc. Also, because changes in attitude are likely to be relatively small for most exhibitions, we recommend assessments that encompass a broad range of possible positions and are sensitive to small shifts (e.g., scales that assess visitors' degree of agreement with statements that include extreme views from various perspectives).

4) Behavior: Some ISE projects propose to change visitors' long-term behaviors, in their lives, beyond the exhibit. This type of impact is particularly targeted in projects that are environmental in focus. Evidence of behavior change might include visitors' self-reported intentions to change their behavior, and longitudinal follow-ups with them (or others) to determine whether such behavior change has, in fact, happened. Sometimes a powerful exhibit experience leads learners to engage with materials and media beyond the exhibition, so this is an impact that might involve

looking across related media within a single project for trajectories of behavior change over time. Evidence may even include measures not directly linked to an individual visitor: e.g., increased sales of high-efficiency light bulbs in hardware stores, following the opening of an exhibition on energy use at a nearby museum.

Like attitude, evidence for behavior change is susceptible to visitors' pleasing bias (their tendency to say what they think the researcher wants to hear), especially when the desired behavior change is obvious. For this reason, follow-up interviews or questionnaires are especially important. Also, because long-term behavior change is notoriously difficult to effect (especially with as short an intervention as an exhibition), assessments should be sensitive to small changes, including visitors doing desirable behaviors occasionally, reducing the frequency of undesirable behaviors, stopping to question their choices, or talking about possible behavior changes with others.

There is a potentially important relationship between the categories of attitude and behavior change. For projects where this relationship is a central focus, evaluators may wish to refer to background theories such as: the Theory of Reasoned Action, Theory of Planned Behavior, Elaboration Likelihood, and Social Marketing.

5) Skills: This category of impact targets the procedural aspects of knowing, as opposed to the declarative aspects captured by the "knowledge" impact (described above). Indicators would include any evidence that visitors have learned to do something STEM-related that they could not previously do. Another form of evidence would be visitors actively using STEM-related skills that they already do possess, insofar as they are reinforcing their capacities through practice and rehearsal, particularly in a social context. Less experienced members of a visiting group (such as children) often learn skills by watching, mimicking, and jointly participating with more experienced members (such as adults).

Typical STEM-related skills include scientific inquiry skills (such as observation, exploration, questioning, prediction, experimentation, argumentation, interpretation, and summarization) as well as more specific skills related to technology and devices (e.g., using instruments such as microscopes or telescopes successfully). They also include skills related to learning in the particular informal environment: how to use interactive exhibits, how to draw relevant information from labels or other interpretive devices, and how to learn effectively with others of different skill-levels – sharing resources, teaching, scaffolding, negotiating activity. They may also include broader skills (such as linguistic, logical-mathematical, spatial, bodily-kinesthetic, or even interpersonal and social skills), as long as these are plausibly related to STEM themes or are linked to STEM learning later in life.

HYPOTHETICAL EXAMPLES

This section lists four hypothetical ISE-funded exhibition projects, and shows how the hypothetical findings from each could be put into the required format involving large-scale impacts and more specific audience objectives.

A) Hypothetical Exhibition: "Plants: unsung heroes of our planet"

Project goals as stated in grant proposal: We aim to help visitors appreciate the fundamental role that plants play in our ecosystems; to encourage visitors to marvel at the role of plants as carbon dioxide consumers and oxygen producers; to realize that, in spite of their immobility, plants are highly complex and sophisticated living things; and, to address some common misconceptions about plants.

Thus, the intended project impacts would be:

1) Knowledge: Visitors will understand aspects of the basic chemistry, properties, and role of plants in ecosystems.
2) Attitude: Visitors will appreciate plants, both in terms of their sophistication as organisms and their vital role on planet earth.

Relevant findings would then serve as evidence for these impacts.

Evidence of impact on knowledge (from hypothetical results):

• Visitors knew that plants create their own food: When asked to sort cards with written characteristics of living things, 60% of adults leaving the exhibition could correctly identify "create their own food" as characteristic of plants but not animals. When asked to explain their choice in more detail, a smaller percent, 40%, understood that plants assembled their food from simpler materials. A common misconception, even among those who knew that plants make their own food, was that this food is sucked by plants from the soil (35% of adults). There was no control or comparison group, but reference to the literature on literacy (citations) suggested that only 25% of adults in the U.S. population believe that plants make their own food.

• Visitors understood plants' role in atmospheric gaseous exchange: In exit interviews, 50% of visitors mentioned plants' role in oxygenating the atmosphere. While they may have known this before viewing the exhibition, 30% quoted specific plants listed in the exhibition as highly efficient oxygenators, showing that they had remembered detailed information.

• Visitors became more aware that plants tie up carbon: Concept maps created by adult visitors on the topic of "ways plants help us" were more likely to include carbon sequestration after visitors had gone through the exhibition than before. Specifically, 20% of adults added this feature to their own concept maps after seeing the exhibition. (There was no control group for this finding.)

• Visitors learned that most of a tree's material comes from carbon dioxide in the air, not from the soil: 25% of visitors who had seen the exhibition correctly identified "the air" as the source of most of the weight of a tree. This number was significantly higher than the 10% from a comparison group who had not seen the exhibition. While answering this question, 15% of adults explicitly mentioned that this fact had surprised them.

• Visitors already knew that plants move: In a card-sorting task carried out by two groups of visitors (those who had and had not seen the exhibition), there was no significant difference in

the number of visitors who correctly identified movement as a behavior of plants (80% versus 83%). However, discussion with visitors suggested that this might have been because the question was misleading: visitors' most common example of plants moving was because of wind, rather than self-initiated movement.

Evidence of impact on attitude (from hypothetical results):

• Overall appreciation: in describing what the exhibition was about, 60% of visitors recognized that it was to help people to appreciate plants.

• Visitors appreciated the sophistication of plants: In exit interviews, 20% of visitors mentioned that plants were more adaptable / flexible than they had realized. Behavioral observations also showed that the time-lapse videos were particularly effective in this way: visitors frequently commented on the cleverness or capacities of plants while watching them (35% of observed groups). A few even described the plants as "smart."

• Visitors appreciated the environmental contribution of plants: In exit interviews, 70% of visitors talked about the environmental role of plants, and 35% specifically mentioned that this was a valuable or even vital contribution. 25% mentioned concern about the fate of the earth's jungles, both as food for animals and as planetary storage for carbon. No comparable data was found from other sources.

• Visitors sustained their appreciation over time: Email follow-up interviews with visitors three months after their visit provided some evidence that visitors had sustained these attitudes over time. Specifically, 50% recalled the purpose of the exhibition as helping people appreciate plants, and this was not significantly different from the 60% found during exit interviews. 75% said they had discussed plants with friends or family since the exhibition and the majority of these mentioned a sense of appreciation for plants or concern about their decline as part of the conversation.

Notes about reporting:

- Hypothetically, suppose that 20% of visitors in exit interviews said they wanted to go home and plant more trees in their yards. This finding would be identified as an unanticipated outcome.

B) Hypothetical Exhibition: "Robotics for all"

Project goals as stated in grant proposal: The project will provide members of the public, especially girls and women, with access to a range of highly engaging experiences in contemporary robotics. The exhibition will provide an introduction for those entirely unfamiliar with the field, will share recent applications in related fields, and will provide a taste of career opportunities in nearby Silicon Valley for those interested in exploring steps beyond the museum experience, including attending workshops, joining local robotics clubs, and considering robotics-related career options.

Thus, the project's overall impact would fall entirely within the category of "engagement."

1) Engagement: Visitors, especially women and girls, will have highly engaging and enjoyable experiences with robots, and will want to explore options for extending their experience.

Table 5-1. Summary of Impacts of *Plants* Exhibition

Impact	Impact Category	Audience Objectives	Evidence
Visitors will understand key aspects of the chemistry, behaviors, and ecology of plants.	Knowledge	Visitors leaving the exhibition will know that plants create their own food.	Card-sort activities showed that visitors leaving the exhibition knew that plants create their own food (60% versus 20% in the U.S. adult population).
		Visitors will know that plants are beneficial in that they generate oxygen and take in carbon.	Exit interviews showed that 50% of visitors learned or were reminded that plants provide oxygen to the atmosphere. Concept maps drawn by adult visitors before and after using the exhibition showed that 20% became more aware of carbon sequestration as a way that plants help us.
		Visitors will know that much of the weight of plants comes from carbon in the air.	After seeing the exhibition, significantly more adults knew that most of the weight of a tree comes from the air (25% versus 10% of the control group).
		Visitors will know that plants may behave in complex ways.	Card-sort activities showed that 80% of visitors knew that plants move; this was a reminder rather than new knowledge (compared with 83% of control group).
Visitors will appreciate plants, both in terms of their sophistication as organisms and their vital role on planet earth.	Attitude	Visitors will recognize that one purpose of the exhibition was to help people appreciate plants.	Exit interviews showed that 60% of visitors understood that the exhibition's purpose was to help people appreciate plants.
		Visitors leaving the exhibition will articulate their appreciation of plants as complex and important members of living systems.	During the exit interviews, 20% of visitors spontaneously mentioned that plants were more adaptable than they had realized, and 70% of visitors talked about the environmental role of plants.
		Visitors' appreciation of plants will last over time.	During follow-up interviews 3 months later, 50% of visitors mentioned at least one property of plants that they saw as valuable or impressive. 75% of visitors self-reported sharing their attitude with friends or family after their visit.

Relevant findings would then serve as evidence for these impacts.

Evidence of impact on engagement (from hypothetical results):

- Visitors were engaged at the staffed exhibit elements for very long holding times: A tracking study showed that visitors spent a median of 4.5 minutes at each robotics station. This compares with approximately 1 minute at typical hands-on exhibit elements (visitor studies citations) and with 1-10 minutes at open-ended staff-assisted floor experiences (visitor studies citations). Girls and women, the target audience, spent a median time of 3 minutes at each exhibit element they stopped at, not as high as boys and men (3.5 minutes) but still longer than typically reported times from the visitor studies literature. Observations showed that males' times were typically longer because they were the initiators of activity, and only when they had participated in the race-related aspects of the exhibits were they willing to let the girls take a turn.

- Exhibit elements were equally attractive to males and females: A summative tracking study showed no significant differences between males and females in terms of the average number of exhibit elements they stopped at (18 and 16 respectively, out of 35). No data were available to compare this with gender ratios at other robotics exhibitions, though data from robotics clubs shows that this is broadly perceived as a male-dominated activity, with females typically constituting less than 25% of active members (citations from online demographics of clubs).

- Visitors reported a mix of emotional responses: Visitors were asked to complete an assessment of their positive and negative emotional responses after viewing the exhibition, and this showed they were not as happy but significantly more inspired than visitors leaving a neighboring exhibition on butterflies (scoring 3.4 versus 4.2, and 5.8 versus 4.3, respectively).

- Girl on a school field trip group was inspired by the exhibition: One teacher who had visited the museum on a field trip self-reported that her students had particularly enjoyed the exhibition. While this was not data that had been systematically collected, the teacher reported that one of her female students had decided to do a project on robotics for the school science fair, based on her experience at the exhibition. It is only one anecdote, but might suggest a line of inquiry worth pursuing as a new indicator of impact.

- Visitors signed up for the workshops: A total of 75 visitors signed up for workshops in the series that accompanied the exhibition. Of these, 45% were female. The number of participants who actually arrived at the workshops showed some attrition (20%), with no significant difference between male and female percentages of attrition.

- The career booths were not highly engaging: Of the *n* robotics-related careers that were represented in the careers section of the exhibition, the only one that received sustained interest (from all visitors) was the one from the MIT Media Lab that featured a robot that displayed emotions. Visitors were very engaged by their interactions with this robot, but they did not frequently follow up with questions for the staff. Only 40 brochures about robotics careers were taken by visitors (an estimated 2% of those who had the opportunity), and the greater

majority of questions asked of the career staffers were about the robots rather than about the career paths to creating them.

Table 5-2. Summary of Impacts of *Robotics* Exhibition

Impact	Impact Category	Audience Objective	Evidence
Visitors, especially women and girls, will have engaging, enjoyable experiences with robots, and will want to explore options for extending their experience.	Engagement	Visitors will be engaged by the exhibits and will spend extended times at the robotics stations.	A tracking study showed that visitors spent a median of 4.5 minutes at each robotics station. The target audience of girls and women spent a median time of 4.0 minutes. Both sets of times are significantly longer than the 1 minute typical of hands-on exhibit elements.
		Girls and women will be engaged by the exhibits as much as boys and men.	The tracking study showed that exhibit elements were equally attractive to males and females (while robotics clubs typically report less than 25% of their active members are female.)
		Visitors will enjoy their experience.	An assessment of positive and negative affect (PANAS) showed that visitors leaving the robotics exhibition were not as happy but more inspired than visitors leaving a neighboring exhibition on butterflies. One teacher self-reported the exhibition had inspired a female student to do a robotics project for the science fair.
		Many visitors will want to extend their experience by attending workshops or spending time at the career booths.	A total of 75 visitors signed up for workshops in the series that accompanied the exhibition, and 60 participated, thus extending their experience. Girls and women constituted 45% of the workshop attendees. However, the career booths were not very engaging, and the brochures about robotics careers were only taken by an estimated 2% of visitors in the area.

C) Hypothetical Exhibition: "After the flush, where does it go?"

Project goals as stated in grant proposal: We aim to inform visitors, especially children, about the plumbing systems and sewerage systems of cities, including various systems for treating waste water. We will present the implications of using drains and toilets as ways to dispose of various unwanted materials, especially toxic liquids and living organisms. We will recommend alternatives to waste water disposal where appropriate, and will encourage visitors to use these alternatives by providing information and resources specific to their home locations.

Thus, overall impacts would be:

1) Knowledge: Visitors, especially children, will understand that drains and toilets lead to open oceans or river systems, with limited filtration and treatment occurring before release into the environment. They will learn what kinds of items should not be washed down a drain or toilet.

2) Behavior: Visitors who use waste water systems indiscriminately will change their behavior to some degree. Specifically, they will be less likely to use toilets and drains to dispose of unwanted aquatic plants or pets, and more likely to reduce the amount of detergent in their washing activities.

Relevant findings would then serve as evidence for these impacts.

Evidence of impact on knowledge (from hypothetical results):

• Visitors understood that waste water is ultimately released: When asked to draw a picture of what happens to the water that goes down drains and toilets, 80% of adults and 60% of children drew pictures that correctly showed water being ultimately released into open water systems such as oceans or rivers. While there was no pre-test or control group for this finding, a front-end study conducted at the Waste-Water Museum showed that 50% of visitors interviewed about waste water said they did not know where it ended up.

• Visitors understood the impact of waste water systems: Exit interviews showed that most visitors (70% of adults and 60% of children) had understood the big idea of the exhibition: viz.., that waste water systems can release undesirable things into rivers and oceans.

• Visitors learned some specific facts mentioned in the exhibition: Exit interviews showed that 25% of visitors self-reported that they had not realized that treated sewerage is released within a mile of public beaches, or that wetlands can be more effective than human-made treatment facilities.

• Visitors connected the exhibits to their own lives: Exit interviews showed that visitors made clear connections between the exhibition and their own lives. A majority of families (70%) reported discussing their own washing and flushing practices while they were in the exhibition.

Table 5-3. Summary of Impacts of *Waste Water* Exhibition

Impact	Impact Category	Audience Objective	Evidence
Visitors, especially children, will understand that waste water systems lead to open water systems, and that some items should not be washed down a drain or toilet.	Knowledge	Visitors will understand the exhibition's main idea: that waste water systems can release undesirable things into rivers and oceans.	Drawings by visitors leaving the exhibition showed that 80% of adults and 60% of children knew that the contents of drains and toilets are eventually released into open water (compared with 50% of visitors in a front-end study). Exit interviews showed that 70% of adults and 60% of children understood the exhibition's main idea: that waste water systems can release undesirable things into rivers and oceans. Follow-up interviews showed that visitors still remembered this idea six months later (60% of adults and 50% of children).
		Visitors will learn about a range of approaches for waste-water treatment, and their implications for the environment.	During exit interviews, 25% of visitors self-reported that they had not previously realized that treated sewage is released within a mile of public beaches, or that wetlands can be more effective than human-made treatment facilities. However, the exhibition's simulated filtration system gave 50% of visitors a misleading sense that releasing the invasive species constituted a successful action.
		Visitors will connect the exhibition's main idea to their own waste water behaviors.	Most families (70%) self-reported discussing their own washing and flushing practices while in the exhibition.
Visitors who use waste water systems indiscrimin-ately will lessen their inappropriate use of these systems.	Behavior	Visitors who previously used waste water systems indiscriminately will change their definitions of acceptable behavior.	Visitors who had seen the exhibition were significantly less likely to find it acceptable to dispose of unwanted medications via waste water systems (compared with a non-randomized control group, quasi-experimental design).
		In their own lives, visitors will be less likely to use waste water systems indiscriminately.	During exit interviews, 30% of adults expressed an intention to change, or consider changing, their behaviors related to waste water disposal. In follow-up internet surveys conducted six months later, 30% of visitors said they had discussed disposal mechanisms with others in their lives, and 10% had taken at least one action to change their behavior.

- Some visitors, especially children, did not understand the threat of invasive species: Tracking and timing showed that visitors were particularly engaged by the simulated filtration system that demonstrated how microscopic organisms could still pass through and be released. However, 50% of visitors (mostly children) seemed not to realize the dangers of invasive species in an ecosystem, and were therefore pleased, rather than concerned, that the organisms were released into open water.

- Visitors remembered the main point: In follow-up phone calls six months after the visit, 60% of adults and 50% of children remembered the main idea of the exhibition. They also remembered particular exhibits in detail, most often the simulation of the filtration system.

Evidence of impact on behavior (from hypothetical results):

- Visitors showed some evidence of change in their definitions of acceptable behavior: A quasi-experimental study compared visitors who had just seen the exhibition with visitors who had not seen it, but said they planned to. The group who had seen the exhibition was slightly less likely to agree with statements such as "It is ok to put down the drain anything that could be healthily consumed by a person, such as medicine" (agreement levels of 3.1 and 3.6 respectively, on a 5-point scale).

- Visitors expressed an intention to change their behaviors: In exit interviews, 30% of adults commented that they intended to change, or consider changing, their behaviors in terms of waste water disposal. 15% could name at least one alternative disposal method that they had found personally compelling.

- A few visitors actually changed their behaviors, and many discussed it: Follow-up internet surveys conducted six months later showed that 10% of the visitors reported taking at least one action as a result of their visit to the exhibition. More commonly, 30% of visitors said they had discussed alternative disposal mechanisms for some of their wastes. 20% reported that their children had expressed concern and reminded them of their experience on some later occasion, such as cleaning out their home aquarium. Note: as with all electronic surveys, it is important to describe the return rates and discuss possible sample biases.

D) Hypothetical Exhibition: "Why do you say that? Articulating evidence in everyday life"

Project goals as stated in grant proposal: We want to encourage visitors to identify and state the warrants for their claims, an essential practice in science that is often overlooked in everyday life.

Thus, impact would be:

> *Skills*: Visitors will learn and/or practice the skill of articulating evidence or reasons for their claims.

Relevant findings would then become indicators for these impacts:

Evidence of impact on skills (from hypothetical results):

- Visitors did articulate evidence or reasoning to back up their claims. Recordings of visitors at a representative subset of exhibits showed that visitors did provide warrants for their claims. Specifically, 60% of visitor groups who used the targeted exhibits stated 2 or more warrants for their claims. While we could not locate any comparable data in the literature, this was significantly more than the number of groups using a representative subset of exhibits in a neighboring exhibition about gemstones (where only 40% of groups stated 2 or more warrants for their claims). Particularly successful for facilitating this skill were the communal whiteboards that encouraged visitors to share their reasoning with other groups; these were used by 40% of family groups and 35% of field trip groups, on average.

- Visitors were not aware that they were supporting their claims. Even though visitors were in fact supporting their claims exit interviews showed that the majority of visitors (70%) were not aware of this. The most common description of the exhibit collection was that it was a place where there were fewer labels, and where one needed to figure out what to do. Several visitors (10%) reported being fatigued by the effort of using the exhibits.

Table 5-4. Summary of Impacts of *Evidence* Exhibition

Impact	Impact Category	Audience Objective	Evidence
Visitors will learn and/or practice the skill of articulating evidence or reasons for their claims.	Skills	While using the exhibits, visitors will be able to state evidence for their claims.	Audio recordings showed that 60% of visitor groups did state 2 or more warrants for their claims while using the exhibits (compared with 40% of visitors groups using neighboring exhibits). However, interviews showed that most visitors (70%) were not aware that they were supporting their claims, many thinking instead that they were "figuring out what to do" at the exhibits.
		Visitors will also state evidence for their claims in their lives beyond the museum visit.	Follow-up email interviews suggested that visitors only rarely thought they had used the skills beyond their visit (2 out of 25 visitors).
		The exhibits will support evidence-based reasoning by visitors with a range of ability levels.	Observations and interviews with two field trip groups of children with moderate cognitive disabilities suggested that the exhibits helped these children make claims, rather than provide evidence for them.

- Visitors rarely reported using their skills beyond their visit. Follow-up email interviews with a small sample of visitors (N=25) showed that in only two cases, visitors felt they had used the skills in their daily lives. Specifically, the examples given were: doing a more careful report for a science fair project, and asking for evidence when arguing with a sibling about what constitutes adequate sleep.

- Children with cognitive disabilities found the exhibition useful for making claims: Two field trip groups of children with moderate cognitive disabilities were observed using the exhibition and interviewed afterwards using naturalistic methods. The data suggested that these children did not spend much time offering evidence for their claims, but did engage in sharing and refining claims about the exhibits they were using, a related and important prerequisite skill.

Issues common to evaluating exhibitions

When evaluating exhibitions, some important considerations commonly arise.

REALISTIC EXPECTATIONS

Most visitors will spend only a few hours in total at a museum, zoo, botanical garden, or aquarium, and a significant part of that time will be taken up with activities such as navigation, eating, visiting restrooms, enjoying social conversations, and keeping track of other members of the group. Serrell (1998) found that a single interpretive exhibition typically holds visitors for a maximum of about 20 minutes. Because this constitutes such a brief engagement with the project deliverable, it may be unrealistic to expect a single visit to a single exhibition to have large learning impacts in any category. It is especially unrealistic to expect a single exhibition experience to affect students' school test scores, which depend on a multitude of factors beyond the control of the ISE project team, and which are usually designed for a different purpose (viz. assessing concept learning and factual memory of very specific scientific content after weeks or months of teaching). When reporting evidence of impact, it may be possible to use several small aspects of an exhibition experience to point toward a larger impact, especially if the project is based on a cumulative model of learning, with experiences designed to build on or reinforce each other.

USING NON-TRADITIONAL ASSESSMENTS TO MATCH VISITORS' INTENTIONS AND ACTIONS

Public audiences use exhibitions as a form of leisure experience, not to pass an examination or cover a curriculum. Because of this, school-based tests of conceptual understanding or factual memory are unlikely to capture the kind of learning that has occurred. The most valid assessments are likely to be brief (or relatively invisible), non-intimidating, and open-ended, allowing visitors to share their chosen journeys and connections with the materials presented. It

may be helpful to look for evidence of impact on the visual, spatial, and kinesthetic abilities of visitors, in addition to the more commonly assessed verbal and logical abilities. And impact does not require that learners do something entirely new – it can involve reminding, retelling, practicing, or connecting to the mundane, as families explore and reinforce their knowledge and identities. Learning may be apparent only in a certain physical or social situation, or may be impossible for the learner to even talk about. Overall, exhibition assessments may need to be particularly flexible and responsive to visitors' agendas and actions, to capture the kinds of learning impact that may be occurring.

TYPICAL IMPACTS OF EXHIBITIONS

In terms of the categories of impact in this guidebook, exhibitions most often show evidence of engagement, interest, and emotion, and some forms of knowledge (especially personal connections and associations, reinforcement of previous knowledge, and making inferences). Measurable changes in attitudes, behaviors, or deep conceptual understanding, though certainly possible, are rarer.

VISITORS' MOVEMENTS AS EVIDENCE OF ENGAGEMENT

One type of study method ("tracking and timing") has been developed simply to try to characterize visitors' movements and actions as they move through an exhibition, and such studies can provide useful evidence of engagement, especially if compared with studies of neighboring or comparable exhibitions.

VISITORS' INTERPRETATIONS AS EVIDENCE OF KNOWLEDGE / UNDERSTANDING

Because each visitor only samples an exhibition (generally fewer than half the elements), it is a significant mental feat for visitors to be able to synthesize their experiences into a main idea that they can state in their own words, especially if they have not been cued (i.e., previously warned) that they will be interviewed.

THE DIFFICULTIES OF EXPERIMENTS

The "free-choice" nature of exhibitions makes experimental studies particularly challenging to conduct without changing the fundamental nature of the experience. There are usually trade-offs between the rigor of an experimental design and the authenticity of the learning experience being

studied. We recommend that any experimental study of exhibitions should be thought through well in advance with the help of an experienced evaluator.

STRETCHING TIMESCALES OF STUDY

A recent trend in exhibition study has been the extension of the timescales of interest in both directions: shorter and longer. On the shorter side, researchers and evaluators have been studying visitors' movements, gestures, and short snippets of conversation, to better understand what they do and how they interact with each other and with the exhibition. On the longer side, researchers have been studying what visitors remember about experiences they had weeks, months, or even years earlier. Most of these studies are in-depth research projects rather than exhibition evaluations. However, even for evaluations, it may be feasible to conduct follow-up studies a few months after a museum visit, to study impact over time.

REFERENCES

Serrell, B. (1998). *Paying attention: Visitors and museum exhibitions.* Washington, DC: American Association of Museums.

CHAPTER 6 EVALUATING MASS MEDIA

Barbara N. Flagg

Mass media provide a rich source of science information and news for adults, youth and children. A 2006 telephone survey reveals that 41% of American adults "get most of their science news and information" from television, 14% from newspapers or magazines, and 4% from radio (Pew Internet and American Life Project, 2006). In addition, in areas with giant screen theaters, one-quarter of the general population is reported to visit a giant screen theater in a year (Kennedy, 2004). In an effort to capture the public's eyes and ears, the NSF funds mass media projects in various formats: television series and single shows; long and short format radio series; 2D and 3D giant screen films; planetarium digital dome programs; as well as companion print books and children's magazines. Mass media, as defined in this chapter, involve one-way communications to an extremely large demographically diverse audience.

Evaluation associated with NSF-sponsored mass media progresses through various stages, from front-end analyses that gather baseline information about target audiences to formative evaluation that tests treatments, storyboards, and/or rough-cuts with audiences, culminating in summative evaluation that assesses the impact of the media deliverables on the public (Flagg, 1990). Summative evaluation is also not necessarily the end-of-the-line activity for a project, since television and radio shows can be on-going series; thus, the summative evaluation findings for one season can play the role of formative evaluation by modifying how the series develops in subsequent seasons.

Typically, NSF-sponsored mass media projects also produce other deliverables including, for example, interactive web sites and outreach activities or curriculum materials for informal and formal venues. Summative evaluation of other deliverables is considered in other chapters: see chapter 7 for outreach with youth and community, chapter 8 for interactive technology issues, and chapter 10 for evaluation focused on the added value or synergy of *multiple* deliverables. Through a purely fictional example, this chapter reviews the process of specifying the intended impacts, audience objectives, research designs and data collection techniques associated with summative evaluation of mass media. Meet the fictional production house: MM Communications, Inc, which is organizing a mass media project about global warming and climate change (a topic chosen only to serve as an example for this chapter). Among the many activities of the project, mass media products are planned including a television series, a giant screen film and radio shorts.

Impacts. To guide the design of the media deliverables, the MM Communications staff first discuss with advisory scientists, outreach educators, and the formative and summative evaluators what might be the expected impacts of the project. Broad impact categories, described previously in Part I, include: (1) awareness, knowledge or understanding; (2) engagement or interest; (3) attitude; (4) behavior; (5) skills and (6) other. All of the mass media deliverables of

our example project are intended to deal with global warming and climate change as content, but each medium is focused on one impact category for illustration purposes only:

- A two-hour television series intends to increase knowledge and understanding of global warming and climate change.
- A 40-minute giant screen film will modify public audience's stereotypical attitudes towards scientists.
- Ten 2-minute radio shows will encourage listeners to change their behavior to reduce their carbon footprint.

Each media product will be differentially designed so as to ensure that its impact is attainable by the public audience. However, to guide production and to evaluate the effectiveness of the project, each impact statement is elaborated into statements of objectives that describe how the audience will be different after exposure to the media. What will the audience know, understand, believe, or do differently after seeing and listening to the mass media products? How will we know that changes have occurred? These questions are illustrated, plausibly if not comprehensively, for each of our three mass media examples.

TELEVISION SERIES

The two-hour PBS television series planned by our fictitious MM Communications is intended to increase adult viewers' knowledge and understanding of global warming and climate change. Drawing on a front-end review that identifies public conceptions and misconceptions related to this content area, the team aims a portion of the series toward explaining the dynamics of our climate system. They generate audience objectives to guide script writing, animation development and eventually evaluation. Here are three learning objectives related to climate system dynamics:

1. Viewers will understand that the earth's surface temperature is mainly governed by the balance of the sun's energy heating the earth and the energy the earth radiates back into space.
2. Viewers will learn that some radiated energy is absorbed by greenhouse gases in the atmosphere, giving us a livable temperature.
3. Viewers will recognize the role of inertia in the climate system such that changes in CO_2 emissions do not immediately impact global temperature.

Upon completion of the television series, a summative evaluation is implemented in which the impact on knowledge and understanding is assessed by measuring achievement of the above three audience objectives, among others. A wide variety of evaluation designs are applicable to mass media, each with advantages and disadvantages (Gunter, 2000). The summative evaluator works with the production team to design an evaluation study that can effectively assess a variety of impacts and audience objectives, is feasible within the permitted timeframe, and realistic given budgetary constraints. To demonstrate that changes in understanding are attributed to exposure to the television series and not some other experience or factors, our example evaluator

and project team chooses to implement a type of experimental study (Campbell and Stanley, 1966). Our experiment controls who is exposed to the TV series and when as well as who gets what measurements. We randomly assign potential viewers to two groups: a group that sees the TV series (treatment) and a control or comparison group that does not see the series. Random assignment means that each viewer has an equal chance of being put in one group or the other. Viewers can watch the broadcast in real time or be given videos to watch at home at convenient times, but care should be taken that the control group does not have access to the shows. If it is thought that the experience of viewing or the exposure to the media format itself is a possible factor in impact, then the control group could view a different TV series to provide an equivalent media experience. Alternatively, comparison groups might include one that watches the TV series and is also exposed to some additional deliverable like a web site, or a group that watches the series and has a community outreach experience like a forum discussion or science café. (See Chapter 10 for more discussion of evaluation of deliverable combinations.)

The rationale of including a control group is to have a random sample that experiences everything the viewing group might experience (except viewing the show), thereby controlling for effects that confound the measured effect of the TV series. For example, during the viewing period, there may be relevant real-world events occurring (e.g., climate change news announcements); there may be changes in respondents (e.g., decreasing interest in completing the viewing task over time); and data collection procedures may influence participants (e.g., previewing measures may cue participants' to view for certain information). In fact, in our experiment, we opt to omit the pretest because of cueing effects and trust that random assignment assures initial equivalence of comparison groups. In other situations a pretest can be unlikely to sensitize or cue viewers; for example, the summative evaluation of a television series that teaches children the process of science inquiry could assess viewers and non-viewers with hands-on science tasks both before and after viewing a series of shows.

Control or comparison groups help rule out alternative explanations for changes in our audience's knowledge and understanding of climate dynamics. Logical argument also can be used to rule out these alternative explanations in some situations. For example, because of the limitation of time in a focus-group viewing of a one-hour television show, there are few other explanations for changes in outcome measures (Morgan and Krueger, 1998). It's unlikely that the respondents themselves will change physically or leave the short session or that outside events intervene to influence viewers, but comments and actions within the focus group can influence results. On the other hand, a focus-group viewing puts limitations on generalizing the results. Viewing with a pretest in a group setting does not reflect an at-home viewing situation and raises a question of how well one can generalize the results to the natural viewing environment. Every evaluation design has its strengths and weaknesses in its varying applications, and these need to be considered in choosing a design and in interpretation of the results.

Our cognitive audience objectives as stated previously for the TV series use the terms "understand," "learn," and "recognize." In order to measure impact of the series on viewers, the evaluation must develop operational definitions of these terms. What do we mean when we say a viewer will "understand" climate system dynamics and how will we measure that

"understanding"? Table 6-1 presents evidence for *one* objective of our fictional evaluation of a television series. In this table, we have one audience objective with multiple pieces of evidence, all assessing different aspects of the learning objective. [Please don't try these at home because they have not been pilot-tested!]

Table 6-1: Impact of Television Series on Knowledge and Understanding

Impact:	**Impact Category**	**Audience Objective**	**Evidence**
The TV series will increase adult viewers' knowledge and understanding of global warming and climate change.	Awareness, knowledge or understanding	Viewers will understand that the earth's surface temperature is mainly governed by the balance of sun's energy heating the earth and the energy that earth radiates back into space.	Asked to draw a picture that explains their understanding of global warming, adults who viewed the TV series included more elements involved in the climate system than those who did not view the TV series. Those who performed better also rated the TV series as more appealing. In response to a face-to-face interview, adult viewers of the television series were significantly better able than non-viewers to describe the action or movement of energy in earth's climate system and used appropriate scientific terminology. However, the results showed an interaction with gender whereby the male non-viewers demonstrated a better understanding than the female viewers. A set of multiple-choice questions assessing verbal factual knowledge about the mechanism of earth's climate system and global warming revealed that adults who viewed the TV series scored significantly higher than those who did not view the TV series.

In Table 6-1, you will note a gender difference under Evidence – "male non-viewers demonstrated a better understanding than the female viewers." This is a negative finding for our project, because the TV series intended to help close the gender gap. In addition to assessing "what" the impact of the series is on understanding, a summative evaluation can also include methods to explore "how" and "why" the series succeeds or fails in achieving its intended impacts. Such lessons-learned contribute to the growth of the field of informal science education. In our gender difference, for example, further interviewing revealed that adult female viewers, who typically have difficulties with visual-spatial relations, misinterpreted the three-dimensional diagrams describing the climate system.

The giant screen film planned by MM Communications follows the trials and tribulations of climate research scientists as they collect data in the coldest, driest and windiest continent, Antarctica. In our film example, the 40 minutes of panoramic views and an emotionally-grabbing story of the scientists' adventures with nature are intended to modify stereotypical attitudes towards scientists. A front-end study for this project reveals that most museum visitors imagine that scientists who study climate change typically sit at computers all day creating complex and unreal models. Our film's audience objectives are to revise viewers' stereotype of scientists to include the ideas that scientists work outside a laboratory, that scientists' work can be dangerous and life-threatening, and that scientists are persistent and dedicated.
To measure impact on attitude, our summative evaluation uses a quasi-experimental study. An experiment study, as described for our TV series, would be the best method for establishing causality, but it is not the only method nor the method that will suit many natural field situations in which such control is impossible. For these situations, an alternative is a quasi-experimental design, of which there are several varieties (Cook and Campbell, 1979).

A quasi-experimental design is effective for making causal inferences about a giant screen film or planetarium presentation in a natural theater setting. One type of quasi-experimental evaluation design involves selecting random viewers in the pre-show line or ticket line to participate in one of two groups: one sample to complete the assessment procedure prior to viewing the film and a different sample to complete the procedure after viewing. There are other quasi-experimental designs, but several characteristics of the population and treatment (i.e., a film) leads MM Communications' evaluator and production team to the decision to use this particular version.

First, the population we wish to generalize are self-selected museum visitors whose intention is to view a film. The best comparison group for our viewers is a sample of museum visitors who intend to view the film but have not yet done so; there are no comparable museum visitors from whom the treatment (the film) could be withheld. Second, we must assume that museum visitors already have an attitude of some sort towards scientists, thus it is important to include a pretest that establishes what the audience's attitude is prior to seeing the film. However, interviewing just prior to viewing may sensitize an audience to our desired outcomes and affect their viewing and the posttest results, so one group is interviewed prior to seeing the film and another randomly selected group interviewed only after seeing the film. Third, each group is selected at random so that they are representative of the same population, and random selection is logistically simple in the theater environment where the audience lines up or buys tickets before show time.

In our example, the two audience groups are interviewed individually about their attitude toward scientists. Again, we need to define what we mean by "attitude" in order to measure it. There are numerous theories about attitude change to guide the operationalization of the outcome objective. Attitudes are unlikely to change in 40 minutes so we will focus on assessing precursor or prerequisite behaviors to attitude change. For our example, the team posits that to modify scientist stereotypes requires that the audience finds our scientist characters appealing, that the viewers think their actions are believable, that the audience engages in and recalls the scientists' actions and identifies discrepancies between their own pre-viewing idea of what a scientist is like

and what they see and hear in the film. The qualitative data of audience interviews are coded by researchers unaware of group-designation, and the results are compared across pre-viewing and post-viewing groups to look at significant differences. Table 6-2 presents our audience objectives and evidence for our fictional giant screen film.

Again, a giant screen film may not be a stand-alone product, and a quasi-experimental evaluation design could include comparison groups that have been variously exposed to outreach activities, an interactive web site, companion exhibits, a guest scientist presentation, and so forth. Chapter 10 considers evaluation of deliverable combinations.

Table 6-2. Impact of Giant Screen Film on Attitude

Impact	Impact Category	Audience Objectives	Evidence
The giant screen film will modify public attitude toward scientists.	Attitude	Viewers will like the film's scientists and believe that their actions are credible. [Likability and credibility of the attitude object (scientists in our case) are important precursors to persuading an audience to change their attitudes.]	Interview rating questions revealed that eight out of ten viewers rated the film's scientist characters as "very appealing" and "very believable."
		Viewers will remember the scientists' actions in the film. [Comprehension of the reality of scientists' fieldwork is a precursor to attitude change.]	Open-ended interview questions showed that the viewing audience recalled the scientists' data collection activities in detail.
		Viewing the film will lead the audience to describe the activities and personalities of scientists differently from the scientist stereotypes of non-viewers.	In response to open-ended interview questions, viewers as compared with non-viewers described climate scientists using different descriptors and categories. Viewers' descriptions included significantly more frequent references to collecting data in the field rather than doing work in a lab, included more references to scientists risking their lives in data collection activity, and described scientists more frequently as tenacious or persistent.

RADIO

The final component in MM Communications' fictitious mass media project is a set of ten two-minute radio shorts intended to change audience behavior. The impact is for listeners to modify their behavior in ways that will reduce their carbon footprint. Each radio short suggests a particular behavior to decrease the production of greenhouse gases; for example, replace

incandescent light bulbs with compact fluorescents, turn off lights and electrical appliances when not in use, and use a clothesline instead of a dryer.

The short duration of the radio pieces presents a challenge to assess causal impact. An experimental or quasi-experimental design could be used. We might, for example, email daily MP3 files of the radio shows to a treatment group and not to a control group. But what if we want to generalize to the natural listening experience of hearing a 90-second piece by chance in a commuter's day? We could use instead a naturalistic or descriptive survey study, in which we have no control over exposure or intensity of exposure to the treatment (in this example, our short radio shows) (Fink, 2003). Some people hear all shows that are broadcast, some people hear a few, some hear none, and some may hear repeats. There is self-selection into different listening groups.

Our fictional series is broadcast several times a day for two weeks on a commercial radio station, which has a mailing list of listeners that it has acquired through running various on-air contests. To a random sample of the list, we mail a short written questionnaire with a return envelope and a guilt-provoking one-dollar bill incentive, and our return rate is a respectable 60%. Assuming that our respondents have not misclassified themselves, we can compare listeners and non-listeners on an operational measurement of our outcome objective; that is, whether or not respondents report doing the behaviors suggested by the radio shows. However, the use of self-selected comparison groups increases the possibility of some spurious variable biasing the results. For example, it could be that the shorts are aired during a long-format environmentally-oriented show. Those who voluntarily tune into the longer show may be more interested in the environment than non-listeners and thus prone to modify their greenhouse gas-producing behavior with or without hearing our 90-second segments. In our example, however, our radio shorts are integrated into the local news broadcast. Table 6-3 presents our audience objective and evidence for the behavioral impact of our fictional series of radio shorts.

Table 6-3. Impact of Radio Shorts on Behavior

Impact	Impact Category	Audience Objective	Evidence
Listeners will change their behavior or acquire new behaviors that decrease their production of greenhouse gases.	Behavior	Listeners will replace their incandescent lights with compact fluorescents, will increase their efforts to turn off lights in unused areas, will turn down their heating thermostat or turn up their air conditioning by 2 degrees, will utilize a clothesline, and [6 more behaviors as suggested in the radio shorts].	Respondents to the mailed questionnaire were classified into three groups according to their self-report: high listener (heard 5 or more shows), low listener (heard 4 or fewer shows) or non-listener (heard no shows). High listeners indicated in response to a behavior check-off list that they had recently purchased compact fluorescent lights, whereas low listeners and non-listeners were significantly less likely to check off this behavior. None of the other listed behaviors showed a relationship with listening to the radio shorts.

In addition to gathering evidence regarding the planned impact and objectives, evaluation designs and methods should be open to uncovering unintended effects, unexpected outcomes, or outcomes peripheral to the intended impact. For example, in our fictional radio series evaluation, an open-ended question discovered that a significant percentage of show listeners had rented the climate change film, *An Inconvenient Truth*, **after** the radio series had aired. The short format of the radio series was not very effective in changing behavior that decreases greenhouse gas production, but it did encourage listeners to reach out for more information – an unintended positive behavioral outcome.

CONCLUDING REMARKS

Through a series of fictional examples, this chapter focuses on a limited set of issues relative to the impact of mass media on public audiences, considering only the impact categories intended to be aggregated across ISE projects. In addition to assessing what impact a deliverable has on interest, knowledge, attitude and behavior, summative evaluations also explore other aspects to interpret *why and how* the effects might have occurred or *not* occurred. Depending upon one's theories of communication processing and media effects, an evaluation might record audience demographics or environmental factors as effect mediators; identify media components that attract and hold visual and auditory attention; measure perceived reality or credibility of content, storyline and portrayals; consider personal relevance or identification with characters; assess density, pace and clarity of media elements; and so forth (Bryant and Thompson, 2002). Summative evaluation is not theory-testing basic research; nonetheless, well-designed individual evaluations across many instances of similar genre can contribute as a group to our knowledge base of how mass media communicates effectively (Fisch, 2004; Flagg, 2005, Spring).

REFERENCES

Bryant, J. and Thompson, S. (2002). *Fundamentals of media effects*. New York: McGraw-Hill.

Campbell, D. T. and Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally & Company.

Cook, T. D. and Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Boston: Houghton Mifflin Co.

Fink, A. (2003). *The Survey Kit, Second Edition*. Thousand Oaks, CA: Sage Publications.

Fisch, S. M. (2004). *Children's learning from educational television*: *Sesame Street and beyond*. Mahwah, NJ: Lawrence Erlbaum Associates.

Flagg, B. N. (1990). *Formative evaluation for educational technologies.* Mahwah, NJ: Lawrence Erlbaum Associates.

Flagg, B. N. (2005, Spring). Beyond entertainment: Educational impact of films and companion materials. *The Big Frame*, 22 (2), 50-56, 66.

Gunter, B. (2000). *Media research methods: Measuring audiences, reactions and impact.* Thousand Oaks, CA: Sage Publications. Mohr, L. B. (1992). *Impact analysis for program evaluation*. Thousand Oaks, CA: Sage Publications.

Kennedy, M. K. (2004, Winter). GSTA's 2003 worldwide viewer and nonviewer research programs: Key results and how to use them. *The Big Frame*, 40 – 59.

Morgan, D. L. and Krueger, R. A. (1998). *The focus group kit: Volumes 1-6.* Thousand Oaks, CA: Sage Publications.

Pew Internet and American Life Project. (2006, November). *Exploratorium survey: Final Topline.* Retrieved March 5, 2007, from http://www.pewinternet.org/pdfs/PIP_Exploratorium_Science_Topline.pdf

CHAPTER 7 EVALUATING YOUTH AND COMMUNITY PROGRAMS

Patricia B. Campbell

INTRODUCTION

Informal youth and community programs funded under the National Science Foundation's (NSF) Informal Science Education program fall into a variety of often overlapping areas, each of which has its own evaluation challenges. These areas include:

- Program and materials development and testing. These efforts often focus on young people after school or in other out-of-school experiences. They often target groups under-represented in the sciences, including girls of all races/ethnicities; African American, Hispanic, and American Indian boys; and young people with disabilities.

- Developing and testing the effectiveness of models for training populations who are doing informal science education activities with others. They include after school leaders, teachers, community group members, parents, and older youth. Training of trainer models should also be included in this area where work is done with universities, museums, and other science organizations who then work with after-school leaders, community groups, or other groups who will be involved in informal science education with youth and the communities.

- Developing and testing the effectiveness of different types of collaborations to build capacity among partners in youth and community informal science education efforts at all levels and to learn more about ways of increasing the sustainability of informal science education efforts.

- Research on informal learning for youth and/or in community contexts.

In addition, many informal science media and technology projects do outreach involving after school programs and other out of school venues. In many cases, this outreach fits under informal science youth and community programming.

In this chapter, we examine the public and professional audience impact categories and look at how different impacts specific to youth and community programs fit into the impact categories. In addition, we will discuss some evaluation issues that frequently effect informal youth and community programs.

The core evaluation question is "What works for whom in what context?" The original evaluation question was "What works?" With increased interest in diversity, the question expanded to "What works for whom?" This was a major step forward. What works with

suburban white boys, for example, might not work for urban African American girls, and unless the data are being broken out into these variables, we would never know. There is, however, another piece to the evaluation question for youth and community programs, and that is context. The context in which the program is being conducted, the resources that are available, and the skill of the trainer or leader involved are all important. Summative evaluation of youth and community programs needs to be clear not just about the impact of projects on those involved, but also about the characteristics of the youth and/or adults involved, the context of situation, and any impact that the program has had on the community organizations themselves.

The following impact categories have been defined as the general categories of "what works" for informal science education projects, presented in Chapter 3. They are:

1. Awareness, knowledge, or understanding.
2. Engagement or interest.
3. Attitude.
4. Behavior.
5. Skills.
6. Other (specify).

EXAMPLES OF IMPACT CATEGORIES APPLIED TO SAMPLE EVALUATIONS OF YOUTH AND COMMUNITY PROGRAMS

In the following examples, we will look at some impacts and participant objectives used in different types of youth and community projects, categorize them into the impact categories, and suggest sample measures and designs that can yield evidence.

Example 1: Mobilizing Community: This project's primary goals are to both prepare and mobilize the members of a national membership organization group of college graduates to serve as life-long informal science education organizers for family and community science events or as parent educators. Participant outcomes include changes in (see Table 7-1):

- group chapters' missions and activities to include more of an informal science focus and more informal science activities;
- participating group members' science and math attitudes, especially regarding doing hands-on science activities with children and families;
- the number of informal science education activities conducted in under-represented communities;
- participating group members' involvement in science and math education reform efforts.

Descriptions of the different designs and their strengths and weaknesses can be found in Chapter 4.

The data that are collected using the measures and following the designs becomes the evidence. The following is an example of the evidence, the findings that are used by the project to indicate whether or not given participant objectives were achieved.

Table 7-1 Mobilizing Community Worksheet

Participant (Audience) Objectives Changes in:	**Target Audiences**	**Impact Categories**	**Possible Measures**	**Possible Designs#**
Group chapters' mission statements	Public audience adults	Engagement or interest	On-line surveys; Questions added to existing surveys or required reports	Time series design
Group chapters' activities	Public audience adults	Behavior	On-line surveys; Questions added to existing surveys or required reports	Time series design
Participating group members' science and math attitudes	Public audience adults	Attitudes	Member ratings of their attitudes; Member responses to open ended questions	Pre/post measures
Participating group members' science and math activities	Public audience adults	Behavior	Member self report of frequency of participation in science activities	Pre/post measures
Informal science education activities in under-represented communities	Pubic audience adults and children	Behavior	On-line surveys Questions added to existing surveys or required reports; Review of documentation	Time series design
Participating group members' involvement in math and science reform efforts	Public audience adults	Behavior; Engagement or interest.	Member self report of frequency of involvement	Pre/post measures

Chapters with pre- and follow-up surveys became significantly more involved with math and science education both in terms of being more apt to do it and much more apt to have chapter goals dealing with math and science education. All responding chapters were, with one exception, now doing math and science activities.

Receiving training had a positive impact on members' attitudes toward science and math. In both the pre and the follow-up surveys, participants were asked the degree to which they agreed with a series of statements about math and science. After participating in training and being a part of the program for at least four months, participants' attitudes toward science became significantly more positive. For example, they were apt see science as useful for solving everyday problems, more apt to feel that everyone can do well in science and in math if they try, and were less apt to feel that learning math is mostly memorizing.

The program has been able to increase the presence of science in African American communities, training over 1,200 adults in how to do informal science activities in the community and holding over 1,100 informal science events that reached more than 90,000 people.

After being trained, the percentage of organization members doing activities related to math and science reform with the community doubled, increasing from 19.2% (50) to 39.5% (103). The percentage doing activities related to math and science reform with parents also doubled from 11.5% (30) to 23.4% (61). Also increasing was the number of participants running after-school programs and/or doing informal science with children or families (Pre: 53/20.3%; Follow-up: 85/32.6%).

Table 7-2 Fostering Inquiry Worksheet: Public Audiences

Participant (Audience) Objectives Changes in:	**Target Audiences**	**Impact Categories**	**Possible Measures**	**Possible Designs**
Learner science attitudes in both in- and out-of-school science	Public audience: youth	Attitude	Learner ratings of their attitudes; Learner responses to open ended questions	Pre/post measures with comparison groups*
Learner knowledge of science concepts covered in the units	Public audience: youth	Engagement or interest	Learner ratings of their interests; Learner responses to open ended questions. Learner self report of frequency of voluntary participation in science activities	Pre/post measures with comparison groups*
Learner problem-solving	Public audience: youth	Skills	Taping and coding of learners doing problem-solving tasks	Pre/post measures with comparison groups*

*See the note on maturation issues later in this chapter.

Example 2: Fostering Inquiry: This project's primary impacts are (1) to develop, implement, and test hands-on, inquiry-based units of activities for out-of-school programs for children ages 8-12 and, (2) to develop a support structure for after-school or out-of-school programs with science centers and children's museums. Participant objectives include changes in (see Tables 7-2 and 7-3):

- learner science attitudes and interest in both in- and out-of-school science;
- learner knowledge of science concepts covered in the units;
- learner problem-solving;
- after-school center programming;
- after-school staff science attitudes;
- museums/science centers behaviors;
- science center/museum and after-school program relationships.

Table 7-3 Fostering Inquiry Worksheet: Professional Audiences

Participant (Audience) Objectives Changes in:	**Target Audiences**	**Impact Categories**	**Possible Measures**	**Possible Designs**
After-school center science programming	Professional audience: after school leaders and directors	Behavior	Observations; After school staff report Learner self report	Pre/post measures with comparison groups
After-school staff science attitudes	Professional audience: after school leaders	Attitudes	Interviews; Science attitude scales	Pre/post measures
Science center/museum and after-school program relationships.	Professional audience: after school leaders and directors; science center/museum staff	Awareness, knowledge, or understanding	After-school program leader interviews Museum/science center interviews Review of documentation	Pre/post measures Case studies
Museums/science centers community outreach	Professional audience: science center/museum staff	Behavior	Community interviews Museum/science center interviews; Review of documentation	Pre/post measures; Case studies

Example 3: Studying the Impact of Visual Representation.

This project's primary goal is to study the impact of using digital cameras while doing informal science education activities on young people's knowledge of the science concepts covered in the activities (see Table 7-4).

Table 7-4 Visual representation Worksheet

Participant (Audience) Objectives Changes in:	**Target Audiences**	**Impact Categories**	**Possible Measures**	**Possible Designs**
Knowledge of science concepts covered in the unit	Public audience: youth	Increase in awareness, knowledge, or understanding	Science content items from the National Assessment of Educational Programs (NAEP) or from the science tests from states such as Massachusetts, Florida and Colorado	Pre/post measures with two groups doing the unit; one of which uses digital cameras

Example 3 raises some interesting evaluation issues. The project is a research project, not an evaluation. Program evaluation is not the same as research, although they share many characteristics. While they can use similar methods and provide similar information, "program evaluation focuses on decisions. Research focuses on answering questions about phenomena to discover new knowledge and test theories/hypotheses" (Young, 1997).

It is not clear what an evaluation of a research study should look like. The peer review process, where anonymous reviewers and editors decide if the quality of a study is high enough for it to be published, is one form of evaluation, i.e., scholarly benefits and research standards. Another form of evaluation can be the impact of the results of a study on performance of the field. Since that usually takes a significant amount of time to happen and publication and dissemination of study results tend to occur at the end of a project, such an evaluation can be difficult to do during a project period. One alternative (Table 7-5) is to provide research results to professional audiences who can then be questioned about their opinion of the value of the research and the ways, if any, they would use the results of the research.

Table 7-5 Visual Representation Worksheet: Professional Audiences

Participant (Audience) Objectives Changes in:	**Target Audiences**	**Impact Categories**	**Possible Measures**	**Possible Designs**
Knowledge of the impact of digital cameras in informal science education activities	Professional audience: curriculum developers	Increase in awareness, knowledge, or understanding.	Self report of changes in knowledge	Post-test only
Application of the knowledge learned from the study.	Professional audience: curriculum developers	Change in behavior	Self report of changes in behavior	Post-test only

ISSUES OF PARTICULAR INTEREST TO THOSE EVALUATING YOUTH AND COMMUNITY PROGRAMS

While the following issues can apply to a variety of different types of formal and informal science programs, they are especially frequent in and of particular concern to those evaluating youth and community programs.

Maturation

Maturation, or just getting older, is a key issue for evaluations of youth programs. As children age, they learn and change independent of any programs in which they participate, and do so more rapidly than they will as adults. Because of this, a design where over time young people in programs are tested twice (pre/post design) or more than twice (times series design) is not an adequate measure of change for most evaluations. Any changes of youth in programs need to be compared to changes of young people of similar ages and in similar environments to better see if any changes are due to the program rather than to maturation.

Real vs. Ideal

Many curriculum development projects funded under youth and community programs provide those who are piloting the curriculum with benefits such as training, materials and other resources that are not part of the final curriculum as marketed. Evaluations of the curriculum and its impact are most often done under more ideal circumstances, with people who have been trained and provided other resources. However, most informal science education curricula will be used primarily by people with no special training, who will be providing their own materials. The results of curriculum evaluations done under the more ideal conditions may not hold when the curriculum is used in more realistic environments. Evaluations may want to include a component that tests the usability and impact of the curriculum in more realistic situations.

Informal Science Education vs. Formal Science Education

There is often interest in finding the impact of informal science education on formal science education, especially student achievement. If this is done, then it is important to look at the content covered by any of the formal education measures/tests used. The question to be answered is whether the content of the formal education measure/test reflects the content of the informal science education program. Another concern is that there is a risk of alienating young people coming to an informal science education program by having one of their initial program activities be a formal science test. Care can be taken to devise the assessment tool so it feels like part of the program itself. For example, the evaluator can use a typical project activity to see if skills practiced earlier in the program are used spontaneously by the participants in the test activity at the end.

Cultural Competency

As noted in the first bulleted item opening this chapter, many of the youth and community programs specifically target groups that are under-represented in the sciences, most often girls and women of all races/ethnicities as well as African American, Latino, and Native American boys and men. Recommendations from a 2001 NSF workshop on cultural competency and evaluation that are particularly useful for evaluations of youth and community programs include:

- Cultural awareness of the environment from which the participants are drawn must be emphasized.
- Evaluations must recognize that the culture of learners influences how they respond to the assessment process and assessment items.
- Non-minority evaluators should be trained to evaluate programs that target minority learners (National Science Foundation, 2001, p. 53).

Sustainability

In youth and community programs, sustainability, that is the continuation of the program and its impact, can pertain to individual or to institutional change. Without studies done over a period of years, it is very difficult to assess the sustainability of individual change, particularly in geographic areas where there is a great deal of mobility. Sustained change is easier to track for institutions, including community-based organizations, science centers, museums, colleges, and universities. Indications of institutional change may include:

- Reallocation of resources;
- Continuation of program activities;
- Changes in professional development;
- Changes in mission;
- Continued changes in institutional practices and policies.

A bibliography of evaluation resources for youth and community programs is in Appendix B.

REFERENCES

National Science Foundation: Directorate for Education and Human Resources, Division of Research, Evaluation and Communication. (2001). The cultural context of educational evaluation: The role of minority evaluation professionals. Arlington, VA: Author (NSF 01-43), p. 53.

Young, Jean (1997). Program Evaluation: Background and Methods. http://ed.fnal.gov/trc/program_docs/eval.html accessed April 21, 2007.

CHAPTER 8 EVALUATING LEARNING TECHNOLOGIES

Barbara N. Flagg

The last thirty years have produced an abundance of technologies used as learning tools. Some technologies made waves and then died, like interactive videodiscs. Some newer technologies serve the same purpose as older ones; for example, an audio wand tour can now be downloaded as an MP3 file to a portable player and called an electronic guidebook. Technology has morphed from single functionality (e.g., telephone) to multi-functionality (e.g., cell phone cum camera cum web browser). Some technologies have moved the locus of activity from the individual (e.g., individual computer diary or game) to multiple users (e.g., social networking forum or multi-user game). The common trend in the development of learning technologies has been to pass greater control to the learner and to free the learner from a particular physical space. This movement is most radically realized with online Web 2.0 applications in which the users, not designers, generate most if not all content – in wikis, blogs, forums, and uploaded photos, stories or videos.

As learner control and freedom increase with learning technologies, the challenge of evaluating the affective, cognitive, attitudinal and behavioral impacts increases. Nevertheless, implementing an effective summative evaluation of learning technologies requires the same information and decision-making as for any other informal education program: definition of the impact goals and more specific audience or user objectives; description of the treatment and environmental and social contexts; identification of the target users; deciding what evidence indicates success and selecting appropriate measures of impact to use in a summative evaluation design that will permit causal inferences.

The role of the evaluator through the phases of development of a learning technology is also similar to that for other informal education deliverables. During the planning phase, front-end evaluation helps define (a) the purpose of the technology and the content focus; (b) the knowledge, interests, and attitudes of the target users and their familiarity with the technology; and, (c) the advantages and disadvantages of the technology in the intended environment for the intended target audience and what kind of infrastructure is necessary to support the technology. In the design and production phases, formative evaluation assesses early concepts, paper proofs and/or prototype versions of the learning experience, looking at technology appeal, feasibility and usability as well as traditional issues like content appeal, clarity and comprehension (Rubin, 1994). Evaluations of learning technologies during the development phases are particularly important because technology that is "unfriendly" will likely not have a positive impact on users in a summative evaluation. When the technology is completed and implemented, the summative evaluator steps in to assess impact on the intended audience or users.

WHAT IS THE INTENDED IMPACT AND HOW WILL YOU KNOW?

Each decision a designer of learning technologies makes will have some sort of impact on the target user – either intended or unintended. The more completely you can define your intended impacts early in the design process, the more likely your design decisions will lead to preferred outcomes. Knowing what your audience or user objectives are influences how you design your program as well as how you evaluate it. By defining what it is you want your users to feel, know, think, believe, and/or do after exposure to your product, you have a better chance for success.

Let's consider a technology product that supports a collaborative virtual learning online environment in which users work together to accomplish collective goals. These environments support social networking, providing tools for collecting and sharing knowledge. In the virtual environment, users are represented by their avatar (the user's graphical representation). The user's avatar can act individually or work and converse with other avatars of remote live participants via chats, or receive guidance or listen in on conversations of computer-controlled characters.

The impact for our fictional product is to increase preteen understanding of the scientific process. To make strategic design decisions, and eventually to evaluate the product, we need to characterize the impact clearly and unambiguously: what do we mean by 'understanding;' and what do we mean by 'scientific process'? Precisely defined audience or user objectives meet the needs of both the designer and the evaluator. The technology designer needs defined objectives in order to develop a format that will attract a user, teach or reinforce certain knowledge, and elicit specific user behaviors or interactions; whereas the evaluator uses the objectives in order to develop instruments (e.g., questionnaires, observations, and weblogs) to measure user outcomes and gather evidence about impact.

To move from the impact to the user outcome objectives in our example, we ask ourselves "how will we know" that our preteen user understands the scientific process? In our example, we more specifically define "scientific process" as a creative process of discovery that includes asking the right questions; developing models; and carrying out experiments and making observations. We define "understanding" in our example as being able to describe the steps of scientific method.

With objectives in hand, our designer now produces an engaging storyline and interactions in which the user, through her avatar, experiences aspects of the scientific process. In our example, the user's avatar is placed in Our Home Town, where she can learn from a variety of sources (reading the newspaper, conversing with the grocer) that the city's water is being polluted. Other avatars, representing other online users playing simultaneously, are trying to figure out where the pollution is coming from. All players can interview citizens of the city and each other, read documents, collect water from various wells and run tests, and share their findings (or not) on a database.

With explicit outcome objectives, the summative evaluator can develop measures to assess impact. Numerous measurement techniques, both qualitative and quantitative, could be used to assess this objective. In our example, evidence of success might be that a majority of our preteen

sample succeed in implementing aspects of the scientific process in the virtual environment and come to an acceptable conclusion about the source of pollution. Success is measured within the product itself. To step outside the product, we are successful if - given a defined research problem in an interview - our preteen users (compared to non-users) generate significantly more steps in more detail to design and carry out an experiment. The latter measure calls for transfer of knowledge and may be more difficult to achieve.

There is no right answer for stating objectives, but it is important to be realistic about your objectives. Submitting prototypes to formative evaluation helps hone the statements of intended outcomes to be more realistic and guides product revisions to achieve the planned outcomes. Additionally, evaluators need to be open to the possibility of unplanned outcomes, both positive and negative. For example, our preteen users may have such a good time within the deliverable's virtual environment that they are motivated to join other virtual environments, or they may become disaffected about science and decrease participation in school science.

EXAMPLES FOR FIVE IMPACT CATEGORIES

The five fictional case studies below focus on one of each of ISE's impact categories: (1) awareness, knowledge or understanding; (2) engagement or interest; (3) attitude; (4) behavior; and (5) skills. Each project would likely address more than one of these categories, but for simplicity in this chapter we are limiting the discussion to one impact category per project. These case studies do not begin to present all the possible formats of learning technologies - an arena that is changing daily - nor all the possible configurations of objectives, evaluation designs and measurements. Many NSF-sponsored projects also include other deliverables besides learning technologies. See the other chapters for discussion of exhibits (5), mass media (6), youth and community programs (9), and combinations thereof (10).

AWARENESS, KNOWLEDGE, UNDERSTANDING

Our first fictional technology example takes advantage of the newest personal digital assistant (PDA). Through the processes of front-end literature review, expert feedback, and formative evaluation with users, the designers produce an interactive multimedia program whose impact is to increase awareness and knowledge of science, technology, engineering and mathematics (STEM) careers. Museums offer the program and technology free of charge to teens during their visit.

The teen user first answers an interactive questionnaire that gathers information on the teen's interests and talents. The program identifies STEM careers matched to the teen's responses. Upon clicking of those careers that most interest the user, the PDA presents a brief day-on-the-job video and a museum map and photos indicating where the teen can obtain more background on the field of interest. So, for example, a teen whose answers indicate an interest in medicine is directed to the museum's health gallery or someone whose personality inventory demonstrates an

analytical mind is directed to the forensics exhibit. Once in the gallery, the PDA's wireless mapping program recognizes where the visitor is and offers the teen more information about careers related to specific exhibits within sight.

The summative evaluation for this technology program is a post-test only experiment with random assignment to groups (Campbell and Stanley, 1966). Other evaluation designs are possible, but an experiment permits us to show that our PDA program has an impact on users that cannot be explained in alternative ways. Visiting teens are recruited to participate in the study and randomly assigned to one of two groups: the experimental group, which is given the PDA to use; and the control group that experiences the museum without the interactive technology. Random assignment helps us make sure that the impact is due to the program and not some systematic differences in the makeup of our comparison groups. Both groups visit the museum as was their intention and upon exiting are interviewed about STEM careers. Table 8-1 displays our fictional objectives and evidence.

Table 8-1: Impact on Awareness, Knowledge and Understanding

Impact	Impact Category	Audience/User Objective	Evidence
Users of the PDA program will increase in their awareness and knowledge of STEM careers	Awareness, Knowledge or Understanding	Users, as compared to non-users, will experience more museum exhibits related to careers, for longer periods.	Tracking of the teens through the museum revealed that the experimental group viewed significantly more career-associated exhibits for a longer average dwell time than the control group.
		Users will become aware of a greater number of STEM careers than non-users.	In response to a face-to-face exit interview, the experimental group described a greater number of STEM careers than the control group, indicating a greater awareness.
		Users, as compared to non-users, will be able to describe more careers.	Those who used the PDA program described a greater number of activities that STEM workers do and described them in more depth, indicating positive impact on knowledge.

ENGAGEMENT OR INTEREST

Learning technologies that encourage user-generated content present significant challenges in terms of evaluating impact in part because the treatment or user's experience may not be well-defined and controlled by the designers and may change based on what is contributed by users.

In this fictional Web 2.0 project that accompanies a television nature series, the impact is to increase children's and adults' engagement and interest in the process of observing animals. The web site presents a shell for family units to post pictures, videos, drawings, maps, observational data and written descriptions about animals in their natural environment. The site provides numerous social networking features; for example, web registrants can view postings and write reviews; tag or add keywords to materials; write to or comment on forum postings; contribute to an animal wiki; link to other animal web sites; or chart posted data. When material is posted, the program asks users questions to encourage engagement with the observation activity (e.g., what do you think your animal feels like to the touch; what does your animal eat; what sounds does your animal make). The questions are different with each posting, and as a registrant increases frequency of posts, the questions become more complex to promote future participation and deeper engagement in the observational process. Additionally, forum moderators comment on and reinforce participation, and reward points are given to specific observation keywords (a nod to behaviorism).

The summative evaluation is a case study using content analysis of the web artifacts produced over time by a random sample of the population of user registrants. This unobtrusive analysis will permit the evaluation to make inferences about increased interest and engagement over time by objectively and systematically analyzing the quantity and quality of posted materials (Weber, 1990). The challenge in our content analysis is defining indicators that are valid measures of interest and engagement. Table 8-2 presents user objectives and evidence for our fictional project.

Table 8-2: Impact on Interest and Engagement

Impact	Impact Category	Audience/User Objectives	Evidence
Web site registrants will increase in their interest and engagement in the process of observing animals.	Interest and Engagement	With each visit, users will increase their duration on the web site.	Tracking of individual registrants over time showed that users increased in the amount of time spent interacting with the website.
		Users will demonstrate an increased emotional response to observing animals.	Emotion-laden keywords (e.g., awesome, cool, surprising) increased over time in written observations, descriptions, reviews and tags.
		The materials users post will increase in quantity, complexity and variety.	Content analysis of postings over time revealed an increase in the quantity, complexity and variety of posted content, reflecting increased interest and engagement in animal observation.
		Over time, users will ask more questions in the forum section.	Analysis of forum content shows an increase in questions asked of the moderator over time, indicating increased engagement with the activity.

Although we may document change over time in our registrants' emotional and participatory involvement in the website, we have no unexposed comparison group so we are not able to conclude that any observed change is actually caused by accessing our website. Other causative factors might be parent mediation, exposure to a school activity, or interaction with other web sites. On the other hand, our content analysis can generate hypotheses that could be assessed in an experimental or quasi-experimental evaluation. Note also that we have no way of verifying in our current evaluation design who the user registrants actually are. We could alternatively recruit families to increase control over our study sample.

ATTITUDE

A set of online interactive math games are designed for elementary school children in home settings with a stated impact of changing their attitude towards math. Formative evaluations help ensure that the math games are user-friendly and fun for the target age group. The summative evaluation implements a pretest-posttest experiment in which individuals are assigned randomly to play math games (treatment group) or to play a set of games of non-math content of comparable length, appeal and usability (comparison group). Parents ensure that their children play the games at home for one hour on one weekday every week for four weeks.

The impact on attitude is redefined as focusing on two user objectives that are precursors to attitude change: (1) increase interest in doing math and (2) increase confidence in ability to do math. Evidence of attitude is inferred from a set of quantitative and qualitative measures. Our measures for both outcomes include youth self-ratings on a standardized math attitude scale and homework diaries kept by parents. Parents keep a log of homework activity on the night before a game session and the night after a game session. Evidence of interest might be revealed in the sequence of homework activity, and evidence of confidence might appear in the frequency of complaints about homework. Table 8-3 presents our hoped-for results at the end of four game-playing sessions.

BEHAVIOR

This fictional case involves a 3D simulation game that addresses the impact of modifying adults' behaviors with respect to water conservation. With a variety of possible starting budgets, users design a house, making myriad decisions that impact water usage, water loss, and water waste treatment. Choices of toilets, showerheads, appliances, landscaping and so forth, influence the feedback graphs, which show levels of water usage over time and financial expenditures or savings. In appliance purchases, users consider tradeoffs that go beyond just water use; for example, an incinerating toilet may use no water but introduce less-than-desirable sounds and odors into the house. Players learn how much water is used with various everyday activities in a house and can adjust the behavior of house residents. Spontaneous events such as dripping

Table 8-3: Impact on Attitude

Impact:	Impact Category	Audience/User Objective	Evidence
Users' attitudes toward math will become more positive.	Attitude	Users, as compared with non-users, will report higher interest and confidence in math.	Self-ratings of those who played the math games were significantly higher for math interest and confidence than the comparison group who played non-math games.
		Those who play the math games will be more confident about math homework than those who do not.	Analysis of parental diaries showed that completion of math homework moved earlier in the sequence of all homework for the math game-players but not for comparison group, indicating increased interest and confidence in math. Parental diaries also noted that frequency of complaints about math homework decreased for the treatment group but not for the comparison group.

faucets, freezing pipes, septic system clogs, drought and a new baby intervene in the simulation to complicate the water conservation equation.

Because knowledge acquisition alone does not translate into changes in behavior, we recruit for our summative evaluation those for whom the information is personally relevant: adult homeowners who think they have high water bills. They are randomly assigned to interact with the simulation game or assigned to a comparison group that does not play. Having a comparison group is particularly important because our evaluation looks at behavior over a year and many other factors could influence our treatment group in that long time period. Interest, knowledge and attitude are measured with written questionnaires given to both groups (Bradburn, Sudman, and Wansink, 2004). In our focus on behavior, users and non-users complete an online home water appliance inventory and water usage behavior inventory prior to game-play and then revise their inventories four months and twelve months after the experience. Restricted budgets and timetables make assessment of long-term impact less likely particularly for a short treatment like our game; however, if the game were part of a larger multi-deliverables project, then evaluation over an extended period may be desirable. Table 8-4 displays results from our fictional evaluation.

Table 8-4: Impact on Behavior

Impact	Impact Category	Audience/User Objective	Evidence
Users will modify their behavior to increase water conservation.	Behavior	Over a period of one year, those who play the game as compared to those who do not will report modifying their water usage behavior by, for example, fixing dripping faucets, changing to low flow showerheads or planting drought-resistant plants.	Game players and non-players do not differ significantly in their home water usage in the pre-evaluation period; thus, the comparison groups are equitable at the beginning of the study. At four months, significantly more game players than non-players reported having looked for (and repaired) dripping faucets and more players had modified shower heads. More game players reported changing their water usage behavior only with respect to turning the faucet off during brushing teeth and washing dishes. Significantly more players had considered but not actually planted drought-resistant plants. No differences appeared in home water meter records between the two groups. At twelve months, no behavior differences appeared between groups, or differences in water meter records. A small but not significant percentage of players had discussed, looked at, or purchased water-conserving appliances.

SKILLS

As part of a larger museum camp program on inquiry-based investigation, campers interact with an online tutorial to learn basic technical electronic skills. Guided by the tutorial, campers make their own probe sensors to use later with a networked handheld computer in field investigations. Our indicator of successful mastery of the new skills is not a pre-post test but a simple observation: do the probes work. Table 8-5 displays our user objective and evidence.

Table 8-5: Impact on Skills

Impact	Impact Category	Audience/User Objective	Evidence
Campers will acquire basic technical electronic skills.	Skills	Campers will learn electronic safety rules.	All campers passed a short quiz about electronic safety rules.
		Campers will acquire basic electronic skills of wire cutting, wire stripping, and soldering, sufficient to make a probe.	All campers successfully made a working light probe and a temperature probe, demonstrating their skills of wire cutting, wire stripping, and soldering.

Our one-group posttest design, although not recommended for most projects, is appropriate in this case because we can logically eliminate alternative explanations for acquisition of the new skills. For example, we know from self-report in the registration information that our campers are not coming into the program with any electronics experience, and we can verify this information by observing in the initial part of the session that campers lack basic skills. We know by observing sessions that the only instruction is via the online tutorial and not from other intervening events such as direction given by camp counselors. It is possible that campers mature, change physically or psychologically, in the timeframe of the probe-making sessions, but since the sessions are relatively short that seems an unlikely alternative explanation for acquiring the basic technical skills to make a successful probe. On the other hand, our ability to generalize from our evaluation is limited to probe-making and to the population of self-selected campers interested in electronics and inquiry-based investigations.

CONCLUDING REMARKS

The fictional examples presented above play out only a few of the combinations that are possible for evaluating learning technologies, and the examples focus only on the impact categories intended to be aggregated across ISE projects. Beyond showing what change occurs, summative evaluations also collect information that helps us understand *how and why* the learning technologies affect which people under what conditions.

In measuring impact, we ask the question: did using the technology make a difference to users' lives? In measuring usability, we ask the question: how do users interact with the technology and can they interact in the manner in which the technology was meant to be used? Usability should be maximized through formative evaluation techniques prior to your summative evaluation so that you are looking at the ideal use in the impact evaluation. Summative evaluations measure usability because the manner in which users interact with your technology and the ease with which they do so affects the success of achieving your outcomes of change in interest, knowledge, attitudes, behavior and skills.

Additionally, a long-standing practice in media studies is uses and gratifications research that focuses on *what people do with media* rather than on *what media do to people*. With the introduction of each new type of media, evaluators and developers must come to understand the medium itself and its audiences in order to design the technology most effectively: who uses the technology; how do people use it; what needs does the technology fulfill; how do those needs influence people's response to the technology. Answering these questions adds to our understanding of why our deliverable has its various impacts on users and also contributes to the foundational knowledge of the field of learning technologies.

Technology not only serves the goals of education but also the goals of evaluation. Evaluators of informal learning experiences have explored the uses of technology as tools to aid their data collection and analysis tasks since the 1930's, when an electronic continuous response measure was developed to gauge affective responses to broadcast music (Levy, 1982). Also, collecting user information and feedback via the learning deliverable *itself* is becoming more common. For example, an online survey integrated into a web site or mobile phone experience. On the other hand, as labor and budget-saving evaluation tools are introduced, we must also recognize their limitations and pitfalls. Volunteer respondents to a web survey, for instance, may not be representative of the audience to whom we want to generalize.

Finally, we know that the technology available to support informal learning experiences is changing constantly, and the practices, treatments and possibilities that we discuss today will move in unknown directions tomorrow. However, human learning capabilities evolve much more slowly, and evaluators who utilize multiple methods to obtain a rich picture of their user-technology interactions and impacts will be rewarded, even as the technological ground shifts beneath them.

REFERENCES

Bradburn, N. M., Sudman, S. and Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design.* San Francisco: Jossey-Bass.

Campbell, D. T. and Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally & Company.

Levy, M. R. (1982). The Lazarsfeld-Stanton programme analyzer: An historical note. *Journal of Communication*, 32(4), 30-8.

Rubin, J. (1994). *Handbook of usability testing: How to plan design, and conduct effective tests.* NY: John Wiley and Sons, Inc

Weber, R. P. (1990). *Basic content analysis* (2nd. Ed.). Newbury Park, CA: Sage Publications.

CHAPTER 9 EVALUATING COLLABORATIONS

Randi Korn

This chapter is about evaluating the collaborative component of your ISE project. In the funding world, collaboration is garnering interest and support, especially NSF's ISE program. At one time collaboration was viewed as a good idea; today many believe collaboration is imperative (Gajda, 2004). NSF's ISE program lists collaboration as one of the key elements to be reviewed in considering funding requests. Noted business guru Frances Hesselbein of the Peter F. Drucker Foundation wrote that ". . . businesses and non-profits in today's interconnected world will neither thrive nor survive with visions confined within the walls of their own organizations" (Hesselbein, 2000). This sentiment is also expressed in *Evaluating Collaboratives: Reaching the Potential*, a book worthy of review by those engaging in and evaluating collaborative efforts (Taylor-Powell et al., 1998). Authors note:

> ". . . in order to obtain the legitimacy, power, authority, and knowledge required to tackle any major public issue, organizations, institutions, and citizens must join forces . . . Organizations that share objectives must also partly share resources and authority . . . to achieve their collective goals" (Taylor-Powell et al., 1998, p. 14).

Funders believe that when individual entities collaborate, they pool resources (dollars and intellectual assets) and achieve greater good than if they were to each work independently. At one time the collaborative nature of a project was in the background and few paid attention to how the collaboration was functioning; now collaboration is in the forefront, almost at center stage, a requirement and a condition for ISE project design.

COLLABORATION AND EVALUATION

The NSF solicitations indicate, as do those of other funding organizations in the private and government sectors, that highly functioning collaborations strengthen, extend, and deepen project impact. In general, evaluating the collaborative element of a project has the potential to inform the ISE field of how collaborative relationships function. From the perspective of an individual project, an external evaluator can study the inner workings of the collaboration to identify and measure the project's impact on the field. This book is meant to help PIs, evaluators, and proposal writers think about and ultimately articulate their project in terms of impact. This chapter discusses collaboration as a means to an end—that is, this chapter assumes that the collaborative aspect of your project is the key innovative element critical to the success of the project. This chapter will help PIs frame the unique nature of their project's collaborative relationship in terms of impacts, as outlined in Chapter 3. It will also help readers think about their collaborative partners and how the power of their collective intellects, skills, and resources can further their project's impact.

Due to their very nature, organizational collaboration affects professionals who are participating in the collaboration—often furthering their development as informal science education professionals and furthering practice (Inverness, 2005). Some collaborations may be conceived to further informal science education professionals exclusively, such as developing a Web site for professional use; other collaborations may involve creating a product for public audiences, such as an exhibition, but the unique collaboration of the partners involved is the primary innovative feature of the project. While the examples here may not explicitly relate to your ISE project, they should allow you to extrapolate how you might frame the collaborative impacts for your project. This chapter will be useful to PIs who are:

- Collaborating with organizations they have never worked with before;
- Managing a project where collaboration is the key innovative feature;
- Participating in a collaboration to conduct applied research that will further informal science education professional practice.

COLLABORATION: A NEW WORK STRUCTURE

There are many reasons why organizations choose to collaborate. Many do so because the grant guidelines require collaboration for the reasons stated above. However an ISE project will be more successful if collaborators realize that working together will afford them unique opportunities that will help them develop a superior project. A best-case scenario is that collaboration among organizations forms because collaborators share a goal and realize the product will be strengthened if they pool their individual and organizational assets, including intellectual resources, skills, financial resources, and organizational support and knowledge.

Collaboration also offers the prospect of change—a change in an organization's capacity or a change in participants' perspectives. For example, a few years ago NSF funded an exhibition project that was a collaboration between a science museum and an academic association, including scholars from the academic association who were serving as advisors to contribute their scholarly expertise. At the outset of the project, advisors discussed their expectations for the exhibition; their debates insinuated that the exhibition would include only text panels of their written words. Over time the advisors realized that text, though important, was but one element of many that would constitute the exhibition. The notion of a museum creating *experiences* was a new concept to them. Their understanding and perspective of what a museum exhibition can offer changed through their experience with this project, demonstrating that collaborations sometimes generate surprising results. Some outcomes are not preconceived; instead they are discovered along the way or become evident during the evaluation process. By conducting a summative evaluation that focuses on the collaborative process, the impact of the collaboration will be revealed through measures that identify the quality and amount of change, including the unexpected impacts that the collaboration has brought forth.

EVALUATION IN THE CONTEXT OF COLLABORATION THEORY

Collaborations create complex work environments and therefore require an evaluation strategy that responds to their complexity. When evaluating collaborations, some evaluators strongly recommend that evaluators select a guiding framework to avoid becoming overwhelmed with all the possible interrelationships. There are many frameworks available, but the one most frequently cited is *collaboration theory* (Frey et al., 2006; Gajda, 2004). Collaboration theory models are discussed in two references: Frey et al., 2006 and Taylor-Powell et al., 1998. Collectively, these models identify seven stages of collaboration as follows: coexistence, communication, cooperation, coordination, coalition, collaboration, and coadunation (Frey et al., 2006). These stages are important to consider, if only to help PIs realize that all teams pass through lower levels of collaboration before they reach productive group behavior.

Evaluation practice, like other fields of practice, is evolving. In the past when evaluators were asked to evaluate collaborations, they adapted elements from *process evaluation* (Nightingale and Rossman, 2006). Technically, a process evaluation examines a project's operating environment and the relationship between program providers and program recipients (Institute for Law and Justice, 1997). The funding community, with its emphasis on organizational collaboration, has prompted evaluators to rethink process evaluation and clarify what it should examine, as the relationship between program providers and program recipients is only one small part of a multi-organizational collaborative program. In order to study relationships among individuals who work in culturally distinct organizations involved in large, complicated collaborations, process evaluation now focuses on the human and organizational dimensions of the project and allows the evaluator to examine organizational and personal interactions, the integration of practice across organizations, the integration of organizational culture across organizations, relationship changes, and system changes (Gajda, 2004).

The evaluator may study the collaborative process from different perspectives—independently and interdependently (Nightingale and Rossman, 2006)—because as the number of collaborators increases, the number of possible relationships and affects of those relationships on organizations also increases. The evaluator could choose to look at each collaborating organization independently, citing the outcomes of the project on that one organization—without reference to the elements that may have caused the outcomes (e.g., the relationship that was forged with another collaborating organization). Similarly, the evaluator could examine the entire collaborating *network*, demonstrating how the interdependence of the organizations offered many outcomes—each one dependent and building on the other.

The evaluator may also study the organic nature of the collaboration over the life of the project. Accordingly, the evaluator will identify data collection times and sources that will provide information throughout the collaboration and devise measurement tools to capture the continuum of collaboration over time. For example, the evaluator may design a series of questions that ask respondents to rate various collaborative behaviors and activities on 7-point scales. If the evaluator wanted to measure how collaborative team members are interacting, the question might look like this:

Please rate your experience in today's meeting on the following scales based on your <u>honest opinion</u> of your experience <u>today</u>. (*Circle **ONE** number on **EACH** scale below.*)

I was not invited to participate in problem solving activities today.	1	2	3	4	5	6	7	My opinion was frequently sought when the team was problem solving today.
I did not contribute to the conversations today.	1	2	3	4	5	6	7	I contributed to the conversations today.
Today's activities did not include knowledge sharing.	1	2	3	4	5	6	7	Today's activities included knowledge sharing.

The evaluator might administer such questions twice annually over the life of the project. Ideally, over time, respondents' ratings would change, indicating a strengthening collaborative relationship.

FRAMING IMPACTS: COLLABORATION EVALUATION

During the planning phase of a project, PIs probably discuss their expectations for and potential impact of their project, but they may not have considered expectations for and impact of the *collaboration*. Preferably, *early in the project's life,* PIs must respond to questions about their expectations for the collaboration to prepare for the summative evaluation for two reasons:

- Measuring impact suggests there has been a change and baseline measures are helpful for comparison; and,
- If the project team seeks baseline measures, the evaluator should collect data at the beginning of the project's life—even if the evaluator's role is to conduct a summative evaluation.

Measuring impact requires that an evaluator participate throughout the project's life—even if the evaluator is only studying the collaborative element of the project (although many ISE projects have one project evaluator responsible for conducting all evaluations associated with the project). Evaluators often ask many questions to seek clarity about the project because thorough and deep understanding helps them do their work. Questions are provided below to illustrate what an evaluator might ask (note the alignment between the impact categories presented in Chapter 1 and questions 1 and 2 below). Answers to these questions help shape the framework the evaluator will use to determine the evaluation design and data collection instruments. How PIs respond will help the evaluator understand the intent behind the collaboration. Question # 3 and its sub questions are different; they focus on procedural issues—*how* you will achieve what you want to achieve. In process evaluation, evaluators are interested in the procedural structure you will impose among collaborators because they use it to identify data collection opportunities across the span of the project.

1. What **organizational** changes do collaborating organizations hope to stimulate through this collaboration? How will you know if you have been successful?
    - What is each collaborating organization's goal for pursuing the collaboration?
        - What are the differences and similarities among the goals of collaborating organizations?
        - What practices/resources (intellectual, in particular) will each organization offer?
    - What *knowledge* do collaborating organizations hope to develop through this collaboration?
    - What *skills* do collaborating organizations hope to develop through this collaboration?
    - What *attitudes* and *behaviors* do collaborating organizations hope to change and/or develop through this collaboration?

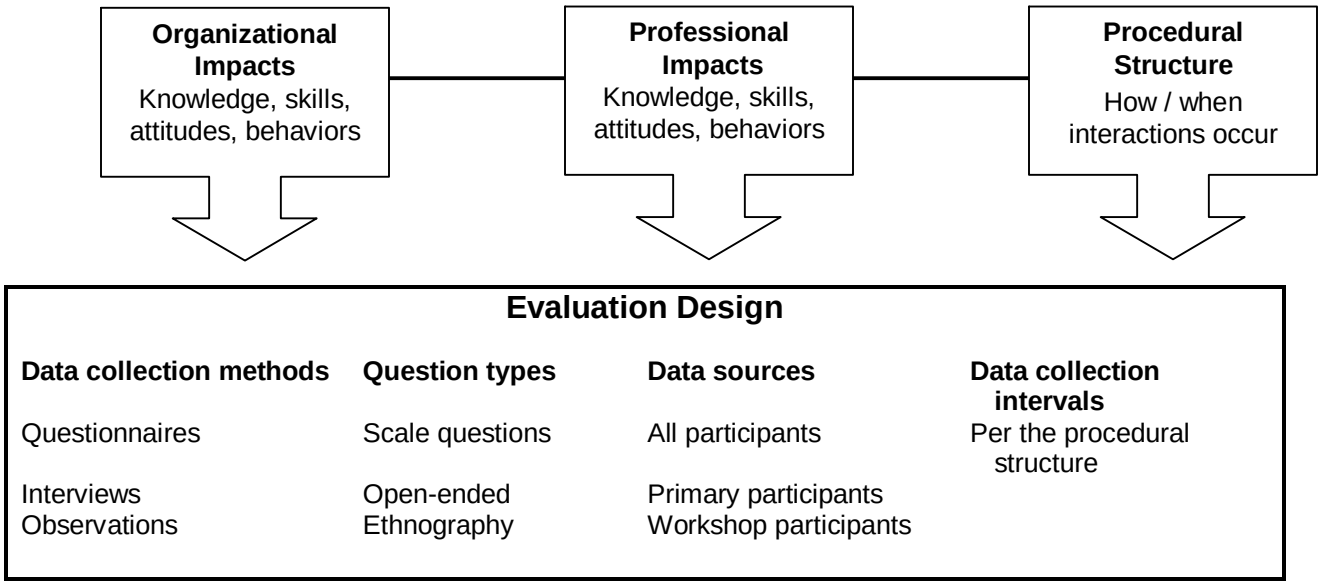
2. What do collaborating informal science education **professionals** hope to gain from the collaboration?
    - What *knowledge* do collaborating professionals hope to develop through this collaboration?
    - What *skills* do collaborating professionals hope to develop through this collaboration?
    - What *attitudes* and *behaviors* do collaborating professionals hope to change and/or develop through this collaboration?

3. What is the **procedural structure** of the collaboration—**how** will collaboration happen?
    - What processes will you use to facilitate collaboration?
    - What strategy will you use to align collaborating organizations' expectations?
    - What strategy will you use to align collaborating professionals' expectations?
    - What communication strategies will you use to facilitate in-person collaboration?
    - What communication strategies will you use to facilitate remote collaboration?
    - How will you address organizational and personnel challenges if they emerge?

Discussing these questions will help PIs in three ways: 1) if PIs work through them at the outset of the project, they will reach clarity on important issues that will affect the success of the collaboration; 2) clarifying expectations leads to clarity of purpose and vision; and 3) the questions represent three simultaneously-operating frameworks in collaborations—the organizational, the individual, and the procedural. The evaluator can use these three frameworks to think through the evaluation design, as the evaluation design will take into account intended impacts from organizational and individual perspectives and the procedural structure of the collaboration. For example, how will this collaboration affect organizational knowledge or behaviors? How will this collaboration change professionals' knowledge, attitudes, or skills? What procedural structure will PIs impose to stimulate, facilitate, and nurture collaboration and the intended impacts? What is the schedule for work sessions, professional development

activities, telephone communications, Intranet development, etc.? The evaluation design also includes identifying data collection opportunities (as per the procedural structure), data collection tools, specific questions, and data sources (e.g., the leadership of collaborating organizations; participating professionals, etc.). See Figure 9-1 for a graphical representation.

Figure 9-1. Relationship between the Structure of a Collaboration and Evaluation Design



ALIGNMENT BETWEEN PROJECT IMPLEMENTATION AND EVALUATION DESIGN

To effectively measure the impact of the collaborative aspect of a project, PIs and the evaluator each have distinct but highly interdependent responsibilities. Ideally,

PIs should:

- Create procedures and processes that allow collaboration (both in person and remotely) and respect the dynamic, evolving nature of collaboration;
- Reach consensus among all collaborators on impacts for each organization and participating professionals (using the questions presented earlier as a deliberation framework);
- Identify milestones of achievement throughout the collaboration and indicate how you will reach each milestone so the project stays on track; and,
- Acknowledge that conflict is a natural result of collaborative behavior and identify strategies to manage conflict for positive change (evaluation can serve as a tool for learning about and working through conflict).

Ideally, the evaluator should:

- Thoroughly understand impact statements and collaborators' meanings of them;

- Design evaluation instruments to:
    - capture transitions in the collaborative process (collect data at intervals over time coinciding with the collaborative procedural structure);
    - invite open conversation about the collaborative experiences (periodically conduct one-on-one interviews with staff throughout all participating organizations and facilitate group discussions);
    - measure change—if change is a desired project outcome (administer a standardized measurement device to track change numerically over time);
    - document the collaborative process (analyze meeting notes and observations);

- Implement data collection strategies throughout the collaborative process, following the procedural structure of the collaboration;

- Use actual milestone achievements to realign data collection activities; and,

- Offer assistance to PIs when complex situations arise, as the evaluator is on the edge of the project and can facilitate meetings during incongruous times.

As implied above, impact statements are extremely important to evaluators, as they influence the data collection tools and their questions. Also implied is the importance of the procedural framework, as it determines the data collection schedule and from whom and how the data will be collected.

Selecting Data Collection Methods

Chapter 4 lists the range of data collection methods available to evaluators and identifies their pros and cons. When evaluators select which data collection strategies they will use, decisions are based on a number of variables including impact statements (what is the best way to measure what needs to be measured); the product and process (e.g., exhibition, media, collaboration); data sources (e.g., who is receiving the deliverables—the public [specifically—adults, children, families] or professionals); data type required to capture achievement of impacts (quantitative data, qualitative data, both types of data); duration of project; procedural framework; and number of data collection periods. Many evaluators use several methods, mixing qualitative and quantitative strategies in an effort to capture the full impact of a project.

Identifying expectations for a project and selecting organizations that can assist in achieving expectations, while furthering the practice of *all* collaborating organizations and professionals is a difficult and complicated endeavor. However, when success occurs, it is useful to examine the variables that contributed to that success. Three ISE projects are summarized below to exemplify the importance of articulating intended impacts early in a project's life and the relationship between impacts and evaluation. All three examples involve museums, and indeed museums have been highly active in forming collaborations for the past several decades, especially as traveling exhibition development became so costly that consortia projects were needed to reduce the cost per institution participating.

1. A collaboration between a history museum and a science museum to develop an exhibition

A research center in a history museum sought to develop an interactive exhibition about creativity, invention, and play. The exhibition would include history exhibits as well as interactive exhibits to promote inventive and creative play. While staff could have worked with its exhibition designers, they decided to look outside the museum to further their practice in developing and prototyping interactive exhibitions—new territory for this particular history museum. They chose to collaborate with a science museum that was expert in developing and testing interactive science exhibitions. Historians and educators from one museum worked with exhibition developers from the other museum, each providing a skill set the other did not have. Together they created an extraordinary exhibition that achieved the majority of its projected visitor experience goals. Table 9-1 describes two intended impacts from the collaboration.

Table 9-1. Collaboration Impact of Example 1

Impact	Impact Category	Professional Audience Objective	Evidence
Museum integrates prototyping into its institutional culture	Knowledge	Exhibition developers describe the kinds of questions one asks during formative evaluation.	In a post-workshop debriefing, all staff participated in mock formative evaluation session, demonstrating the kinds of questions one asks during formative evaluation so their colleagues could realize the instruction qualities of formative evaluation.
	Attitude	Exhibition developers describe how formative evaluation helped them improve the exhibits they were testing.	Observations indicated that exhibition developers participated fully in the week-long evaluation workshop, observing and interviewing visitors as part of their participation, debriefing at the end of each day, and changing exhibits based on evaluation findings.

Summary of Impact Categories

- Impact of the collaboration: Staff in the research center at the history museum demonstrated to their museum colleagues how formative evaluation can work to the museum's advantage (change in staff members' attitude towards formative evaluation).

- Data collection methods: Reporting-back sessions to museum staff (not involved in the project) after formative evaluation, followed by an in-depth roundtable discussion to field questions about formative evaluation.

2. A collaboration between a scientific research society and a science center to develop an exhibition

Research scientists have much to offer the public but rarely have an opportunity to present their knowledge in a public venue. Science museums, on the other hand, have a dedicated audience of people interested in the work of scientists but often do not have scientists on staff to participate in the development of exhibitions or programs. The research society, with scientist members around the U. S., submitted an NSF proposal to collaborate with a science center to design and travel an exhibition: the scientists would provide the content for the exhibition as well as member scientists to volunteer in science centers that would host the exhibition; in turn, the science center would develop and design the exhibition and test interactives. Table 9-2 describes intended impacts for the collaboration.

Table 9-2. Collaboration Impact of Example 2

Impact	Impact Category	Professional Audience Objective	Evidence
The science center will integrate the work of current science researchers into their exhibition and program development.	Attitude	Staff members describe the value of working with practicing research scientists.	Post-project in-depth interviews with science center staff indicate that staff recognize the innovative quality behind many of the interactives that they co-developed with research scientists.
	Behavior	Each project staff member collaborates with one scientist and develops an exhibition activity.	Interviews among staff indicated that more than one-half of exhibits in the exhibition were co-developed between a scientist and a science center exhibit developer.

Summary of Impact Categories

- Impact of the collaboration: Before the collaboration, the science center did not have a regular practice of working with scientists. Given that the project included a society of practicing scientists, this project provided the ideal environment for exploring how to best collaborate with science researchers. The original intent behind the project was to impact science centers (integrate current research scientists' work into the center's programming), which was achieved, but research society member scientists also gained something from the experience—a new appreciation for science centers, as they learned about informal science learning and the role science centers play in supporting life-long learning opportunities for interested adults and children (an unintended impact).

- Data collection method: Qualitative interviews with all collaborators. (At the end of this project the PI had requested that the evaluator conduct a process evaluation; therefore, only interviews were conducted. Had the PI requested process evaluation at the beginning of the project, the evaluator would have designed data collection tools and integrated data collection strategies into the procedural structure and conducted a more comprehensive evaluation of the collaboration.)

3. A collaboration among small science museums

Small museums often find it challenging to develop interactive science exhibitions due to small staff and limited organizational capacity. For the first round of NSF funding, five museums collaborated to build and circulate a small traveling exhibition and associated educational programming (see Table 9-3). A second round of funding was sought by the original five museums and three additional museums to form four mentor partnerships to more fully develop exhibition design and evaluation capabilities. The third round of funding added a research component and studied the conversations between young museum visitors and their adult counterparts to understand the construction of science learning.

Table 9-3. Collaboration Impact of Example 3

Impact	Impact Category	Professional Audience Objective	Evidence
Build capacity among small science museums to design and travel science exhibits for small science museums to host.	Knowledge	Staff will describe the steps involved in developing, building, and traveling an interactive exhibition.	The exhibition that traveled to small museums was the evidence of success. In post collaboration group discussions, staff from collaborating museums described exactly what they learned about each stage of exhibition development, including the importance of identifying goals and objectives and testing ideas throughout development with colleagues and visitors.
	Behavior Skill	Participating museums will collaborate to plan, design, build, test, and travel an interactive science exhibition.	Data points indicated that participating small museums designed and built an exhibition that traveled to other science museums.

Summary of Impact Categories

- Impact of the collaboration: A number of small museums have strengthened their organizational capacity to plan, develop, design, test, and build exhibitions (increase in knowledge; change in behavior among professional staff; new skill development).

- Data collection methods: Qualitative round-table discussions, qualitative interviews, notes from facilitated meetings, focus groups, participant essays, exhibit planning documentation, and listserv.

As evidenced by these examples, collaboration projects are often designed to provide professional development opportunities. Each of these examples also suggests that a collaborative project may produce a product, such as an exhibition, program, or multi-media production for public audiences. A collaborative project also may produce research that generates knowledge in a particular field of study. When a PI asks an evaluator to evaluate the collaborative element of an ISE project, the PI perceives the project's collaborative element as innovative and vital to the success of the project. So while NSF ISE projects will all likely involve more than one organization, conducting an evaluation of the collaboration should generally take place if the collaboration—in and of itself—represents an innovative component of the project.

Traditionally, summative evaluation is designed to report how well a project achieved its impacts, but it is essential to include the evaluator in the project from the outset. This is especially true if the collaboration is going to be evaluated in a summative evaluation, because then data collection should take place throughout the life of the project—not only at the end.

REFERENCES

Frey, Bruce B., Lohmeier, Jill H., Lee, Stephen W., and Tollefson, Nona (2006). Measuring Collaboration among Grant Partners. *American Journal of Evaluation, 27 (3)*: 383 – 392.

Gajda, Rebecca (2004). Utilizing Collaboration Theory to Evaluate Strategic Alliances. The *American Journal of Evaluation, 25:1*, 65 – 77.

Hesselbein, Frances and Whitehead, J. (2000). Foreword to J. Austin, The collaboration challenge: How non-profits and businesses succeed through strategic alliances. San Francisco: Jossey-Bass.

Institute for Law and Justice (1997). Urban Street Gang Enforcement. Department of Justice, Bureau of Justice Assistance Inc. Washington, D.C. Retrieved April 25, 2007 from http://www.ojp.usdoj.gov/BJA/evaluation/guide/documents/process_evaluation_gangs.ht m.

Inverness (2005) Teaming Up: Ten Years of the TEAMS Exhibition Collaborative. Retrieved March 24, 2007 from http://www.inverness-research.org/reports/teams/2005- 04_rpt_teams-summative_eval.pdf.

Nightingale, Demetra Smith, and Rossman, Shelli Balter (2006). Collecting Data in the Field. In Wholey, J. S., Hatry, Hl, and Newcomer, K. (Eds.). *Handbook of Practical Program Evaluation*. Jossey-Bass: San Francisco, 363 – 395.

Taylor-Powell, Ellen, Russing, Boyd and Geran, Jean (1998). Evaluating collaboratives: Reaching the potential. Madison WI: University of Wisconsin Cooperative Extension.

CHAPTER 10 EVALUATING PROJECTS THAT COMBINE DIFFERENT TYPES OF DELIVERABLES

Cecilia Garibay

The first four chapters in Part II of this book have focused on evaluating impacts for a specific type of project—an exhibition, a community program, or a giant-screen film, for example. Increasingly, however, NSF awardees are developing projects which combine several deliverables across the various categories discussed, rather than focusing on one component. Like Chapter 9 on collaborations, this chapter will cut across the program areas of informal science education.

It is common, for example, to see NSF-funded exhibitions that include related programming and an on-line component; an educational television series with dedicated website content viewers can access; or collaborative projects among organizations that include components for both public and professional audiences. In some cases, projects combine three or more components. Some deliverables are developed as "value added" pieces, intended to complement or extend the primary experience (such as an exhibition), perhaps for a subset of the audience. In other projects, a suite of integrated components are designed to work together as a whole to achieve impact.

This chapter focuses on evaluating impacts of the latter kind: projects in which the use of multiple deliverables working together is viewed by PIs as central to the project's success.

When embarking on a project with multiple deliverables, a project will be much stronger if the components are well integrated. Therefore, it is imperative that such projects be based on a working hypothesis of why each deliverable—and the interplay among them—is necessary to achieve the intended impact. Doing so will make for a solid project that potentially can inform the field about ways in which combined deliverables work toward achieving impacts, the types of learning possible from such strategies, and how different components (such as varied media) can work together.

For projects with multiple deliverables, evaluation plays a key role in understanding how the integration of specific deliverables achieves specific impacts.

THE RATIONALE FOR COMBINING DELIVERABLES

During the conception stage of a project, PIs and other team members may have outlined impacts for the overall project, but may not have discussed in detail how multiple components are expected to interact with each other to achieve a desired impact. Yet understanding this

interaction is critical to developing appropriate evaluation strategies that accurately measure effects.

While developing various deliverables for a project may be tempting, it is important that a clear rationale exists for doing so. Consider these two examples:

- An exhibit about health and nutrition includes a website containing on-line activities based on the exhibit as well as related content for users to explore in-depth. The website is intended to extend visitors' experiences by allowing them to further explore the topic (e.g., making healthy food choices). It is expected that the general public can also access the website to learn about health and nutrition, thereby reaching more people than the exhibit alone.

- An exhibit about health and nutrition is combined with an on-line component where visitors may plan and track their physical activity and nutritional choices over time. Visitors create a user name and password during their initial visit to the exhibit, then can continue to use the on-line component to monitor their ongoing activity and eating habits. Tracking these data, long after the exhibit experience, is intended to help visitors make better health choices as well as develop awareness of the important components of health and the importance of longitudinal data collection in scientific investigations.

These examples offer significantly different approaches to developing a project combining different deliverables as well as differing rationales for doing so. In the former, the *exhibit* is the focus; the website experience, while a nice "value added" component, is not central to the target audiences' experiences or to achieving the goal of helping viewers learn about health and nutrition. It is still possible, of course, that in the first example the experiences of those visitors who access the website may differ from those who only visit the exhibition; but in this example, there is no *specifically articulated* hypothesis of how the website component actually adds to the overall impact.

In the second example, the on-line experience is intended not only to extend visitors' engagement with the content, but also to provide on-going monitoring and tracking of data about physical activity and nutrition in order to increase awareness and understanding of the use of longitudinal data collection in scientific investigations. In other words, the integration of the two components is necessary in order to achieve the intended impact.

Ultimately, the *intent* behind combining deliverables is a major step to guiding evaluation decisions. If, as in the first example, the intent is for certain components to act as companion pieces giving broader reach to the general public, then it may be less important for summative evaluation to devote significant resources to evaluating these accompanying pieces (although it would, of course, be expected that the evaluation would at least assess these deliverables in earlier evaluation phases).

Of course, a project with one or more value-added deliverable components can still succeed and significantly impact the intended audience. If your project focuses on one key deliverable with other components as secondary, it would be most useful for you to review the specific sections of Part II that discuss your particular project deliverable without further need to review this chapter.

Projects like the second example, however, will find value here as well as in the previous sections, which focus on a single deliverable.

In some cases, projects with multiple deliverables are attempted because PIs believe that bringing together different groups of professionals or institutions will advance informal science education. For example, collaboration among professionals can prove fertile ground for innovation and result in effective conceptual change and reform. A project focused on conceptual or systemic change will often include deliverables not only for the public, but also for professionals, with the intent of advancing dialogue around learning goals behind the public's experiences. If your project falls in this area, you should read this chapter in tandem with Chapter 9 (which addresses collaborative projects, including those for professional audiences).

PLANNING FOR EVALUATION

When teams are clear about the ways each deliverable—and the interplay among them—is expected to contribute to the project's impact, the evaluation process benefits. One potentially useful step is to develop, early in the planning and development process, a logic model with the evaluator. As illustrated in Chapter 4, a logic model is a picture of an organization's working theory and assumptions underlying a project; it links impact with program activities or processes and the theoretical assumptions of the project. Logic models can help "facilitate thinking, planning, and communications" about a project (Kellogg Foundation, 2001).

While a logic model is important in any project (see Chapter 4), it is critical for one combining deliverables. The process of developing a logic model can help articulate the project's working hypothesis about the interrelationship among deliverables. Ideally, development of a logic model takes place, in conjunction with the evaluator, early in the project, well before conducting any evaluation (front-end, formative, remedial, or summative)—one important reason, among others, to bring in an evaluator in the planning stages of the project.

An evaluator may ask the following questions concerning projects that combine multiple deliverables:

- What is the rationale behind multiple deliverables? What is the working hypothesis of *why* combining these specific types of deliverables will achieve the desired impact?
- How is each component expected to contribute to achieving a project's goal? What is the role of each deliverable?
- What is the interplay among deliverables? How do they interact *as a group* to achieve the intended impact?

In the previous example in which the two components are integrated, the working hypothesis might be that: a) engaging in an on-line component where users track their *own* physical activity and nutritional choices will motivate users to continue exploring exhibition content; and, b)

monitoring this information over time leads users to make better health choices and also helps them develop a greater awareness of the ways longitudinal data are used in scientific investigations. The contributions of the exhibition would be: 1) to provide visitors with an initial opportunity to encounter and explore the specific health and nutrition content; and, 2) introduce the long-term tracking component, giving visitors a chance to create a user name and password for the on-line experience—which is what allows for extended tracking.

Interrelationship of evaluation phases

This book deals specifically with summative evaluation, and therefore this essay discusses impacts in the context of projects combining multiple deliverables. It is important to recognize, however, that long before the use of summative evaluation, a project will have included several evaluation phases (front-end, formative, and remedial) aimed at increasing the likelihood of the project achieving its goals, as discussed in Chapter 2. The bibliography in the Appendix B provides sources for each phase of evaluation.

This notion may raise questions about the relative emphasis on evaluating the various project deliverables as well as about which aspects of the project are evaluated, and at what stages. Must a project evaluate *every* deliverable? If so, when does evaluation occur? If not, which deliverables are the most critical to evaluate? Decisions rest on careful consideration of intended project goals. A PI, in consultation with an evaluator, should focus on the major impacts desired and determine the extent to which each component contributes to those goals.

In most cases when deliverables are truly integrated, components will likely be evaluated in formative and remedial phases, because achieving intended impact depends on ensuring that all components work as a whole. This requires clarity about the difference between a promised project deliverable (such as an exhibition, website, or film) and an output that may or may not result from the intended audiences' engagement with such a deliverable.

Take, for example, a project in which teens participate in a four-week environmental science program intended to increase understanding of ecosystems and to help participants develop positive attitudes about science. The project hopes to achieve these goals through hands-on activities, interactive on-line challenges, and field trips to hear from and interact with scientists. Teens also develop a culminating project, prepare a poster summarizing their research, and present results to parents and community members at a community festival.

Here, the three key deliverables are the hands-on activities, on-line challenges, and field trips. The projects developed by participating youth are a function of their engagement and are a type of output rather than a project deliverable. Thus, formative and remedial evaluation would focus on assessing each of the three aforementioned components to provide feedback to the project team.

IMPACTS

The potential impacts of projects effectively combining deliverables will vary widely depending on specific deliverables and goals. These projects will generally focus on achieving impacts in the same areas that single deliverable projects do: an increase in knowledge or awareness; an increase in interest or engagement; a change in attitude or behavior; and/or development of new skills. When the multiple components are truly effective in an integral way, however, the results will show a synergy of impacts greater than the sum of the individual effect of each component.

This may appear as a purely quantitative difference (e.g., knowledge gain is far greater for learners using two components than for learners using either one), or it may appear as a difference in the quality of the learning (e.g., two components individually increase the knowledge of learners, but when used together they also change learners' attitudes). Overall, these types of projects may advance our understanding of how integrating deliverables may be especially successful in achieving specific types of impacts.

METHODOLOGICAL CONSIDERATIONS

As the chapters in this guidebook have indicated, no one evaluation design is appropriate for all projects. Of the many possible evaluation strategies for assessing project impacts, the most appropriate depends largely on the nature of your project.

Projects combining various deliverables, however, are likely candidates for mixed-methods studies. In mixed methods, evaluations utilize both qualitative and quantitative techniques. This approach is especially useful for understanding complex phenomena and can provide a more holistic understanding of different facets of a project (Green and Caracelli, 1997). Your evaluator will work with you to determine an appropriate mix of methods.

In some cases, projects with integrated multiple deliverables benefit from experimental evaluation designs. This is because one must closely examine as well as isolate the independent variables (in this case, the deliverables) to understand what influence each—and the various combinations—had in achieving the intended results. (Note that if a project has one main deliverable, with other components serving more as value-added elements, an experimental or quasi-experimental design may *not* be most appropriate or useful for your project.)

As noted in previous chapters, an experimental design involves what is typically referred to as a pre/post study, in which users are assessed before and after participating in an "intervention" (i.e., your project). In addition to some sort of "before and after" assessment, experimental designs involve comparing results with a sample group similar to your target audience, but that did not "use" your project. This is known as a "control group," and they also are assessed twice, but without actually engaging with your deliverables. The reason for using a control group is that it helps eliminate the possibility that changes in, say, knowledge occurred due to some other factor besides your project.

Another key feature of experimental designs is that participants in the study must be randomly assigned to one of these groups. That simply means that participants have an equal chance of being assigned to the control group or to the one receiving the "intervention" (i.e., actually engaging with your project, whether it be viewing your educational program, visiting your exhibition, or participating in your youth program). Randomization makes for a stronger design because one can be more certain that outcomes result from your project rather than because of some other difference between those who engaged with your project and those in the control group who did not.

In informal science education, however, a true experimental design is often very difficult to implement because it is usually not possible to randomly assign participants to an experimental or control group. In this case, then, a quasi-experimental design can be used. An evaluation can use a control group similar to the experimental group. To illustrate what an evaluation using a quasi-experimental design looks like, take the example of a project with a single deliverable, such as an educational science television program. Users are assessed both before and after viewing the program. Another similar group also receives pre/post assessments *without* seeing the program. A visual representation of this design might look like this:

Group	**Pre-assessment**	**Television Program (the "intervention")**	**Post-assessment**
A (intervention)	O	X	O
B (control)	O		O

Projects with multiple deliverables add a level of complexity to the evaluation design; in many cases, each deliverable must be examined both individually and in all possible combinations in order to assess impact. A project with two deliverables would need to include four groups: those who engaged with each of the deliverables on their own, another group who engaged with both deliverables, and the control group. Here is a visual representation of this design:

Group	**Pre-assessment**	**Project Deliverables (the "intervention")**		**Post-assessment**
		Deliverable 1	*Deliverable 2*	
A (intervention 1)	O	X		O
B (intervention 2)	O		X	O
C (intervention 3)	O	X	X	
D (Control)	O			O

One can see, then, that the more deliverables a project includes, the more complex the evaluation design will likely be.

Because of the nature of many informal leaning environments, however, it may not always be possible to implement a pre/post design. In fact, in some cases, pre-tests may not be desirable because they run the risk of "cueing" the audience. The very act of pre-testing sensitizes the user and can affect the outcomes. For example, in a famous study at the Hawthorne Electric Plant in the late 1920s, researchers found that people tended to improve their performance because they

knew they were part of an experimental group (this is commonly known as the Hawthorne Effect).

In some cases, therefore, evaluators will administer only post-assessments to all groups, and will accept the limitations of what can be claimed in the study results. In instances when control groups are not feasible or desirable, a study may simply compare each group (as in the table above) without a control.

One additional challenge in the evaluations discussed to this point is sample size; it may be difficult to find enough participants for the groups that engaged with two or more deliverables. If a project, for example, includes an exhibition and a website, there will likely be fewer participants who viewed both the exhibition *and* the website. Obviously, the more deliverables a project includes the fewer the people engaging with two, three, or more deliverables. This may mean small sample sizes, which in turn makes generalizing results more difficult. (See Chapter 3 for a discussion of sample sizes.)

The extent to which evaluating the interaction of various deliverables is necessary, of course, depends largely on the project's intent. In a project containing deliverables for both the public and professionals, in which the goal is conceptual change, focusing on other aspects may be more appropriate.

For example, a multi-institution project aimed at increasing public understanding of nanotechnology might focus on evaluating the impact of different conceptual change models rather than on specific interactions of media products in all possible combinations. Again, your evaluator will work with you in making these decisions, the logic that guides them and how they are connected to theory.

Given that the goal in the Project Monitoring System is on impacts that can be aggregated and generalized, quantitative or experimental studies will likely be selected. It is important to note, however, that in some projects with multiple deliverables, using other designs besides ones containing control or comparison groups may sometimes be appropriate. Designs that include case study or naturalistic approaches, for example, can be invaluable in understanding the relative impact of project deliverables on a broad range of target audiences. There is considerable value in generating multiple sources of data and analyzing the data for convergence.

Case studies, for example, can "often offset the marginalizing effects that can result simply from focusing on averages when analyzing data" (Allen et al., 2007, p. 240). Similarly, naturalistic approaches can help us identify and understand the "mutually influencing factors" that result in different experiences for individuals (Allen et al., 2007, p. 237). These approaches can also deepen our understanding of the *nature* of experiences for different people and can allow researchers to explore complex effects over time. These approaches are useful in varied circumstances, and might be especially critical when a project hopes to reach new audiences or a wide range of groups.

Overall, projects with multiple deliverables tend to be complex and you should be clear about the rationale for combining various deliverables. A well-developed hypothesis for doing so will lead

to a stronger project. The design of your summative evaluation will largely depend on the specific types of deliverables in your project, the intended target audiences, and the impacts you hope to achieve by combining deliverables. This chapter provided some possible approaches that are likely candidates for projects with multiple deliverables. Ultimately, the most rigorous design appropriate to the nature of your project and the intended outcomes is recommended.

HYPOTHETICAL EXAMPLE

Project summary

A project about photosynthesis had two components: an exhibition and a website. Both used immersive environments (in one case a virtual experience and the other an exhibit) to help visitors explore the world of plants and the basics of photosynthesis. Rather than having the on-line experience serve merely to extend the visitor experience, both components were developed with the same learning goals:

a) Increase visitors' understanding of the basics of photosynthesis.
b) Modify visitors' attitudes about plants, particularly in appreciating the key role they play in life on earth.

Summative Evaluation Design

Summative evaluation focused on assessing the overall impact of the project as a whole, including how the relationship among components contributed to project impact. The study examined the experiences of three groups:

- Those that only visited the exhibition
- Those that only visited the website
- Those that both viewed the exhibit and the website

The study compared experiences among the three groups to determine the effect of the on-line component on those who had visited the exhibition. The experiences of web-only participants were examined to determine how a web experience differs from an exhibition experience.

Results

Table 10-1 charts results for those who visited the exhibition only, web only, and those who used both.

Table 10-1 Photosynthesis Project Worksheet

Impact Category	Audience Objective	Exhibition Only	Web Only	Exhibition and Web
Knowledge	▪ Visitors will understand that plants make their own food	▪ 65% of visitors correctly identify the fact that plants make their own food.	▪ No significant difference in users' ability to identify plants making their own food as a key feature of plants.	▪ No significant difference between those who used components alone or together.
	▪ Visitors understand that plants use air, water, and sunlight to produce sugar.	▪ 30% of visitors can identify the three basic components plants use to produce sugar	▪ 55% of users can identify the three basic components plants use to produce sugar	▪ Users demonstrate a more in-depth and sophisticated understanding of photosynthesis concepts than those that only visited either the exhibit or web alone.
Attitude	▪ Visitors report an appreciation for plants' general contributions to animals and humans.	▪ 80% of visitors indicate that the exhibit helped them appreciate the role of plants in life on earth.	▪ 66% of users indicated that the website helped them appreciate the role of plants in life on earth.	▪ 70% of those using both the exhibit and web indicate that their experience helped them appreciate the importance of plants to life on earth. ▪ 40% indicated they discussed their experiences and appreciation for plants with family and friends

REFERENCES

Allen, S., Gutwill, J. Perry, D., Garibay, C., Ellenbogen, K., Heimlich, J. Reich, C. and Klein, C. (2007). Research in museums: Coping with complexity. J.H. Falk, L.D. Dierking, and S. Foutz (Eds.). *In Principle In Practice*. New York, NY: Altamira Press.

Greene, J.C. and Carcelli, V.J. (1997). Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms. In J.C. Greene and V.J. Caracelli (Eds.) New directions for program evaluation (Vol. 74). San Francisco, CA: Jossey-Bass.

Kellogg Foundation (2001). *Using models to bring together planning, evaluation, and action.* Retrieved August 10, 2002, from the Kellogg Foundation Web site: http://www.wkkf.org/pubs/Pub3669.pdf

APPENDIX A GLOSSARY

Lynn Dierking

(1) *Backward research design approach*—a planning process in which you first identify your desired results and audience(s), determine acceptable evidence for accomplishing these impacts and then plan the activities of the project.

(2) Types of Evaluation

(a) *Front-end evaluation*. This phase of evaluation provides input to decisions about how to develop a program in advance of the planning stage. Generally it provides background information for future project planning. It typically is designed to determine an audience's general knowledge, questions, expectations, experiences, learning styles and concerns regarding a topic or theme.

(b) *Formative evaluation*. This phase of evaluation provides information to improve the program during the design and development stage. Formative evaluation studies typically provide information about how the project can be improved and occur while a project is under development. It is a process of systematically checking assumptions and products in order to make changes that improve the final design or implementation.

(c*) Remedial evaluation*. This form of evaluation provides information to improve a project once it is complete and allows for corrections once projects are underway. Remedial Evaluation is the assessment of how all the individual parts of a project work together as a whole; like formative evaluation the goal of remedial evaluation is to improve educational effectiveness and insure achievement of goals and objectives.

(d) *Summative evaluation*. This form of evaluation assesses outcomes or impacts of a "settled" project. Summative evaluation is conducted after an interpretative media or program is completed and provides information about the impact of that project; what is assessed should be tied to project goals and objectives, however there should be an effort to document unintended outcomes also.

(3) *Indicator*— a precise and measurable indication of impact.

(4) *Construct*—a conceptual idea, such as "intelligence," which cannot be observed directly but is approached by using various tests, measures, and observation techniques.

(5) *Operationalize*--the act of translating a construct such as engagement into its observable manifestation.

(6) *Reliability*—a description of a measure that can be used by more than one person to consistently document an observed behavior in the same way.

(7) *Validity*— a process for improving the extent to which a measure approximates the construct it is intended to assess.

APPENDIX B EVALUATION BIBLIOGRAPHY AND RESOURCES

Randi Korn, Pat Campbell, and Cecilia Garabay

INTRODUCTION TO EVALUATION AND AUDIENCE RESEARCH

Diamond, Judy. (1999). Practical Evaluation Guide. Walnut Creek, CA: AltaMira Press.

Korn, Randi. (1994). Studying Your Visitors: Where to Begin. *History News, 49(2)*:23-26.

Munley, Mary Ellen. (1986). Asking the right questions. *Museum News, 64*(3), 18-23.

Patton, Michael Quinn. (1986). *Utilization-Focused Evaluation* (2nd ed.). Beverly Hills: Sage.

Screven, Chandler, G. (1990). Uses of evaluation before, during, and after exhibit design. *ILVS Review, 1*(2), 36-66.

Front-end Evaluation

Dierking, Lynn D., and Wendy Pollock. (1998). Questioning assumptions: an introduction to front-end studies in museums. Washington DC, Association of Science Technology Centers.

Mager, Robert F. (1975). Preparing Instructional Objectives (2nd edition), pp. 5-7. Belmont, CA: Pitman Management and Training.

Miles, Roger, & Giles Clarke. (1993). Setting off on the right foot: Front-end evaluation. *Environment and Behavior, 25*(6), 698-709.

Parsons, Chris. (1993). Front-end evaluation: How do you choose the right questions? *Visitor Studies: Theory, Research, and Practice, vol. 6*, 66-71.

Formative Evaluation

Borun, Minda, & Katherine A. Adams. (1992). From hands on to minds on: Labeling interactive exhibits. *Visitor Studies: Theory, Research, and Practice, vol. 4*, pp. 115-120. Jacksonville, AL: Center for Social Design.

Flagg, B. N. (1990). *Formative evaluation for educational technologies.* London: Taylor & Francis Group.

Jarrett, Joanna, E. (1986). Learning from developmental testing of exhibits. *Curator, 29*(4), 295-306.

Kennedy, Jeff. (1994). User-friendly exhibit design checklist. *User-Friendly: Hands-on Exhibits That Work,* pp. 69-74. Washington, DC: Association of Science-Technology Centers.

McNamara, Patricia, A. (1990). Trying it out. In Susan McCormick and Beverly Serrell (Eds.), *What Research Says about Learning in Science Museums,* pp. 13-15. Washington, DC: Association of Science-Technology Centers.

Miles, R., Alt, M., Gosling, D., Lewis, B.,& Tout, A. (1988). The design of educational exhibits (2nd ed.). London: Allen & Unwin.

Rubin, J. (1994). *Handbook of usability testing: How to plan design, and conduct effective tests.* NY: John Wiley and Sons, Inc.

Serrell, Beverly. (1996). *Exhibit Labels: An Interpretive Approach,* Ch. 13: Evaluation During Development, pp. 131-146. Walnut Creek, CA: Altamira Press.

Taylor, Samuel, (Ed.) (1991). *Try It! Improving Exhibits through Formative Evaluation,* pp. 9-75. Washington, DC: Association of Science-Technology Centers.

Methodology

Bradburn, N. M., Sudman, S. and Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design.* San Francisco: Jossey-Bass.

Campbell, D. T. and Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally & Company.

Cook, T. D. and Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Boston: Houghton Mifflin Co.

Fink, A. (2003). *The Survey Kit, Second Edition.* Thousand Oaks, CA: Sage Publications.

Gunter, B. (2000). *Media research methods: Measuring audiences, reactions and impact.* Thousand Oaks, CA: Sage Publications.

Ingersoll, Gary (1982). *Experimental Methods (in Encyclopedia of Educational Research (Fifth Edition);* Harold Mitzel ed. New York: The Free Press, *pp 624-631.*

Morgan, D. L. and Krueger, R. A. (1998). *The focus group kit: Volumes 1-6.* Thousand Oaks, CA: Sage Publications.

Mohr, L. B. (1992). *Impact analysis for program evaluation.* Thousand Oaks, CA: Sage Publications.

Patton, Michael Quinn. (1988). *How to Use Qualitative Methods in Evaluation,* Ch.5: Depth Interviewing, pp. 108-143. Newbury Park, CA: Sage Publications.

Payne, Stanley L. (1951). *The Art of Asking Questions*, Ch. 3: Who left it open? pp. 32-54. Princeton, NJ: Princeton University Press.

Serrell, Beverly. (1998). Paying attention: visitors and museum exhibitions. Washington DC, American Association of Museums.

Sommer, R., & Sommer, B. (1980). *A Practical Guide to Behavioral Research: Tools and Techniques*. New York: Oxford.

Stevens, F., Lawrenz, F., Sharp, L. (1993). *User-Friendly Handbook for Project Evaluation: Science, Mathematics, Engineering and Technology Education*. Arlington, VA: National Science Foundation, Division of Research, Evaluation and Dissemination, Directorate for Education and Human Resources.

Weber, R. P. (1990). *Basic content analysis* (2nd. Ed.). Newbury Park, CA.: Sage Publications.

Sample Data Collection Techniques

Exhibit 5, p 44 in *User-Friendly Handbook for Project Evaluation: Science, Mathematics, Engineering and Technology Education* (Floraline Stevens, Frances Lawrenz and Laurie Sharp (NSF 93-152))

Exhibit 2 p 20 in *User-Friendly Handbook for Project Evaluation: Science, Mathematics, Engineering and Technology Education* (Floraline Stevens, Frances Lawrenz and Laurie Sharp (NSF 93-152)). From *Educational Evaluation: Alternative Approaches and Practical Guidelines* by Blaine R. Worthen and James R. Sanders copyright 1987 by Longman Publishing Group.

Web-based Resources on Design:

http://www.socialresearchmethods.net/kb/design.php
A clear, short but comprehensive, on-line overview of quantitative designs covering the following areas:

- Introduction to Design
- Types of Designs
- Experimental Design
- Quasi-Experimental Design
- Relationships Among Pre-Post Designs
- Designing Designs for Research
- Advances in Quasi-Experimentation

http://www.socialresearchmethods.net/tutorial/Mensah/default.htm

A clear, short but comprehensive, on-line overview of quantitative designs covering the following areas:

- Biography
- Phenomenology
- Grounded Theory
- Ethnography
- Case Study

Protection of Human Subjects

http://ohsr.od.nih.gov/cbt/index.html
Short computer-based training, from the National Institutes of Health, on protecting human subjects, one for people who are doing research and/or evaluation and one for people who are members of institutional review boards.

http://www.nsf.gov/bfa/dias/policy/human.jsp
NSF web site which is regularly updated with rules and references:

Evaluation Resources for Youth and Community Programs

Harvard Family Research Project's Out-of-School Time Program Research and Evaluation Database
http://www.gse.harvard.edu/hfrp/projects/afterschool/evaldatabase.html

Informal Science Evaluation Reports and Resources
http://www.informalscience.org/evaluation/index.php

OERL, the Online Evaluation Resource Library: Search for Learner and Parent Instruments
http://oerl.sri.com/search/instrSearch.jsp

What We Know about Girls, STEM, and Afterschool Programs: A Summary
http://www.afterschool.org/sga/pubs/whatweknow.pdf

Web sites with links to multiple resources

http://www.ehr.nsf.gov/rec/programs/evaluation/main.asp

www.informalscience.org

www.visitorstudies.org

www.insci.org

APPENDIX C THE AUTHORS

Sue Allen is the Director of Visitor Research and Evaluation at the Exploratorium, one of the largest in-house museum research groups in the world. Over the last decade, she has conducted studies of teacher professional development workshops, exhibits and exhibitions, and public programs. Her research focuses on questions that are of interest both to museum practitioners and learning theorists, in areas such as scientific inquiry, narratives in science museums, exhibit design, mediation structures, and tools for assessing learning in informal environments. After studying and teaching physics in various contexts, she received her Ph.D. in Science Education from the SESAME program at UC Berkeley. She first joined the Exploratorium through a James S. McDonnell Post-doctoral Fellowship in Cognitive Studies in Educational Practice. She has served on the Board of Directors of the Visitor Studies Association, has taught courses at Rhodes University and UC Berkeley, and is currently a member of the NRC Committee on Learning Science in Informal Environments.

Patricia B. Campbell, PhD, President of Campbell-Kibler Associates, Inc, has been involved in educational research and evaluation with a focus on formal and informal science, technology, engineering and mathematics (STEM) education and issues of race/ethnicity, gender and disability since the mid 1970's. Her BS, from LeMoyne College, is in Mathematics, her MS, from Syracuse University, is in Instructional Technology and her PhD, also from Syracuse University, is in Teacher Education. Dr. Campbell, formerly a professor of research, measurement and statistics at Georgia State University, has authored more than 100 publications including co-authoring *Engagement, Capacity and Continuity: A Trilogy for Student Success*; *What Do We Know?: Seeking Effective Math and Science Education* and *Good Schools in Poor Neighborhoods: Defying Demographics, Achieving Success.* Dr. Campbell was a member of the US Department of Education's Impact Review Panel and was part of the team involved in the development of the National Science Foundation publication *Infusing Equity in Systemic Reform: An Implementation Scheme*. She received the Betty Vetter Research Award from Women and Engineering Program Advocates Network (WEPAN) and the Willystine Goodsell Award from the American Educational Research Association.

Lynn D. Dierking is Sea Grant Professor in Free-Choice Learning, Oregon State University (OSU) and a Senior Researcher at the Institute for Learning Innovation, Edgewater, MD. Dierking is internationally recognized for her research on the behavior and learning of children and families in free-choice learning settings and the development and evaluation of community-based efforts and has published extensively in these areas. Dierking has led a number of NSF-funded projects and evaluation studies and currently is collaborating with a Franklin Institute colleague on a research project retrospectively investigating the long-term impact of gender-focused free-choice science learning experiences on girls' interest, engagement, and involvement in science communities, careers and hobbies. Together with colleagues in the Science & Mathematics Education Department at OSU, she is helping to establish a Master's and Ph.D. program in free-choice learning within the existing K-12 and

college teaching graduate program. She received her Ph.D. in Science Education at the University of Florida, Gainesville, and has worked in a variety of science learning settings, including museums, schools, community-based organizations and universities. She serves on the Editorial Boards of *Science Education* and the *Journal of Museum Management and Curatorship*.

Barbara N. Flagg is Director of Multimedia Research, a national consulting group based in Bellport, NY, which specializes in front-end, formative and summative evaluations of technology based educational products. Clients include television stations, radio producers, filmmakers, software companies, museums, and universities. Recent projects include evaluations of public television and radio series and websites for children and adults, giant screen films, museum exhibits, interactive games and after-school outreach materials. Dr. Flagg served until 1990 on the faculty at Harvard University's Graduate School of Education, where she taught for ten years about design and evaluation of educational technologies. Her doctoral degree was received from Harvard University in Human Development, and her academic research studied how children attend to and learn from media. She is the author of an award-winning textbook, Formative Evaluation for Educational Technologies.

Alan J. Friedman is a consultant in museum development and science communication. For 22 years he served as Director of the New York Hall of Science, New York City's public science-technology center. Under his leadership the Hall won special recognition for encouraging new technologies, creating models for teacher training, serving an extraordinarily diverse audience, and evaluating the effectiveness of informal science learning. His work has been recognized by the American Association for the Advancement of Science's Award for Public Understanding of Science and Technology, the Association of Science-Technology Centers' Fellow Award, the American Institute of Physics' Andrew Gemant Award, the National Science Teachers Association's Distinguished Informal Science Education award, and the New York City Mayor's Special Recognition Award for Excellence in Science and Technology. He was President of the Visitor Studies Association for 2005-2007. Before coming to New York, Dr. Friedman served as Conseiller Scientifique et Muséologique for the Cité des Sciences et de l'Industrie, Paris, and was the Director of Astronomy and Physics at the Lawrence Hall of Science, University of California, Berkeley for 12 years. Dr. Friedman received his Ph.D. in Physics from Florida State University and his B.S. in Physics from the Georgia Institute of Technology.

Cecilia Garibay focuses on audience research and evaluation in informal learning environments, particularly projects aimed at reaching underrepresented audiences. As a bicultural/bilingual researcher, Ms. Garibay frames her work in culturally responsive and contextually relevant research and evaluation approaches. Some of her current research efforts include consulting with the Children's Museum of Houston on exhibit and program initiatives targeted to Latino communities, facilitating a strategic planning process at the Exploratorium intended to develop a long-term vision for building stronger community outreach, and conducting research at various museums on Latino audiences. She has led more than 60 evaluation studies—including research of multi-organization initiatives and collaborations—and is versed in all stages of evaluation. Ms. Garibay has consulted with a wide range of free-choice learning organizations, including the Association of Science and Technology Centers, Exploratorium, Science Museum of Minnesota,

Monterey Bay Aquarium, Phoenix Zoo, Wildlife Conservation Society, Children's Museum of Houston, Morton Arboretum, Chicago Botanic Garden, TERC, and the Conservation Trust of Puerto Rico.

Randi Korn is Founding Director of Randi Korn & Associates, Inc. (RK&A™), a company that conducts all phases of program and exhibition evaluation, mission evaluation, and visitor research in museums. With three offices (Alexandria, VA, Brooklyn, NY, and San Francisco, CA), RK&A's client list is extensive and includes the Peabody Museum of Natural History at Yale, Fort Worth Museum of Science and History, New York Hall of Science, Science Museum of Minnesota, Utah Museum of Natural History, among many others across the United States. Having completed over 500 evaluations over its 18-year history, RK&A is well versed in both qualitative and quantitative methodologies. Before starting her own consulting business in 1989, Randi Korn had worked in several types of museums, including natural history, science, art, history, and a botanic garden. She is the author of numerous articles on visitor studies, evaluation, and audience research.

Gary Silverstein, a senior study director at Westat, has provided evaluation and technical assistance services for such clients as the National Science Foundation (NSF), the U.S. Department of Education, the U.S. Department of Commerce, the Appalachian Regional Commission, the State of Pennsylvania, and the Robert Wood Johnson Foundation. He has developed online monitoring systems for several NSF programs—including the Informal Science Education program, the Math and Science Partnership program, and the Systemic Initiatives. Recent evaluation studies have included an examination of the impact of providing low-income families with computers and Internet access; an assessment of the implementation of the Young Epidemiology Scholars Program; and evaluations of education, vocational education/workforce development, telecommunications, and civic capacity-building projects in rural communities.

David A. Ucko serves as Deputy Director for the Division of Research on Learning in Formal and Informal Settings at the National Science Foundation. Previously, he was Section Head for Science Literacy and Program Director for Informal Science Education. He also is President of Museums+more LLC. Formerly, he served as Executive Director of the Koshland Science Museum at the National Academy of Sciences; founding President of Science City at Union Station and President of the Kansas City Museum; Chief Deputy Director of the California Museum of Science & Industry in Los Angeles; and Vice President for Programs at the Museum of Science & Industry in Chicago. Dr. Ucko was appointed by the President and confirmed by the Senate to the National Museum Services Board. He has chaired the Advocacy Committee and the Publications Committee of the Association of Science-Technology Centers. Prior to entering the museum field, he wrote two college chemistry textbooks while teaching at the City University of N.Y and at Antioch College in Ohio. Dr. Ucko is a Fellow of the American Association for the Advancement of Science and a Woodrow Wilson Fellow. He received his Ph.D. in inorganic chemistry from M.I.T. and B.A. from Columbia. E-mail: DUcko@nsf.gov.