Interview on Evaluation in Informal Science Education: Alan Friedman

Suggested Citation: Peterson, L. (Interviewer) & Friedman, A. (Interviewee). (2013, August 26). *Interview on evaluation in informal science education.* [Interview transcript]. Retrieved from http://www.informalscienceevaluation.org/
Interviewee: Alan Friedman, Consultant for Museum Development and Science Communication Interviewer: Lisa Peterson, SK Partners Note-taker: Archana Kannan, SK Partners Date and Time of Interview: August 26, 2013, 10:30am to 12:00pm (Pacific) Location: Phone/Skype

As part of our efforts to understand current evaluation issues in informal science education (ISE), we conducted interviews with leaders in the field. We purposely selected a sample of individuals who could provide insights from a range of perspectives; collectively, they have experience with ISE and ISE evaluation as practitioners, evaluators, researchers, funders, and institutional leaders. Several participants generously agreed to share the transcripts from their interviews.

Please note:

- These are transcripts of oral interviews, *not* polished or written remarks prepared for publication.
- These transcripts have been edited for clarity, brevity, and ease of reading. Participants were also provided with the opportunity to remove any potentially sensitive material.
- The views or opinions expressed are solely of the individual interviewee and do not necessarily represent those of their affiliated organizations.
- We intend for these transcripts to serve primarily educational purposes. We believe that others may benefit (as we did) from the rich insights provided in these interviews.

Interviews were semi-structured: we used a protocol that ensured asking key questions in a comparable fashion across interviews, but there was ample flexibility to allow for interesting and unpredicted turns in conversation. The coverage and order of questions varied across interviews. Interview topics included but were not limited to participants' views on evaluation uses, methodologies, "best practices," and challenges. Interviews were conducted in-person or by phone, and each lasted approximately 90 minutes.

In these transcripts, the following conventions are used:

- Initials indicate who is speaking. Blue text is used when interviewer is speaking.
- *Italics indicate paraphrasing or researchers' comments/interpretations.*
- 'single quotes' indicate hypothesized thoughts or questions; e.g., And I asked 'what have you had done before? And what did you think of it? And what do you need?'
- - single dash indicates an interrupted thought or change in thought; e.g., It's just been - I was just so happy to have had that opportunity to work with them.
- … ellipses indicate overlapping speech, deleted sections.
- [brackets] indicate non-verbal observations and other clarifications added by SK Partners.

A Note about this Interview:

Some of this interview focused on an evaluation that Alan Friedman and colleagues conducted of the *Star Games* exhibition at Lawrence Hall of Science, University of California, Berkeley. For more information about the evaluation, see:

Sneider, C. I., Eason, L. P., & Friedman, A. J. (1979). Summative evaluation of a participatory science exhibit. *Science Education, 63*(1), 25–36.

[BEGIN INTERVIEW]

LP: As we mentioned in our email, we are interested in learning from leaders in the field of informal science education and we are so pleased to be speaking with you today. So to warm up, I was wondering if you could just spend a few minutes walking through some personal highlights of your career, especially in relation to informal science education and evaluation.

AF: Sure. Well I had never heard of informal science education by the time I got my PhD, which is in physics. But I knew that there were things like museums and planetariums, I just never thought of them as primarily educational. They were more like, well, they were an academic discipline on their own and the public was occasionally allowed to wander in. Planetariums I got interested in because I just like looking at stars. The Hayden Planetarium in New York, where I grew up until I was age 5, was the first one I visited.

But anyway, I went ahead and got a PhD in physics, and then a few years later I was teaching physics at Hiram College in Ohio. There was not a planetarium, but an observatory. And a volunteer high school teacher opened it up to the public a couple of nights a month. I was on the physics faculty and no one else was much interested in it and they needed someone to supervise it, so I said I would supervise it.

So I wound up learning a lot from this high school teacher about how to run a star party for the public. That was really the beginning of my interest in informal science education, because if you've never been to a star party, it is outdoors, under the stars, with at least one telescope and people come and wait in line to look through it. It's really an exciting experience. It can be indeed a life changing experience when you see Saturn for the first time, when you see the craters in the moon, moons of Jupiter, and then you see these spiral galaxies, and nebulae. And all of a sudden it's becoming real, more real than it is looking at a picture in a book and more real than it is watching something on television. That began my interest.

About four years later I was on a fellowship at the University of California at Berkeley and at a dinner party I heard that they were getting a planetarium at the Lawrence Hall of Science and I knew someone there so I asked, 'What you going to do with it?' Cuz I'd had some things I thought about doing ever since I was running this little observatory. And they said, 'Well we don't know. Do you have any ideas?' So I show up the next morning with some ideas and that began basically 11 years of work at the Lawrence Hall of Science. And I've been in informal ever since.

LP: It's so interesting to see the route that leads you to your path...

AF: Right, but for the purposes of your project, it was a couple years later, maybe two, when we did our first exhibit to go around the planetarium and my office mate, named Laurie Eason said,

"Why don't we evaluate the exhibit?" And I said, "Well, I have looked at it, it looks good to me." And she said, "Well no, there is something we can do a little more rigorously than that."

So she was the one who introduced me to evaluation: front-end, formative and summative, and that *Star Games* paper was the product of that. It made a huge difference to me because suddenly it was not just education or entertainment. It was real science. You can actually learn something from the process of exciting people about astronomy and trying to teach it through the medium of an exhibit or a planetarium show.

So that was the big shift for me. It was not just playing around science; you can actually do science in the process of informal science education. And that was not at all obvious to me before.

LP: Ok, that is really interesting. And you said that was kind of a turning point for you?

AF: That was the turning point. Most of the faculty, this was at the University of California at Berkeley, the faculty regarded teaching primarily as a chore, something to be gotten through as quickly as possible; something to think about as little as possible and not as a science at all. It was simply, 'Well, there are students, and some of them will become graduate students, and some will become our colleagues--it is up to them to figure out how to do this. It is sort of a Darwinian survival of the fittest. And we might even upset things if we became good teachers--people who are not cut out to be researchers could find themselves getting PhDs.' So this is quite the opposite tack, saying, 'We can learn to become better educators just as we can learn to become better physicists or astronomers or whatever.' So that was a turning point for me.

LP: And that turning point for you, tell me more about that and how that affected your approach and your thinking.

AF: Basically, my initial appointment at Berkeley was 6 months, ¾ time after my fellowship ran out. At the end of 6 months, my term was ending but I was offered fulltime for another year. After about two years I had to make a decision as to whether to go back to my faculty position at Hiram College, where I started, or whether to stay. It was the experience of doing evaluation, of seeing the science behind science education, that convinced me to stay. At that point I had to say, 'I am not going to do research in physics anymore, I am going to research in how people learn physics and other sciences.' So I resigned from my position in Hiram, which they had been holding open for me, and stayed on in Berkeley on the research staff.

LP: So that was a significant turning point for you. So then was *Star Games* your first experience in evaluation?

AF: Yes

LP: Great, that leads kind of perfectly into my next question, which is do you have an exemplary or favorite evaluation project?

AF: Well of course that was my first one. So that's one I still feel very affectionate towards. And it became the model for a lot of what I did later. That was the last one where I was really in charge of the evaluation. After that, I was basically hiring either in-house or external evaluators and working with them to develop the evaluation plan and the analysis.

So I have many other favorites I've been involved in. The evaluation of the audio tours is one that we learned a great deal about. A lot of it was surprising. The evaluation of an exhibition called *Understanding AIDS* was a very important evaluation because of the importance of the exhibit and the topic and its impact around the world, well at least around this country.

The *Star Games* exhibit had a worldwide impact. There are over a dozen copies of that exhibition that made use of what we learned in the evaluation to convince people to replicate the exhibit and to do things the way we would have done it had we known the results of the evaluation.

But the AIDS exhibit was very important because it was a controversial exhibit; it was one that people doubted could have an impact. The evaluation that John Falk and Martin Weiss did. Similarly the exhibition that came just before that one was called *Hidden Kingdoms: the World of Microbes* and that also had a really good evaluation. Followed by an exhibit called *Marvelous Molecules: The Secret of Life*, also dramatically improved by front-end, formative, summative and in that case, remedial evaluation. So we did all 4 of the possible forms of evaluation. All three exhibits--*Hidden Kingdoms*, *Understanding AIDS* and *Marvelous Molecules*--I think are really marvelous exhibits in terms of their impact on the public, but also their attractiveness, their value for bringing people into a science that they did not even know existed. Those are among my favorite exhibit evaluations.

Planetarium shows, we also began doing quite a lot of evaluation. There were at least 3 PhD theses produced evaluating the style of planetarium show that my team developed at Berkeley and that's had an impact around the world. In particular, in helping people understand how to use small planetariums, the inflatable ones, like the Starlab, that all began with an evaluation of the program in that small planetarium in Berkeley in the 1970s. Materials from that project are still being used and they have been translated into 6 or 7 languages. So that evaluation is one that has had deeply gratifying long-term impact.

LP: Thank you so much for those examples. You had sent the AIDS evaluation in the packet you sent us so I was able to read that and it was very interesting. I'll look forward to maybe looking for some of these other ones as well. As we move forward in these questions we'll try to tie some of those examples back in. You mentioned impact a couple times and also kind of the spectrum of evaluation for the *Measuring Molecules*---front-end, formative, summative, remedial, we'll try to tie those in as we go through some of these other questions that talk more specifically about evaluation.

So we have been conducting a brief review of the summative reports on informalscience.org. Our general impression, and we're trying to check this with some quick coding, is that many of the studies are limited in their use of comparisons, such as either comparisons with other groups, or over time, or with other similar interventions. Do you agree or disagree with this general impression?

AF: I think that is a fair impression from what is currently posted on informalscience.org. I don't think it is a fair representation of the work that is actually being done. There are a couple of reasons for this:

First place, unfortunately, the majority of studies of informal science evaluation are not published at all; they are performed for in-house use only. There are various reasons why they are not published. There is resistance on the part of almost everybody involved, to publish something, especially when it is probably not applicable anywhere else. That's because these are evaluations of a very particular exhibit or program, with a very particular set of goals and objectives, and likely there is not much that is generalizable in the results of those evaluations. So that's the first reason, people just don't see that it does any good.

Second there's never, [pause] rarely is there adequate funding for evaluation. Putting something in a publishable form adds many months and thousands of dollars to a project. So if you don't think it's going to do any good and it is going to add to the time and the cost, why do it?

A third reason, is that often there are negative findings. So maybe we accomplished this and that, but in fact did not make any significant difference in some instances. If we publish that result, people might look it up and say, 'Well gosh, they only did half the things they thought they were going to do.' Now who exactly would say this I don't know because the only people who would look at an evaluation like this would be people like a granting agency and most of them are smart enough to know that a very good result would be that you accomplished-. Beverly Serrell has a wonderful phrase, 'If half of your exhibit accomplishes half of its goals with half of the visitors, you are very successful.' Half times a half times a half is $1/8^{th}$, but still that in fact is successful. But other people will look at the $7/8^{ths}$ and say, 'Wow that was a disaster.' A mistrust on the part

of audiences who might read a report has kept many of them from being published, because we don't believe many readers would understand this report.

Finally, informalscience.org was started as an NSF funded project. People who didn't do NSF funded things pretty much have ignored it. I don't think it yet has incorporated all of the studies published in Visitor Studies, and there are far more there than are on informalscience.org. They were supposed to be merged together and maybe they will be. But if you go to the Visitor Studies website, and you can get to it from informalscience.org, all of their papers that were published for the last 20 years of Visitor Studies are available--they are not all science, but many of them are experimental studies with comparison groups. There are a lot of pre-post, there are a lot of quasi experimental studies and there are a lot of studies that are true randomized controlled trials.

LP: You mentioned some of the studies on Visitor Studies that were published studies and more of those, it sounds like you're saying, are experimental designs, comparison groups, true RCTs. How important do you think it is to have those kinds of designs in summative evaluation?

AF: I think it is very important for summative evaluations. Summative evaluations that simply use self-reports of learners or visitors or people who receive the treatment are not all that reliable. So if in fact we are doing a summative evaluation for the purpose of knowing whether to do more exhibits or programs like this, or whether to change directions, or whether to recommend this strategy to someone else, then I think it is important to have really good data, by which I mean more rigorous than just surveys and self-reports.

The question is, what's the purpose of doing the evaluation? Again I've expressed to you some of my frustration that most of this work is not generalizable. I'll drop in here one of the things I've been thinking about in terms of recommendations: we need some modules that would be generalizable. Gil Noam's project to create a universal instrument is one of these. No one imagines you would use an instrument like Gil's as your only evaluation but you could fold it into a summative evaluation and then you would have at least one piece of your work that people could compare to the results they get from very different projects. Therefore, Gil's questions are not questions that are tied to a specific set of learning goals for a specific exhibit, they are more general learning goals. Some of this learning is affective domain learning--developing curiosity and identity. That is the area where I think we can do the most good right now, is to create some generalized protocols for looking especially at affective domain impacts and making these available for free use, but also maybe with a requirement that you publish your results. This would, I think, dramatically advance the whole field.

… [Skype connection glitches. From now on, turned off Skype video and used Skype audio only.]

LP: That leads into another question we had about what should summative be. So that recommendation talks about a piece of it. If you look at the bigger picture of what summative evaluation should be, how would you describe that?

AF: Ok, so what should summative evaluation be? Let's consider some different reasons for doing summative evaluation:

One is the funder requires it. So in this case the funder wants to know, 'What did you learn by doing this project and how effective was this project in meeting its goals?' Those are two separate things we can learn from evaluation. That's perfectly reasonable.

Some funders just want to hear nice things, 'Oh it was a great success, the exhibit drew of lot of people, a thousand people tried to sign up for 100 places in the after-school program.' That is important but that doesn't tell you much about impact of the program.

So I would see it being very desirable to be able to measure some kind of impact–that is, not only did people come to see it or participate in it, but they went away changed in some way. It could be they learned the following 5 facts. It could be they increased their interest. It could be that the following year they watched a science TV program that they would not have otherwise watched, 'cuz they never used to watch science on TV and now they are watching it. These are various kinds of impacts all of which are measurable, and all of which, since I am also a funder now, I think funders find very gratifying.

And you also want to know from your grantees, just for your own education as a funder, 'What did you learn?' One thing you might have learned was, this technique was not effective. It produced no measurable impact, or even more interesting, it produced a negative impact. We built this exhibit about some topic, people saw the exhibit and they were less interested in that topic afterwards. I'm sure that happens. I've been to many exhibits and thought, 'Well, that's all I need to know about that.' So what the impact was and what lessons you learned, which are two related but somewhat different things. That is one purpose of a summative evaluation.

A second purpose would be to make recommendations going forward in the future. We tried a hands-on exhibit about astronomy, could we then recommend doing additional hands-on exhibits about astronomy or anything else on the basis of what we know about how that exhibition behaved with visitors? And the answer was yes. All the exhibits I've done since then, I've made use of lessons learned during that first one.

Every time I do a summative evaluation on a project, I get something that I can use for the next iteration, either of that project or a completely different project. What's effective and what is not. And there are some things that turn out in general to just not be effective. Long text on the walls

is just not effective. A few people do read them, but in general, we know the longer the text, the less of that text gets read. That's something we've learned from summative evaluations. Interaction can be very effective but not necessarily. Sometimes people can get hung up on learning how to make the interaction work without even noticing what it is about.

Just an anecdotal example of this was an exhibit at the Lawrence Hall of Science about earthquake safety and we took a pinball machine, a real pinball machine with levers and bumpers and balls, and changed all the graphics so that it was about the steps you needed to take to protect your home in an earthquake. So attach the bookshelves to the wall, back strap the water heater to a wooden beam or a brick wall. And to win this pinball game you had to hit the ball to the little responder things for each one of these. It was extremely popular, people waited in line to play it, and not a single soul even noticed it was about earthquake safety. [LP laughs] They were just listening to the bells and the lights flashing and trying to rack up as many points as they could.

We had a similar exhibit in France that I've written about--a roulette table to teach probability theory. You played this exhibit, it was interactive, you played it on a computer, you had some play money to bet. You picked a theory of probability and then you saw whether you lost money or won money. The catch is in real probability, and this was a very good simulation, there is a randomness and because of that even bad strategies, ones based on a flawed understanding of how probability works, will nevertheless win sometimes. It turned out in the evaluation of that exhibit, people were just convinced that whoever won must have had the right strategy, without appreciating what the exhibit was supposed to show, which is, for example, that the roulette wheel has no memory. So if 20 red shows up 5 times in a row, it is neither more nor less likely to show up in the next spin of the roulette wheel. But no one, or very few people, learned the history-independent nature of a roulette wheel (or a coin toss) from the exhibit. So, we can make recommendations for going forward in a similar area or sometimes in a different area. That is the second use of summative evaluation, to get ideas you can use in future projects.

There is a third more pedestrian reason, which is just, 'Do we keep this exhibit, do we phase it out, or do we make copies of it and sell them to others or recommend this exhibit or program to others?' That's the third thing that can come out of summative evaluation, very specific, 'What's to be the future of this project based on its summative evaluation?'

LP: Great, thank you for outlining those three different areas.

AF: Three different uses of summative evaluation in the informal science world: first, to please the funder; second to learn something to use going forward; and the third to determine the future of that particular project.

LP: And with that I think we will go ahead and move into the specific example that we had picked out, which was the paper that you sent, *Summative Evaluation of a Participatory Science Exhibit*, and that one described an experiment conducted to evaluate the success of *Star Games* in meeting its educational goals. So some of the things that we have already talked about came into play with that. So we haven't seen, in the evaluations posted on informalscience.org, and you kind of clarified, you know, 'That's not all that's out there.' But we haven't seen a lot of experimental designs being used in exhibits. Why do you think we don't see as many in the sample that we've been looking at?

AF: Ok, well again, most of them are simply not published. I gave you a bunch of reasons why they don't get published. I had to fight with my own public relations staff and development staff who said, 'We would prefer that these not be published.' Because there are some kinds of funders, many of them that we had, who were not academics, who were not interested in academic use of a project, who only want to hear what happened with their money was good. And if we publish an evaluation that says, 'this is what didn't work,' it may freak out some of these funders. And to some extent I think my staff was right about damage that could be done by publishing mixed results of an evaluation.

But I went ahead and published anyway because I thought we could find replacement funders if we had to, but in terms of what we could learn in the long run to improve ourselves and the field, it was worth publishing. But in general, I think for all the reasons I gave–the cost, the potential downside of negative results, and especially this notion that it will not do any good anyway because the evaluation is so specific to a particular treatment, exhibition, planetarium show, program afterschool, whatever, so specific that it really probably is not useful.

Star games: this was the 1970s when this work was begun; and I think the several publications grew out that work, but I sent you the longest one, that was really, on my part and on the part of my institution, that was our first big project so we threw everything at it. We spent as much money and time doing the evaluation as we did building the exhibit. You can't do that very often. We were lucky because we got really good funding for it--NSF funded this not on the basis of an astronomy exhibit, but on an experiment in evaluation of a science exhibit. So there was a lot of money for it, and we did it very carefully, and we learned a whole lot.

I've had a few other projects like that, the audio tour project was very similar. We spent more money on the evaluation than we spent on the audio tours themselves. But this is rare. Most of the time you are given money to develop a 30 minute planetarium show; to create a 10,000 square foot traveling exhibition, and that's the purpose and most of the funders want to see their money "on the floor," in the exhibit or program itself, not in the form of an evaluation report. They want to see the exhibit. If you give them a choice between a more elaborate evaluation or more elaborate exhibit, many funders will pick the more elaborate exhibit every time.

And I have had at least one funder said they had no interest at all in doing evaluation and they refused to pay for any. If we knew what we were doing, we would do a good job. And they trusted we knew what we were doing, so if we wanted to evaluate this we could, but we couldn't use any of their money for it. So this attitude is out there. And as long it is out there, we are not going to see a whole lot of stuff published.

It also is expensive to do this. I figure 20-30% of most projects should be devoted to all forms of evaluation, but as I just described it can be 50% or more if you are really trying to do a thorough job and the evaluation is your main point in doing the project in the first place. But I have argued, and there is a paper called "Convincing the Director," if I haven't sent it to you it's on my website, you can download it, in which I say, 'Well, if you spend 30% maybe your exhibit will be 30% smaller, but it will be 60% better.' And if what counts is the impact as opposed to just the square footage, then evaluation is the best investment every time. That's my argument. So if it is impact per dollar, evaluation is a great way to spend. If you are only looking for square feet per dollar, then you have a case for not evaluating or doing just a minimal cursory evaluation. But if impact is your measure, then you are not even going to know if you succeed without evaluation. And I can personally guarantee that you will have more impact if do front-end and formative evaluation than if you don't.

LP: That actually reminds me another question we had for you, which is, thinking about all the different kinds of evaluation that can be done, how do you allocate your evaluation dollars to get the most bang for the buck?

AF: Ok, again, depends on purpose. If you want to produce the best exhibition or program, with the most impact on the audience, I would put maybe a quarter of my evaluation budget into front-end and the remainder into formative. And that will produce the best product. You won't know that it is the best, because you didn't do summative evaluation. I am just claiming it will based on almost 40 years of doing this. If you want to know just how good it was, then you add summative. And if you want to make it one step better then you add remedial, which is basically just summative evaluation, but you are going to use the result to make some final incremental improvements.

LP: If you had the choice, if you were not required by a funder, if you did not have to do the accountability piece for the funder, would you still allocate funds towards summative?

AF: Well I would because I am particularly interested in learning and improving the field as a whole. So I would do the kind of summative that will look at things that might be universal lessons and might be applicable for other projects. So I would do it every time because that's my reward. I do like seeing crowds happily enjoying an exhibit or a program but I am even more

interested in knowing whether 5 and 10 years from now there would be more crowds elsewhere enjoying things because of something we learned. Since that's my reward, whenever I can, I want summative evaluation.

But again, you don't always have a choice. The last major exhibit that I worked on, it opened about a month ago, the funder had an extremely tight budget and they were only interested in an expert judgment evaluation. They did not want to interview anybody. So I basically was the curator, a designer specified it, a fabricator built it, and then I spent 2 days watching visitors and noting their behavior and giving my personal report, my judgment on what was working and how it was working and then some lessons that I learned which I would use if they built more exhibits like that one. So that's not a rigorous evaluation by any means, and it's not what I would have done if it had been my choice. But in this case, that's what the funder wanted. Especially since, this is an exhibit which is not open to the public, it is only for use by employees of a research organization; I did not feel so bad about not having the chance to do more rigorous evaluation.

LP: Thank you, that's really useful. I think I want to circle back and dive a little bit deeper into the nitty gritty from the *Star Games* project. Did you consider any alternative designs for *Star Games*?

AF: For the exhibit or for the evaluation?

LP: For the evaluation.

AF: Yes. We actually did little pilots of pre and post, post only for-, because we were concerned about cuing the visitors-, if they know what we are going to ask them in advance, they might behave differently than if they don't know. But we did little pilots and eventually came up with the design which is post-test only for a randomly selected control group and an experimental group.

We looked at a couple of ways of doing the affective domain impact evaluation, the most successful of which turned out to be what came to be called the [Dennis] Schatz Raffle Inventory, described in that paper. We did it once raffling off posters and once raffling off books and we got a very different result raffling off books. There was no significant difference between the control and experimental group when the prize was a book, but a very significant difference when the prize was a poster. We concluded that the exhibit made people want to look at astronomical images but not necessarily read books about astronomy. And of course that was a disappointment. I really hoped they would want to read books. But at least the exhibit increased interest in looking at astronomical images.

LP: It sounds like you used some kind of pilot testing to come up with your final design and you took the time to do that. And did you encounter challenges in implementing that design?

AF: Yes, there were some logistics issues about how to prevent the control group from seeing the exhibit. And there were issues about, particularly the skills test where we asked people to focus a telescope and then timed how long it took them. We had difficulty with some of our intern staff – we only discovered after they had collected their data that they had just estimated the time, even through we had given them stopwatches. So we had to go back and check their data to see if was consistent with the data from people who actually used a stopwatch. Because there was a fair amount of data collected--several hundred visitors each tracked and timed, and interviewed, and given this skills test, and then the raffle test. We used a lot of non-professional data collectors. And I have since learned you really have to do that more carefully than we did it then. And when we used college students for most big evaluations at the New York Hall of Science and we actually train them, we watch them do a practice evaluation and we check up on them periodically to make sure they are following the protocol.

LP: Those are great lessons. Thank you for that. You talked about some of the measures that you used. You had kind of a psychomotor skill measure with the Focus Time Procedure, and then a choice instrument with the Raffle Choice. A lot of the reports and evaluations that we have been seeing use a lot of self-report.

AF: Right. I don't even remember if we collected self-report data. The other part of our test was the cognitive part. It was a sort of a multiple choice, but all with images. So one says, 'If you are looking through a telescope and you put a doughnut shaped piece of cardboard to reduce the diameter of the telescope where the light is coming in, how will the image you see change?' Now the right answer is, 'The image will get dimmer, but will not otherwise change.' But there were other choices, which is, 'the field of view would get smaller,' 'the magnification would become less,' and 'there would be no difference at all.' So that was basically a cognitive test about whether you understood the importance of the aperture of the telescope. It turned out that in the control group very few people, I think I am remembering this right, you've got to remember this was over 30 years ago. But very few people in the control group got that right but most of the people who visited the exhibit got it right. And there was one other group that got it right it turned out and that was photographers, amateur and professional photographers all know about f stops. They know that aperature changes the brightness of an image but does not otherwise change the image. So they were getting that right whether they saw the exhibit or not.

LP: Yeah we were struck by the design of that. How did you decide on that particular design for getting at those cognitive measures?

AF: Well we thought about just asking questions and check a box, but then we decided that any kind of terminology we use in the questions people might misinterpret. So when you say you stop down the aperture of a telescope, what does that mean? So we decided to go with the cartoon-like illustrations. Eventually, in the final version of the exhibit, we actually used cartoon-like instructions. It started with just written instructions. So it was basically a little pilot testing and then judgment that this was a visual exhibit and so the more visual we can make the assessment, the more it's likely to show us what's going on.

LP: And for the other measures that you had, the focus time and for the raffle choice, did you have alternative measures that you considered?

AF: For the raffle choice, I know we considered just asking people how interested they are in astronomy. So that would be a self-report. And then we decided we didn't believe what they'd say. Because the people who we'd just shown an astronomy exhibit would think we wanted them to say they were interested in it. So there is phenomenon of trying to please the interviewer, maybe to get the interview over as soon as possible or just because you appreciated their interest in what you thought and so you want to be nice. So went with the raffle inventory, and a very pointed thing was that they didn't have to, in front of us, make their choice. They could make their choice, write it on the back of the raffle ticket and drop it in a box. So we thought we were getting a less biased sample.

…[*There was a Skype audio glitch and we missed part of what AF said here*]

AF: …A major challenge which required formative evaluation was that aiming a high-magnification telescope at an object and finding it was one of the things we know is very challenging. To see if a telescope design was usable, you'd have to have people point the telescope, tell us what they were seeing, or first we would have to tell them, 'see if you can find the red dot in the telescope,' then we would have to look through the telescope to see if it was pointing at the red dot. All that would take a lot of time. So we went with the focus time test, because it didn't matter what they were looking at, when they told us it was in focus, we believed them and focus time was something we could measure pretty quickly.

LP: Actually your sound went out for just a second. You said the alternative you considered was having them do what?

AF: Oh, point, the accuracy and speed with which they could point the telescope at an object.

LP: Interesting. In terms of the design, did you consider designs other than experimental design?

AF: We knew we wanted an experimental design of some kind. It was a question of pre-post or post only. As you know, we went with a control group with post only test. We considered a variety of alternative models.

LP: In terms of use, who did you see as the primary stakeholders of that particular evaluation project?

AF: It was targeted at teenagers. Basically they were all middle school to high school students coming in class groups. Then we would randomly assign class groups to the control exhibit or the experimental exhibit.

LP: Those were the users of the intervention itself correct?

===[more Skype audio glitches. Both parties hang up and resume interview via telephone]

LP: It sounds like the teenagers were the stakeholders or the users for the intervention. I am wondering who the stakeholders were for the evaluation.

AF: Well it was a National Science Foundation project. It was primarily-, the original name of the grant was not astronomy exhibit; it was something about evaluating the impact of an interactive exhibition. So NSF, was a primary audience.

And then we always planned to publish the evaluation. There were a couple of versions of it published--the one in *Science Education* is the most formal, but there was also a version published in *Planetarium*, which is the Journal of the International Planetarium Society and there was a piece in *Sky and Telescope*, which is a well-read amateur and professional astronomers' monthly magazine. So it wound up having reports in three places. So that meant it was our professional colleagues–in astronomy, astronomy education, and science education.

LP: And you said you had always planned to publish.

AF: Yes.

LP: How did that decision come about?

AF: Because, again, this was, for us, a great learning experiment to see how much we could actually learn from doing these evaluations. The Lawrence Hall had done one similar rigorous evaluation, and that was comparing two formats for an exhibition and that was the first published exhibit evaluation from the institution and this was the second.

LP: Do you know about how the results of the evaluation were actually used for different stakeholder groups?

AF: I certainly know that over the next 10 years a dozen museums built copies of the exhibition, either small ones like the Exploratorium, or full blown, as big or bigger than the original. I still see new exhibits with elements that were in this exhibit appearing and they may not even know where the original one came from. We encouraged that, that's why the article's in *Sky and Telescope* and *Planetarium*, we encouraged it and we kept citing the evaluation as the encouragement by saying, 'This really works. People really do learn something from it and they get more interested.' In terms of NSF, I know that we were cited in many other exhibition proposals as an example of the kind of evaluation they would do.

LP: And then internally also?

AF: Oh, well yes. Internally this was the model that to this day is still used whenever funding is available, at the Lawrence Hall and then when I came to New York, at the New York Hall. I should say that there were other people at the same time or even earlier learning to do these kinds of evaluations. So this was not the only example, but this was just a particularly rich example because it was so visual. People at the Franklin Institute, at the Exploratorium, they actually came in to this kind of rigorous formative evaluation work somewhat earlier or later, but they all do this kind of work now. Some of it was inspired by the work at the Lawrence Hall and some of it again was underway already. But we all talked to each other. At that point, in the 1980s, I probably knew everyone doing this kind of evaluation in the country.

LP: That sounds like a valuable network.

AF: Oh it is. Well the Visitors Studies Association, which is international, and the Visitor Studies Group based in London, a number of others, they are really terrific networks. They are small, few hundred people in each one.

LP: So another question, one of our impressions that we've been seeing as we've been reading and talking to people, is it seems challenging sometimes to meet evaluation needs while remaining authentic to the free-choice nature of informal learning experiences. What are your thoughts about that?

AF: My thoughts are that yes, it can be challenging. But it's an essential component. For example, I remember because I had to say it so many times, the exact wording we used: 'We'd like you to take a look at this exhibit,' same words for the control group and the experimental group, 'we'd like you to take a look at this exhibit, stay for as little or as long as you like, and

when you leave, would you stop back by this desk and I'd like to find out what you thought about it.' So that notion of 'spend as little or as long as you want,' that's an explicit recognition that this is a free choice environment. In a classroom, we would say, 'your lesson is going to last 20 minutes and then we'll have a test.' Because you can make it last exactly 20 minutes. We can't do that. What we can do is to time how long people stay and see if there is correlation between how much they learn, or what the impact is and how long they stayed. But we have to let them make that choice.

So the one thing that we sort of broke the free choice paradigm or protocol with was, we did say, with each person who was in the experimental group and the control group, 'We'd like you to look at an exhibit that is under development and tell us what you think. So if you are willing to do this', which everyone was, 'then come with me and I will show you where the exhibit is.' It would have been better if we had been able to randomly pick and track people as they came in the door and saw what percentage of them voluntarily went to that exhibit. That would be another piece of information. And we might have distorted things somewhat.

Maybe this exhibit, for example here is a possibility, the subject matter is not inherently attractive but once people spend some time in it, it really works and they get excited about it. We couldn't learn that because we didn't give people the opportunity to discover it on their own. So yes, you have to make some compromises, you have to work remaining as true as possible to the free choice nature, and if you are distorting it too much, your results are just not going be worth as much. So we thought we came up with a reasonable compromise on this one but that is something that we have to think about all the time.

LP: Do you think there are particular measures or assessments that are a good fit?

AF: One is to use completely unobtrusive, that is by tracking people, by watching what they are doing, even by wiring up the interactive device so you see which way they pulled and pushed and how hard, you can do that. So they don't even know there is an evaluation going on except that you had to go through some sort of IRB protocol.

For example I saw an interactive exhibit at the Exploratorium about a month ago and there were signs all over saying, 'This exhibit is-, everything in this exhibit is being videotaped and monitored as part of an evaluation to learn how to make better exhibits. If you object in any way, please do not enter this area.' And there was no way to get in without passing one of those signs. So that distorts things to some extent, but once you enter, then you don't know anything is going on. You can't see that there is a camera looking at you, you see there is a camera, you don't know if anyone is in the other end of it, or if it is being taped, or what. So these are compromises designed to respect the free choice nature of the learning and yet still get information out of it.

LP: One of the things we're curious about is how summative evaluation can play into decision-making. And you can answer this for any of your experiences, but also curious, when you were at the New York Hall of Science, how evaluation fit in with your strategic planning.

AF: For one thing, a strong summative evaluation makes it much easier to get the next grant from some funders, so like from the National Science Foundation, as long as we can prove that we really learned something from doing this and it might be useful for the field.

For example, we did the first audio tour with a grant from NSF, and then we got a follow-up grant from another division of NSF to test out the audio tour with visually impaired visitors. We probably wouldn't have gotten that second grant if we had not learned a whole lot from the first grant. So that's one way in which we use it for certain funders who care about evaluation. It is a mechanism for getting follow up funding or even funding for a very different project, but it is now a learning experience.

There is one funder that I am actually now on the funding side of because I'm on the board of the Noyce foundation. One of our tests is, 'Will making this grant make us a smarter funder?' If we don't know what the results of the grant were, it can't make us a smarter funder. So we insist that all of our grants explain to us how it is going to make us better funders. What are we going to learn from it? And a negative finding can be just as useful as a positive finding.

It also, I think, and I've written about this separately, if you are doing a summative evaluation and if everyone working on the project knows that it's happening, is a part of it and knows the results, it helps all of them focus on why we are really in this business in the first place. Otherwise it is very easy to focus on, 'I design beautiful exhibits. It is an art form. I am going to make surprising-looking exhibits, amusing exhibits, elegant exhibits. That's well and good but that's not why science museums exist. We exist to have an impact on our visitors and we can have an impact through being elegant or funny or beautiful, but if we are not having that impact it doesn't matter how beautiful or funny or elegant. So the evaluation helps everybody focus on the real bottom-line, which is, "What impact are we having on the communities we are serving?"

LP: Thank you for those different perspectives too, kind of the funder perspective and then kind of the broader perspective and kind of getting everyone on the same page. *[LP does a time check, mentions ways that AF can submit additional thoughts after the interview.]* …I guess the last question that I want to make sure I get in is, what do you see as the most critical evaluation-related challenges and issues that need to be addressed in informal science education? Kind of, where does this field need to go?

AF: Right. I've got several points to make:

One is, it is really important that those of us who work in informal science learning, or informal learning in general recognize the importance and value and necessity of evaluation. We see that because some of our stakeholders, like board members, the federal government at the moment, many private funders are not sure they want to keep funding us. They are not convinced that they are getting their bang for the buck. I think the proper way, not the only way, but the best way to assure our future is to have really good evaluation and to show it around as much as possible.

So the funder wants to know, 'Did my funding make a difference?' If you look at government for example, government agencies will say, 'Well, it makes a difference if it gets the politicians who support us re-elected. Can you show me you do that?' The answer is no. So we've got to make an argument to them and say, 'Getting re-elected is vitally important, we understand that, but one way to get re-elected is to be able to show that you've done good things for people. So if we can show that your support of informal science education at NOAA or NASA or the National Science Foundation made people more comfortable to live in a world of science and technology, that is something you should brag about and it will help you get re-elected I think. And here is the hard evidence that the funds, the support you give for that function at that agency is having that impact.' So it's got to be first the people in the field recognizing the importance of this and how to use it. And then they've got to persuade their stakeholders–funders, governing agencies, their own employees--they've got to persuade them that this is important.

The second thing is, everyone has to feel it is so important that you are going to invest time and money in it. Obviously the funders have to feel that way, but a lot of projects happen that are funded out of earned revenue. In that case it's your own director, and your staff, and your board that have to be willing to spend part of their money on evaluation. This is the argument I told you about before--if you spend 30% of your money on evaluation, you won't get a bigger exhibit but you will get a better exhibit. If it's impact per dollar that really counts, which I think it is, then you're better to spend money on evaluation. But that argument has to be made. It is not automatically accepted.

So, number one, you've got to convince everyone that it is important and valuable, number two, they've got to come up with the resources to make it happen.

And number three--and this is the hardest one--we all need to be doing more generalizable work. That is the universal items, the modules we can add into our evaluation. So we don't just learn about our ability as an individual, as an organization, but we learn about the field as a whole and how it can advance. So, CAISE, the Center for Advancement of Informal Science Education, has as part of its mission, promoting evaluation for this purpose, but to do that they've got to have more studies that make use of common protocols. So that's the third really important thing I think has to happen.

…

LP: Ok, wonderful, thank you Alan so much and have a wonderful day.

AF: Ok, you too. Bye.

[END INTERVIEW]