The Palo Alto Convening on Assessment in Informal Settings

Synthesis Report

October 2014

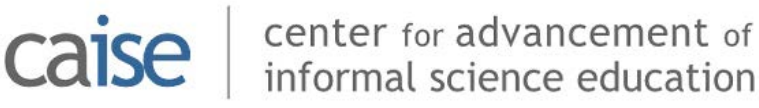
Authors:

Patrick Shields, SRI International
Eric Greenwald, SRI International

James Bell, Center for Advancement of Informal Science Education (CAISE)
Kevin Crowley, Center for Advancement of Informal Science Education (CAISE)
Kirsten Ellenbogen, Center for Advancement of Informal Science Education (CAISE)

Suggested Citation:

Shields, S., Greenwald, E., Bell, J., Crowley, K., & Ellenbogen, K. (2014). The Palo Alto Convening on Assessment in Informal Settings: Synthesis Report. Washington, DC: Center for Advancement of Informal Science Education (CAISE). http://informalscience.org/research/ic-000-000-010-051/Palo_Alto_Synthesis_Report



SRI International
333 Ravenswood Avenue • Menlo Park, California 94025-3493 • 650.859.2000

Participants

Organizers

Facilitators
- Kirstin Ellenbogen, CAISE
- Patrick M. Shields, SRI International

CAISE
- James Bell, CAISE
- Kevin Crowley, CAISE

SRI International
- Eric Greenwald, SRI International

Critical Friends

- Sue Allen, Independent Consultant
- Leslie Goodyear, Education Development Center (EDC)
- Ellen McCallie, National Science Foundation (NSF)
- Vera Michalchik, SRI International
- Martin Storksdieck, National Research Council

Practitioners

- Linda Kekelis, Techbridge
- Lynn Schmitt-McQuitty, 4-H

Funders

- Ann Bowers and Ron Ottinger, Noyce Foundation
- Janet Coffey, Gordon and Betty Moore Foundation
- Carol Tang, S.D. Bechtel, Jr., Foundation

Represented Projects

Developing Citizen Scientists
- Brigid Barron, Stanford University

DEVISE
- Amy Grack Nelson, Science Museum of Minnesota
- Norman Porticella, Cornell University
- Tina Phillips, Cornell University

FOCIS
- Xiaoquing Kong, University of Virginia
- Ji Hoon Ryoo, University of Virginia
- Robert Tai, University of Virginia

PEAR
- Gil Noam, Harvard University
- Larry Suter, Independent Consultant

Science Learning Activation Lab
- Mac Cannady, Lawrence Hall of Science, University of California at Berkeley
- Rena Dorph, Lawrence Hall of Science, University of California at Berkeley
- Chris Schunn, University of Pittsburgh

SYNERGIES
- Lynn Dierking, Oregon State University
- John Falk, Oregon State University
- Nancy Staus, Oregon State University

Note: The many contributions by the convening participants to the content of this document were compiled by Patrick Shields, Eric Greenwald, James Bell, Kevin Crowley, and Kirsten Ellenbogen.

The Palo Alto Convening on Assessment in Informal Science Settings

In December 2013, a group of leaders of six informal science education (ISE) assessment projects met in Palo Alto, CA for a 2-day exploration of the state of the art of measuring the impact of informal STEM education experiences. The goals for the meeting were to explore in depth the technical and practical details of the assessments, share and critique findings, and review plans for ongoing work to validate and refine measures.

The need for the meeting evolved from discussions at two larger gatherings on evaluation and assessment in informal science education. One was a previous convening that the Center for Advancement of Informal Science Education (CAISE) hosted on building capacity in evaluation for the field of ISE. Prior to that meeting a gathering was convened by the National Research Council and the Program in Education Afterschool and Resiliency (PEAR) at Harvard to discuss assessment procedures in informal science.

In the CAISE convening leading evaluators, researchers, practitioners, funders and policy stakeholders identified the issues of greatest concern to the community: common measures and aggregation of findings; access to and coordination of resources; professional development; and, advocacy for the value of evaluation. Common measures emerged as the most urgent of the topics as a group of projects in attendance realized that (1) they were working on an overlapping set of constructs and (2) the tools and measures they were developing, as well as related products, would benefit from some collective examination and comparison.

The NRC-Harvard convening resulted in a paper authored by PEAR entitled *Game Changers and the Assessment Predicament in Afterschool Science* (http://informalscience.org/images/research/Noam_Shah_Science_Assessment_Report.pdf). The paper identified these trends:

- After school and summer programming has become essential and pervasive.
- Schools and school science are in the midst of a change towards new standards and curriculum that reflect many of the core beliefs of ISE.
- There is momentum towards collaboration in the ISE field, as opposed to competitive researchers each developing their own version of instruments.
- There is an increasing demand for data—funders and others are looking for documented outcomes.

An additional finding of this meeting was that common instrumentation could be very useful for the advancement of the field. The opportunity to have large samples that cut across different settings and contexts to demonstrate differential outcomes linked to program quality was compelling and a model existed in the Common Instrument measuring interest in science.

Concomitantly, a group of funders who have been investing in assessment of informal STEM learning—the Gordon and Betty Moore Foundation, the Noyce Foundation, and the S.D. Bechtel Foundation (which had funded the NRC-Harvard convening)—were interested in the continuation of the common measures agenda. Building on momentum from the two earlier convenings, the Palo Alto Meeting on Assessment in Informal Science Settings responded to the need to establish an empirical basis for what works, for whom, and under what conditions in informal science learning environments.

In general there was a collective sense by all involved in the Palo Alto event that this was a timely convening and that there is a clear need for the field to be proactive in strengthening the empirical basis of claims about learning in informal settings. The need is based in the broader goal of the ISE evaluation and assessment community to build a systematic evidence base for science in out of school time. The convening also aimed to transcend the proprietary nature of assessment and evaluation, which in the past has been a challenge for the ISE field.

It was also recognized that the field has important resources where evaluators, researchers and practitioners can receive information about assessments. One such website is InformalScience.org, which has many evaluation reports available that provide good examples of what instruments are presently being used. PEAR, with funding from the Noyce Foundation, created Assessment Tools for Informal Science (ATIS) (pearweb.org/atis), a searchable website with over 60 tools for use in STEM out-of-school learning environments. These websites help define gaps in assessment relevant to the discussion at the Palo Alto convening.

This short document summarizes the convening. In the section that follows, we describe each of the six assessment projects. We then summarize the comments of practitioners attending the meeting. The third section reviews the responses of a critical friends group of researchers and funders. We then synthesize the findings and list a set of takeaways. Finally, we describe some post-convening activity.

Participating Projects

The six projects that participated in the Palo Alto meeting included the following.

Program in Education Afterschool and Resiliency—Assessment Tools in Informal Science (PEAR-ATIS). Director Gil Noam presented the "Common Instrument," a short survey designed to be used for assessment across many out-of-school educational programs. The goal of the Common Instrument project, which involves close collaboration of researchers and practitioners, was to create an assessment that would be immediately useful to the afterschool STEM world. The instrument was designed to be brief and easy to administer, so that programs would be able to use it while implementing their programming. It was also designed to be a pre- and post-measure, in order for program leaders to track the progress of individual students as well as to identify programs' strengths and weaknesses. Naturally, the project also sought to ensure the reliability and validity of the instrument.

The instrument currently contains 10 self-report questions focusing on interest, and engagement of child and adolescent participants in afterschool and summer STEM learning programs. There is a science, engineering and technology version; a math version is in development. The Common Instrument has been administered in programs across 18 different states, with more than 16,000 respondents. Noam briefly overviewed findings (currently under peer review for publication), including the relationship between science interest and program quality. The PEAR team at Harvard is currently working on studies that investigate associations between scores on the instrument and measures of socio-emotional and 21st century skills, program content and quality and demographic

dimensions (e.g., age, gender, ethnicity, location of program). PEAR has also developed with a program quality observation tool, the Dimensions of Success (DoS), which is used widely and can be combined with the Common Instrument and other outcomes measures.

One of the lessons from the PEAR work is the willingness of program staff to collect data when the instrument is short, to the point and generates important information for them and their funders. Programs generally are also willing to let their data be aggregated so that it is possible to make statements about the field as a whole. The Harvard team has also found that when practitioners have access to assessments they begin using them in ways tailored to their needs. This sometimes pushes the work forward in unexpected ways, like adapting an assessment designed for older children for use with younger children. This collaboration between practitioners and researchers is extremely productive and led to the creation of the Common Instrument in the first place.

***Developing Citizen Scientists through Face-to-Face and Networked Learning Opportunities*.** PI Brigid Barron introduced work focused on interest as a catalyst for development. One goal is to identify the "sustaining moves that keep one in the game of learning," i.e., important next steps that grow from educational experiences. Hence the project is studying the opportunities that are provided socially around a learner that enable them to extend and deepen their interest. The project has been tracking interest via case studies to identify multiple opportunities for interest and expertise to develop. A key research question is: What makes someone go from interest to opportunity to engagement?

The project takes an ecological framework approach, with an explicit interest in how one can design ways to intervene across multiple settings over longer scales of time. Funded by the Cyberlearning Program at NSF, the activities take place in the context of a program in Maine called Vital Signs, a citizen science project that monitors invasive species. The program leverages a statewide laptop program, an inquiry-based curriculum, a professional development infrastructure and a focus on network learning.

The project is addressing how to conceptualize and measure participants' interest in the program. Four schools with students varying in SES have been assessed. Items for some of the constructs measured were borrowed from the Activation Lab (see below); others came from other assessments. Findings include:

- Interest in science was not strongly correlated with SES.
- Interest in science did not predict which resources students would use to do school science, but interest was associated with the frequency and range of resources that students would use when they did science for fun.
- The parents of high interest students were teaching about science at home; they were also more likely to be learning about science from their children and were more likely to encourage their children to learn science.
- Interested students were more likely to be noticed as interested by their teachers and more likely to talk to their friends about science.
- Interested students were more likely to report using what they learned in the Vital Signs program outside of the program context (e.g. interested kids were more likely to say that they subsequently looked at plants differently, or noticed species that they had not noticed before).

- There were also some differences in what students in the program projected that they would choose to do in the summer, especially around categories closely linked to Vital Signs' program content.

Barron is interested in exploring the implications of the work for designing more powerful program contexts, particularly around "sustaining moves"—interventions that can provide next steps for students who have been identified as interested and engaged. Baron also is interested in exploring the idea of developing indices of sustaining moves to use as a tool for formative and summative assessment.

Developing, Validating, and Implementing Situated Evaluation Instruments (DEVISE). Evaluation Manager Tina Phillips presented this work designed to develop and promote capacity building for evaluation of citizen science projects through the use of common instruments. The growing field of citizen science projects that were finding it challenging to define and document the learning experiences of their audiences identified the need.

The project began with a literature review that revealed that most existing informal science education assessment work was not appropriate for citizen science projects without modification. DEVISE developed and tested their own instruments for individual outcomes: interest, efficacy, motivation, knowledge of the nature of science, skills of science inquiry, and behavior and stewardship. They have developed different scales for these constructs, which they are in the process of validating.

Much of the validation work has been conducted with a population of bird watchers that the Cornell Lab of Ornithology has ready access to. So far, they have surveyed about 10,000 people on-line as part of the scale-development process. The respondents include school students, but primarily adults who pursue bird watching as a hobby.

Phillips highlighted the self-efficacy construct, made up of eight items. Four of the items are focused on learning and understanding; the other four are focused on the actual doing of scientific or environmental activities. They are currently testing whether this scale can be customized to different topics.

Some of DEVISE's findings so far include:

- Program participants have a variety of connotations for science; some do not identify their interest (bird-watching) as science. Adults tend to define science more broadly, while children tended to report that science is something they do in school.
- Measurement skills are what citizen science projects would most like to see assessed, and these may best be measured through embedded assessments.
- The assessments are currently being developed for pre/post and retrospective use. DEVISE hopes there will be opportunities to customize some of the scales for a variety of content and audiences.

SYNERGIES: Understanding and Connecting STEM Learning in the Community is a 5-year longitudinal study that is supporting a community-wide redesign of a (STEM) learning ecosystem in a diverse, under-resourced community in northeast Portland, Oregon. The goal is to address declining interest in science during early adolescence. Project Coordinator Nancy Staus from Oregon State University described the project's

instrument development process. Community education leaders (broadly defined to include informal and formal educators, parents, youth leaders, and anyone involved in facilitating STEM learning interventions) as well as a team of high school-aged youth researchers were actively involved in both developing and reviewing items. The result is a 10-page instrument to longitudinally track youth interest, participation in STEM-related activities in and outside of school, who encourages participants to do activities, self-efficacy in STEM, and participants' future STEM-related aspirations.

At its inception, the Synergies project identified the need to deconstruct and parse what was meant by STEM interest. Based upon factor analyses of responses to a range of questions related to youth interests, the project was able to discern four content domain categories: earth and space science, human biology, technology/engineering and mathematics. Each participant in the study receives an index score in each of the four areas. From 5th to 6th grade, project researchers found an increase in interest in those indexed areas; but by 7th grade interest had significantly declined across all four domains. However, youth still were interested in earth/space science, human biology, and technology/ engineering, although average math interest was neutral (neither liked nor disliked). In 7th grade, girls were significantly more interested in human biology than boys, and boys were more interested in technology/engineering than girls. There were no gender differences for earth/space science or math interest. STEM interests did not differ based on ethnicity.

These general patterns mask some interesting underlying patterns of STEM interests, however. A cluster analysis revealed three distinct STEM interest patterns for 7th graders: those who 'Like all STEM' (LS), 'Dislike math' (DM), or 'Dislike all STEM' (DS). Youth in LS and DM groups had similar interest in earth/space science, human biology, and technology/engineering but youth in the LS group also liked math, while those in the DM group did not. Furthermore, when 6th-grade STEM interest scores were plotted, the only difference was a significant increase in technology/engineering interest for the LS group. In other words, for 79% of the 7th-graders who also had been in the study in 6th grade, STEM interest remained the same or *increased* over time. The decline in STEM interest observed from 6th to 7th grade for youth as a whole was driven by 25 youth in the DS group who reported a significant decrease in interest for all four STEM dimensions.

Synergies is also doing case studies with a subset of 15 students to understand more about interests and how those interests connect with activities in and out of school. For example, the team has found that most activities decrease from 5th to 6th grade and then fall precipitously by 7th grade. Participants with the lowest STEM interest do significantly fewer out of school activities than do those whose interest has increased or remained stable.

These insights are helping guide intervention strategies geared to students' interest planned for the final year of the project. Adult mentors who share the passion for and have expertise in the participants' interests will lead the activities. The project will also coordinate peer-to-peer relationships, placing children with others of similar interests. Finally, as the name of the project implies, efforts will be made to build synergies community-wide, both those occurring in school as well as those occurring outside of school.

The Science Learning Activation Lab. The Lab expands on recent advances in science education, cognitive psychology, social psychology, and educational psychology, by investigating a new construct called *science learning activation* and a conceptual framework of how it supports science learning. PI Chris Schunn summarized *science learning activation* as a composition of dispositions, practices, and knowledge that enables success in proximal science learning experiences. Lab researchers have identified four dimensions of science learning activation that are predictive *and* can be shaped by designed interventions: fascination with natural and physical phenomenon, valuing science, competency belief in science, and engaging in scientific sensemaking. By success they mean: 1) making choices towards science learning opportunities (often informal in nature); 2) positive cognitive, behavioral, and affective engagement during science learning opportunities; and 3) greater learning. Lab researchers hypothesize that successive iterations of proximal successes in science learning, often experienced in out-of-school learning contexts, generate a feedback loop that propels youth on pathways towards consequential distal outcomes such as: persistent participation in STEM, pursuit of science degrees and careers, and scientific literacy.

In order to test the hypotheses embedded in the above framework the Lab has developed a set of measures that are psychometrically sound (in terms of reliability, validity, and fit), continually improving, and functioning well in the context of research efforts. Included among these instruments:

- *The Science Learning Activation Survey*: The assessment of *science learning activation* includes four scales, each of which demonstrate a high degree of internal reliability (Cronbach's $\alpha > .7$) —fascination ($\alpha = .88$), values ($\alpha = .70$), competency belief ($\alpha = .84$), and scientific sensemaking ($\alpha = .75$)—that parallel the dimensions of *science learning activation*. The assessment takes about 25 minutes to complete.
- *Background Survey:* This instrument enables researchers to collect data related to demographic variables and family resources. It also measures two factors related to prior science learning experiences: (1) prior participation in structured science activities and (2) prior participation in unstructured science activities. Each scale has an internal reliability of 0.80 or greater.
- The self-report *Engagement Survey* asks subjects about their level of affective, behavioral, and cognitive engagement in a particular science learning experience or lesson. It takes subjects about 5 minutes to complete and has an internal reliability of 0.87.

Other measures developed by the Lab that are available and relevant to this study include: choice preference survey, student engagement observation protocols, science learning experience observation protocol, and a learning environment inventory/survey. The Lab has continued to refine the surveys and protocols listed above and has completed multiple studies. So far, the dimensions of activation have been shown to be predictive of *choice preferences* (choosing to participate, attend, and engage in the next opportunity for science learning), *engagement* (including emotional, behavioral, and cognitive components), and *learning* (the student has achieved the learning goals for that particular

science experience). The Lab is now engaged in two NSF-funded projects: *The Activation Approach: A Comprehensive Method and Toolkit for Evaluating the Impact of Science Learning Experiences Across Environments* and *Collaborative Research: Studying the Malleability and Impact of Science Learning Activation.* In addition, Lab researchers are involved in numerous smaller-scale evaluation and design studies that utilize the Lab's framework and measurement instruments and investigate the features of STEM learning experiences that support youth to increase their activation towards STEM learning and experience success in science learning experiences.

Framework for Observing and Categorizing Instructional Strategies (FOCIS). Robert Tai, PI at the University of Virginia, described this project, which is surveying students in grades 3 through 12. An examination of curriculum and programs led to the development of FOCIS, which is a learning activity typology. The typology includes seven activities: collaborating, creating/making, caretaking, teaching, performing, discovering, and competing. A core research question is whether youth who have preferences for particular types of learning activities are more likely to select STEM-related career choices than youth who have different preferences (accounting for demographic characteristics).

For example, the project reported that students who prefer to do discovery activities and making activities, but who do not like to be in collaborative activities, are more likely to say they would choose a STEM career. The project reported this was true for students in elementary, middle, and high school. Implications of this finding for program development might be in designing activities that could shift students' attitudes from neutral to positive levels around collaborating, creating/making, caretaking, teaching, performing, discovering, competing, and collaborating.

Practitioners' Voice

In addition to the research and evaluation projects that participated in the Palo Alto meeting, practitioners from two leading out-of-school time programs provided their perspective on the use of measures.

Techbridge. Executive Director Linda Kekelis described Techbridge as a project for girls ages 5-12 that sees program scale up as an opportunity to introduce assessments. The constructs they are currently evaluating are interest in subject matter, career trajectories and persistence (in STEM). An important question the project seeks to answer is, what's predictive? Techbridge has used the Common Instrument, but they wonder whether it is the right approach for learners of different ages. The project values embedded methods of measuring what program participants are learning. An embedded approach has worked better for the project than a separate assessment process because staff and participants sometimes find surveys disruptive. A narrow tool makes sense for practitioners who need actionable program evaluation and improvement information. Techbridge has diverse program components like field trips and professional role models and they would like to see assessments customized to their needs. Language abilities are also important to Techbridge and so far they are not aware of useful instruments that pay attention to

language differences. Sharing results with immediate audiences is important and their funders are asking them for long-term impact data.

4H- Youth Development Advisor Lynn Schmit-McQuitty characterized the STEM practitioners in 4H programs as needing off-the-shelf, widely applicable, simple evaluation instruments. 4H has found that program staff will begin to develop their own evaluation questions if they see some ready to use examples. They would like access to tools that are easily useable for practitioners in addition to guidance on what to do with findings once they have them. A common tool does not provide context, so they would still have to establish their own framework and specify program outcomes.

The 4H program has also found that there are sometimes disconnects between intent and delivery—a common problem in afterschool STEM programs. At the same time, the program participants often do not see evaluation as critical so staffs are challenged to engage students from the beginning. They share Techbridge's challenge in making room for measurement activities during the course of programming, seeking embedded tools rather than external add-ons.

Critical Friends and Funders

Rounding out the group of researchers, evaluators and practitioners whose work was directly represented at the meeting, a group of "critical friends" with deep experience in ISE research and evaluation were also in attendance; Sue Allen, independent consultant, Leslie Goodyear from the Education Development Center (EDC) and Vera Michalchik from SRI International. Ann Bowers and Ron Ottinger from the Noyce Foundation, and Janet Coffey from the Gordon and Betty Moore Foundation were also in attendance, and Carol Tang from Bechtel and Ellen McCallie from NSF joined part of the meeting remotely. These critical friends were supportive of the goals of the meeting and throughout the process, while also voicing some skepticism about the need and purpose for common measures. Among the related issues that they and the funders noted were:

Value, concerns, and possibilities for collaboration, variation, and commonality

- The objectives and approaches taken by the represented projects vary but there is overlap in constructs and items among these and with other projects not present. How can we continue to collaborate and develop measures while being attentive to each other's instruments in the process? Waiting until the development is complete is probably waiting too long. Explicating what the different models are to the field would be a good idea sooner than later.
- How can we build structures and ground rules for sharing instruments such as, "If I use your instrument, I'll give you my data so that you can continue to validate and develop." We could also create a document about what is harder or easier about developing, using, and interpreting the instruments we are developing. (See the end of this document for resulting matrix.)
- What are the pros and cons of the field going towards a more common vs. a more diverse set of measurements? Would common measures in and of themselves be an inherently positive development for the ISE field? Are there other disciplines that have done this before from which we can draw lessons?

- Perhaps it is best to retain a diversity of approaches to ISE evaluation and research at this stage and not converge on things too quickly. At the moment there is a need for robust studies that are convincing. The educational ecosystem writ large has not yet fully embraced informal, and we cannot expect schools to stand up for us given their own challenges.
- We should do more with what is already known: steps could include figuring out what constitutes legitimacy for the policy audiences we are trying to reach and think about how to talk to them. We should move forward on several fronts concurrently.
- Six CEOs of large youth-serving organizations are discussing collaborating around afterschool STEM and assessment is part of that discussion. The Mott state afterschool network and the Coalition for Science After School are major efforts that are about to sunset. There is some urgency in our conversations at this meeting because these programs have been experimenting and will need easy-to -use instruments soon. Which items will get traction? Which assessments?

Technical issues and field-testing

- Benchmarking would make the measures more useful. Getting them into circulation now, even if they were not totally right, is a good way to get started. This worked well in for the TIMMS and PISA processes.
- Validity is not a one step process—it is an ongoing search that is never complete. It is also a multi-faceted construct that ultimately includes face validity in the eyes of the community.
- What are the fewest items and/or constructs that one can use and still get the findings that are important? Should there ultimately be many instruments that one pulls from to make a composite, customized approach?
- Some of the assessments that are usable will go viral which is great, but there might be some unintended side effects of that, e.g. providing easy access to items could lead to making them *the* de facto evaluation.
- We need to be attentive to the difference between evaluation and assessment. We do not want to send the message that anyone can do evaluation if they have the right tool? Practitioners need to use instruments responsibly. There is a tension in evaluation between expertise and broad participation.

Synthesis

We turn now to a brief synthesis of the convening across the different presentations. We begin with the description of an emerging typology for describing and comparing instruments. We then enumerate a list of take-aways and remaining questions.

Emerging measurement typology

As the meeting participants listened to and discussed the presentations, they began to define a set of dimensions along which to describe the different instruments.

Substance of measures:

- Domain (mathematics, technology, sub-fields of science)
- Constructs measured (interest, motivation, etc.)

Utility of measures:

- Purpose (formative, program evaluation, etc.)
- Settings (museum, school, etc.)
- Unit of analysis
- Age range
- Practicality of usage/intended method
- Connection to design of learning environment
- Use history and benchmark data
- Illustrative findings (what kinds of questions the instrument has been used to answer)

Quality of measures:

- Reliability
- Validity (recognizing that validity has a number of different forms)

This typology served as the foundation of the attached matrix.

Key takeaways and remaining questions

1. Researchers explicitly doing assessment development led the six projects represented at this event. Developing assessments is a complex, messy, and occasionally painful process. Research progresses through peer review, debate, and argument. Much of the meeting consisted of presentations on the results of assessments with critical questions about alternative analyses and feedback challenging interpretations. This is a natural, healthy process essential to ensuring rigor. Even if there is ultimately no convergence or collaboration across measures, it is critical that there is sustained and critical dialog about the measures.
2. Among the projects in attendance, there was a consensus that the areas that the ISE field believes it most directly impacts *can* be assessed. This is an important, positive foundation on which to build future work. There was also a perceived urgency to continue the work: there is an ongoing concern that if we are not able to converge on measurements internally, they will be defined externally.
3. There was an interesting set of debates at the convening about how to define constructs like interest theoretically, but also practically in terms of the questions one

might put on the survey. The discussion served as a reminder that we need to be very specific about how we operationalize constructs and items; it requires us to get beyond loose definitions. The specificity has implications for practice as well. Practitioners need survey items that are sufficiently precise to guide their practice.

4. A similar issue relates to the development cycle. The convening participants encountered productive tensions between iterating through multiple versions of instruments to make them more precise and getting measurements into practice. Some measurements may be more appropriate for large-scale efforts because they are easy and short. Yet such instruments might involve trade offs in terms of theoretical depth, context, and precision of learning theory. Probably a range of assessments is needed to serve different purposes.
5. Important data is longitudinal, which requires repeated assessment with an instrument appropriate to such use. Some projects are doing this in the simplest form of pre-post around an activity; others are structured for following individuals over a number of years.
6. Assessments need to be done in the context of an experience. Accordingly, validity arguments are context bound: one cannot assume an assessment with strong psychometric properties in one setting has similarly strong technical properties in a different setting.
7. Related to the above, assessments are not just capturing out-of-school time experiences, and learning is not just about one experience. This implies the goal of understanding learning across settings and life-wide learning approaches.
8. The community of people doing assessment in informal science education is coming together. As an interdisciplinary community, we should leverage diverse expertise and figure out who has what skills, which instruments rest on which assumptions and connect to which parts of the literature. We need to learn how to work together, share knowledge, and help each other tinker with and further develop our items, instruments, and study designs.
9. It is important for purposes of rigor to empirically test the assumptions we are working from. Here our collaboration can be powerful, because many versions of instruments have to be tested with lots of populations. It is an important part of validity. We need to understand how different groups respond to the same instrument.
10. What would other practitioners or policymakers think if they parachuted into this convening? There is a lot of content to be communicated: assessment development, administration, and analysis. How do we convey insights from this convening to people who need the information? How might communication be tailored to people who work in practice or policy?

Post-meeting follow-up discussions

In the months following the meeting, organizers held three follow-up phone calls on the most pressing issues that emerged: validity, measurement, and the question of what the field will use. The *validity* group discussed the need to figure out a mechanism for sharing valid data and how such a mechanism might be funded, as well as specific validation issues in ISE. They also identified the need for instruments that were particularly useful with regard to validity. The call ended with the plan to create a table of the attending projects that would indicate where they currently were with regard to

validity and identified a few problems in the realm of validity for assessing informal learning outcomes in the future.

The measurement group discussion revisited the idea that the projects that met in Palo Alto were interested in reviewing the assessments being created, and to potentially make some recommendations on how to move the effort to create measurements forward in an intentional way. There were a variety of issues of interest around measurement for those on the call—such as the need for more psychometric investigations, what different measurement ideas mean, and connecting specific items in measurements back to theory—but the question was raised that if a group of learning researchers were to make recommendations about measurement, to whom would they direct those recommendations?

This idea was explored further in the final follow up call, which was framed around the question: What will the field use? A series of recommendations were laid out: for each of the projects to post their instruments on the ATIS site and to have this group characterize the instruments and potentially provide more details about them, and come up with a set of criteria—which could be organized in the form of a decision tree, an online guide, or an actual person providing technical assistance—to help practitioners decide which instrument is best for their work. There was recognition that any of these approaches would require substantial investment from a new funding stream.

This material is based upon work supported by the National Science Foundation under Grant Number 1212803. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

CHARCTERISTICS OF SELECTED INSTRUMENTS FOR ASSESING LEARNING IN INFORMAL SETTINGS

INSTRUMENT	CONTACT	PURPOSE	DOMAIN	CONSTRUCTS	SETTINGS (e.g. school, science center)	AGE RANGE	USAGE (e.g. number of items, time to administer)	FUNDING
"The Common Instrument"	Gil Noam Gil_Noam@hms.harvard.edu	Designed to measure students interest and engagement toward STEM	5 different versions - science, technology, math, retrospective	Interest, Engagement	primarily after school programs; being adapted for use in citizen science	4th grade also adapting adult version	10 question survey	Noyce
Dimensions of Success	Gil Noam Gil_Noam@hms.harvard.edu	Observation tool	Features of the learning environment, activity engagement, STEM knowledge and practices, youth development in STEM	interest, engagement (motivational dimension tied to interest)	summer youth program was pilot		Observation	Noyce
Developing Interest in Science	Brigid Barron Barronbj@stanford.edu	Interests as catalysts (choice of classes) and emergent outcomes (e.g., growth of expertise)	Focus on science and technology; Uses Renninger's scheme across time and environments: community, school, home	Ecological framework - interest, opportunity, engagement that develops expertise	Online - across home, school, community (Vital Signs - Cit Sci invasive species monitoring project in Maine)		Mixed - qual, quan - survey and observation - uses some measure from Activation lab	NSF
DEVISE (Developing Validating and Implementing Situated Evaluations)	Tina Philips Tina.phillips@cornell.edu	Improve evaluation quality and capacity across the field of Citizen Science	6 different domains focused on science and the environment: Interest, efficacy, motivation, knowledge of the nature of science, skills of science inquiry, and behavior/ stewardship	Interest in science and the environment; efficacy; motivation; knowledge of the nature of science; skills of science inquiry; behavior and stewardship	Intended for citizen science projects, but many instruments can be customized for other informal science environments	Intended for adults, but some scales will be adapted for youth audience age 10-17 (Interest, efficacy, motivation, skills)	Varies across scales, but most are between 8 and 24 items and take less than 10 minutes to administer, either on paper or online	NSF and Noyce

CHARCTERISTICS OF SELECTED INSTRUMENTS FOR ASSESING LEARNING IN INFORMAL SETTINGS

INSTRUMENT	CONTACT	PURPOSE	DOMAIN	CONSTRUCTS	SETTINGS (e.g. school, science center)	AGE RANGE	USAGE (e.g. number of items, time to administer)	FUNDING
FOCIS	Robert Tai rht6h@virginia.edu	Are youth who have preferences for particular types of learning activities more likely to select STEM-related career choices than youth who have different preferences?	Discovering; creating/making; collaborating; competing; presenting; caretaking; teaching			Grades 3-12	survey	Noyce
Synergies	Lynn Dierking Dierkinl@science.oregonstate.edu	Understanding and Connecting STEM Learning in the Community	Earth and Space Science; Human Biology; technology and engineering; mathematics	STEM interests; Participation in activities; Encouragement of activities; Science self-efficacy; Future aspirations		5th and 6th grade	survey	Noyce; Bechtel
Activation Lab	Chris Schunn schunn@pitt.edu Rena Dorph redorph@berkeley.edu	Science learning activation=a predictor of success in proximal science learning experiences	Primary version in science; but technology, math, and art activation surveys also exist; also have a version of *emerging activation* in STEM (children ages 6-9) not presented at meeting.	Fascination, values, competency beliefs, scientific sense making	Purposely cross setting (especially in school, after school, camp, science center/children's museum, home)	Designed for 5th and 6th graders, also used with 4th & 7th); pilot versions for 9th,adult, and ages 6-9.	survey and scenario-based assessment; other formats in development	Moore; NSF PRIME; some smaller projects funded by private clients