Observational Methods for Assessment of Informal Science Learning and Education

Drew H. Gitomer
Rutgers University

Observational instruments are receiving broad attention as measures of gauging the quality of interactions within formal educational settings (Bill and Melinda Gates Foundation, 2012). Scores from observations are used as critical pieces of information in the evaluation of teachers for purposes of making employment-related decisions and supporting professional development. Given the potential high-stakes use, a large body of research is emerging to address the validity of observation instruments as assessments of quality (Gitomer & Bell, in press). Using this research base, this paper focuses primarily on the lessons learned from observational methods used in formal education and considers implications for the use of such methods in informal science education.

Observation systems for informal science learning will differ from other formal observation protocols in certain details. However, basic processes and criteria for evaluating the quality of observation protocols and the scores they produce are highly consistent. As is the case for other contexts, best practices of formal and informal learning share many commonalities.

Research has focused on two general classes of observation instruments: those that are subject-specific and those designed for use across all subjects. Subject-specific observation protocols designed for K–12 formal settings have been studied in mathematics (e.g., Hill et al.,2008; Marder et al., 2010), English language arts (e.g., Grossman et al., 2010), and science (e.g., Banilower, 2005; Piburn et al., 2000).

Observation tools have also been developed and researched for informal, out-of-school contexts. As summarized by Yohalem, Wilson-Ahlstrom, Fischer, and Shinn (2007), tools designed for informal settings and instruments used in formal settings share many design

features but differ in fundamental ways. First, almost all systems are designed to support inferences about program quality rather than about specific teachers. Second, instruments are designed to capture qualities of settings unique to informal learning environments.

Instruments reviewed by Yohalem et al (2007) are intended to be applied to out-of-school settings across subject areas. Instruments designed specifically for observing informal settings in science are only now being designed and researched. One such instrument is the Dimensions of Success observation protocol developed by Noam (Noam, Larson, & Dahlmer, 2008).

In addition to highly-structured protocols characteristic of formal education and included in Yohalem et al's review, there is a rich tradition of capturing evidence from interactions that occur in informal settings, particularly in museums. Different techniques have been used to better understand what people are doing when they interact within informal environments. Researchers and evaluators have tried to understand how visitors spend their time at a museum, what exhibits they look at, and how long they spend there. Referred to as tracking and timing, this research (Serrell, 2011) has helped informal learning institutions evaluate and design their environments. The advent of new technologies, particularly radio-frequency identification (RFID), has enabled highly-detailed and accurate data about which exhibits people are attending, and the specific features of those exhibits with which they are interacting (e.g., Baldwin & Kuriakose, 2009; Hsi, 2003). These tracking efforts are particularly suitable to providing measures of engagement and the physical focus of the user within the informal environment.

Such measures do not, however, provide much information about what is being focused on cognitively or what is being learned through these different interactions. In order to probe these issues, researchers have closely studied dialogue among visitors, particularly parents and children as they interact with an exhibit (Ash, 2003; Gleason & Schauble, 1999; Leinhardt,

Crowley & Knutson, 2002). This work not only sheds light on the potential of the exhibit to elicit particular types of discourse between parent and child, but also on how the informal environment helps to mediate sense making by and between parent and child.

Measures of interaction within informal environments certainly have implications for observation protocols that could be used to assess particular aspects informal science endeavors. We consider how prior work in observing formal educational settings has implications for the assessment of informal science settings with a new generation of observation tools.

The Nature of Observation in Informal Science Learning Settings

Observation protocols provide a systematic set of procedures and conceptual frames to evaluate the incidence and quality of specific evidence by assigning scores to interactions within a learning setting. For many protocols, the evidence includes more than the explicit interactions between students and teachers[1]. Evidence can also include instructional artifacts, interviews, and documentation around planning and analysis of the observed interactions.

For formal classroom settings, it is important to differentiate the quality of the setting from the quality of the teacher, as there are many contextual factors, including curriculum, school policies, leadership, and the students themselves, that can also affect the quality of interactions in the classroom. This distinction between *teacher quality* and *teaching quality* (Gitomer & Bell, in press; Kennedy, 2010) is even more salient for informal settings as the *teacher* can take many forms, animate or inanimate. While in many cases there is an identifiable facilitator, in many informal settings, the teacher is embodied in the design of an exhibit, computer simulation, or activity, or can at various times be a docent, interpreter, parent, or peer.

[1] For simplicity of communication, the term *teacher* is used to represent any type of adult facilitator in an informal learning setting. However, in most cases, as explained in the paper, it is more appropriate to refer to *teaching quality* in any assessment of informal programs.

Thus, informal science observation methods should be able to account for a large variety of *teaching* conceptions that scaffold and/or guide learning.

In general, observation protocols used in formal settings have a structure providing the conceptual lens through which one makes judgments of the observed evidence. This structure is typically organized by *domains*, which are the primary constructs, or big ideas, that are the focus of the protocol. Domains are usually defined by a set of *dimensions*, which are the features of the observation that are explicitly scored. Dimension scores are aggregated into domain scores.

Observation protocols currently in use generally adhere to the following process. The protocol begins with an observer developing a record of evidence from the classroom for some defined segment of time, normally without making any evaluative judgments. At the end of the segment, observers use a set of scoring criteria or a rubric that often includes a set of Likert scales (one per dimension) to make judgments based on the record of evidence. These judgments result in numerical scores. Observations of the same classroom are usually made multiple times over the course of a school year.

Judgments vary to the degree that inference is required by an observer, both within and across protocols. Protocols can range from those that ask observers to identify fairly discrete actions by a teacher (e.g., The teacher asks open-ended questions during the lesson) to those that require relatively high levels of inference that take into account not just the teacher action, but also consideration of evidence from students (e.g., The teacher asks questions that promote student thinking and reasoning). Even within protocols, we see that observers are much more able to make reliable judgments of certain readily-observable dimensions (e.g., classroom behavior) than dimensions that require higher levels of inference (e.g., cognitive engagement) (Casabianca, McCaffrey, Gitomer, Bell, & Hamre, 2012).

In general, observers are trained to use a protocol by participating in a training session that lasts a varying number of days. During training observers learn about each of the dimensions, how score points within a scale are defined, and the specific procedures for recording evidence and assigning scores. They then score a set of test cases (typically using video) and are required to become certified observers by assigning scores that approximate the scores assigned by the protocol experts. Many systems will then include periodic calibration test cases to ensure that observers are continuing to assign scores as designed.

As these observation systems move out of research environments to large-scale use in teacher evaluation systems, it is likely that time allocated to training will be dramatically reduced for reasons of cost and the time that likely observers (e.g., school principals) can make available. We can speculate that those dimensions requiring lower inference will be those most readily trained. Indeed, a number of observation protocols for informal settings have involved much briefer training, but they have also tended to focus on lower-inference judgments such as those associated with timing and tracking methods.

Bell, Gitomer, McCaffrey, Hamre, and Qi (in press) provide the following framework, based on the work of Michael Kane (2006), for evaluating the validity of observation protocols that can be applied to the informal learning context in science.

Scoring - To what extent are scores meaningful? Do they have a conceptual basis that is consistent with theory and communities of practice? Do different observers, including expert observers, assign similar scores to an observation? Are there unintended factors that bias scores so as to support inappropriate inferences?

Generalization - Because it is not possible to include all observations, inferences are made about some larger universe (e.g., the quality of a program) based on a sample of

observations. Recent large scale observation studies show that inferences about a classroom can be affected by which lessons were observed, who did the rating, and when in the school year the observation was made (Casabianca et al, 2012). This generalization challenge has led to a consensus view that it is important to observe a setting multiple times and to use multiple observers. Relying on any single observation is likely to lead to a generalization about a program that has very limited support.

Extrapolation - Extrapolation refers to the idea that inferences based on observation are related to a broader conception of quality, whether that be in reference to the teacher, facilitator, school or program. Evaluating the extrapolation inference requires examination of other evidence. For informal learning environments, one could imagine collecting interest surveys from students, monitoring participation in science-related endeavors over some extended period, and even exploring evidence of school-based science participation.

Implication - Scores from observation protocols are often used to support particular kinds of decisions. Plausible uses in informal settings could provide support for decisions about evidence of program efficacy, program funding, and program improvement efforts. Relevant evidence can be used to justify and support these uses. For example, if scores are used for program improvement, is there evidence that, over time, the program actually improves?

<u>What Observations Measure (and what they do not)</u>

Observations can provide powerful evidence of the quality of interactions in learning settings. However, not all evidence about informal science settings is best gathered and evaluated through observation. Alternative approaches should be used to the extent that inferences require speculation rather than direct reference to evidence from the observation.

The Reformed Teaching Observation Protocol (RTOP) and a protocol developed by Horizon Research, Inc. (Weiss, Pasley, Smith, Banilower, Heck, (2003)), both designed for formal settings, provide examples of what can be reasonably observed in formal and informal learning settings. Both protocols require relatively high levels of inference on the observer's part. Examples from each protocol are provided in Appendix A. Each dimension in the protocol refers to something that can be explicitly observed in the lesson. Inferences about things that require speculation are avoided in the observation[2]. For example, evidence about questioning strategies, student responsiveness to questions, and the teacher's presentation of content can all be directly investigated through observation.

Evidence of student learning through observation is much more speculative. One student's response is not necessarily representative of all students, nor is that response likely representative of all of the ideas that are the focus of a lesson. Similarly, student engagement with the lesson is observable—student engagement with science more broadly is speculative and better evaluated through measures such as interviews or records of participation and accomplishment in science activities, both formal and informal.

For informal environments, important aspects of out-of-school learning environments can also be evaluated, some of which are included in formal schooling protocols but receive particular emphasis in the informal context. Dimensions such as creative use of space and materials, focus on youth development, and connections to personal and family life are constructs that can be directly observed. In addition, observations can also be used to illustrate engagement, perseverance, communication, and many other dimensions that characterize a

[2] However, the Horizon instrument does ask the observer, upon completion of the direct observation, to speculate on how the observed lesson would support student learning more generally.

person's experience. Of course, the dimensions (and variations on those dimensions) that make up the Horizon and RTOP instruments are also candidates for informal schemes.

The inherent limitation in studying learning through observation of informal settings is well described in a previous National Academy report on informal education (National Research Council, 2009). While it is inadvisable to use observations as a measure of student learning or other outcomes (e.g., long-term interest in science), it is possible to identify interactions that have been demonstrated through other research to be associated with desired outcomes. Thus, evidence that can be legitimately observed can be used to provide measures of constructs such as level of engagement. Studies can then be done to identify the relationship of level of engagement with outcomes of interest that are better measured through other means. Such measures might include surveys or direct investigations of learning through interviews.

No instrument or measure can carry the entire burden of answering questions of program quality. At the risk of stating the obvious, observations are appropriate for investigating those aspects of instruction that can be observed. We can observe whether students are engaged; however, we do not know their interest in science unless we ask them or include other indicators of science interest. Observations can help us identify whether children are engaging in inquiry, but we do not have a good grasp of whether they understand inquiry without much more direct investigation of their understanding of inquiry. That said, programs that strongly engage students in aspects of inquiry are more likely to have desired long-term outcomes that can be assessed through other means.

Ensuring Quality Observations and Avoiding Common Mistakes

Observation protocols require integrated judgments of relatively complex phenomena. To ensure that judgments are reliable, accurate, and meaningful, protocol design and implementation should include the following:

1. *Clear and coherent rubric design* - For every dimension scale in the protocol, there should be a clear definition of the dimension, clarity about what constitutes relevant evidence, and clear distinctions among score points. Score point distinctions should be consistent such that scores have as similar meaning as possible across all dimensions. Common mistakes include murky language that does not carry the intended meaning to an observer using the protocol (e.g., a score point of 2 saying that x *is observed rarely* and a score point of 3 saying that x *is observed infrequently*); dimensions that overlap substantially such that the same evidence contributes to scores on multiple dimensions; and inconsistent use of score points such that a 3 on one dimension represents an exceptional performance while a 3 on another dimension is only pretty good.
2. *Effective training and quality control* - Observers need to be trained to make judgments on the basis of the protocol, not based on their own preferences or beliefs, either consciously or unconsciously. Observers must be taught to justify their scores in terms of the observed evidence and how it relates to characterizations of performance described in the protocol. Common mistakes include observers not understanding and internalizing the meaning of dimensions and score points within those dimensions, as well as allowing personal attitudes about teaching, learning, science, or informal environments to influence judgments in idiosyncratic ways inconsistent with the protocol. Another very common mistake is one in which observers first make an overall judgment (e.g., This is a good program) and then use that judgment to justify a set of scores. Similarly, poorly

trained observers may simply assign very similar scores across all dimensions on the basis of an overall impression. Even with quality training, it is important to continue monitoring the quality of scores to ensure that some of these problems in scoring do not surface or re-surface over time.

3. *Multiple observations with multiple observers* – Even with the best training and the most refined scoring processes, observers will vary in their judgment and quality of the program will vary day to day. Therefore, if the gal is to make judgments about a program or individual, it is important to observe a setting multiple times and use multiple observers. In that way, if there is measurement error, there is a higher likelihood that the use of multiple observations and multiple observers will cancel out some of these sources of error.

Closing Comment

Ensuring quality in formal settings is very challenging and presents difficulties to states and districts as they implement new evaluation systems. Given available resources, the challenges for informal science settings will be even more daunting. However, achieving quality is not simply a matter of satisfying the measurement and evaluation community. The ability to make reliable and valid judgments about programs is an indicator of a shared understanding of what constitutes quality in informal science and an ability to clearly articulate the meaning of quality. If informal science educators are to improve their programs and communicate and advocate for the importance of their work, then the field will need to not only clarify the meaning of quality, but be able to point to clear demonstrations of effective informal science education.

References

Ash, D. (2003). Dialogic inquiry in life science conversations of family Groups in a museum. *Journal of Research in Science Teaching, 40*(2), 138–162.

Baldwin, T., & Kuriakose, L. T. (2009). Cheap, accurate RFID tracking of museum visitors for personalized content delivery. In J. Trant & D. Bearman (Eds.), *Museums and the web 2009: Proceedings*. Toronto, Canada: Archives & Museum Informatics. Retrieved April 17, 2012, http://www.archimuse.com/mw2009/papers/baldwin/baldwin.html

Banilower, E. R. (2005). *A study of the predictive validity of the LSC Classroom Observation Protocol*. Arlington, VA: National Science Foundation.=

Bell, C. A., Gitomer, D. H., McCaffrey, D., Hamre, B., Pianta, R., & Qi, Y. (in press). An argument approach to observation protocol validity. To appear in *Educational Assessment.*

Bill and Melinda Gates Foundation. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Washington, DC: Author.

Casabianca, J., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., & Hamre, B. (2012). *Effect of observation mode on measures of secondary mathematics teaching.* Unpublished manuscript.

Gitomer, D. H. & Bell, C. A. (in press). Evaluating teaching and teachers. To appear in K. F. Geisinger (Ed.), *APA Handbook of Testing and Assessment in Psychology.*

Gleason, M. E., & Schauble, L. (1999). Parents' assistance of their children's scientific reasoning. *Cognition and Instruction, 17*(4), 343–378.

Grossman, P. L., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J. H., Boyd, D. J., & Lankford, H. (2010). *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores*. Washington, DC: National Center for Analysis of Longitudinal Data in Educational Research.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L. & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition & Instruction, 26,* 430–511.

Hsi, S. (2003). A study of user experiences mediated by nomadic web content in a museum. *Journal of Computer Assisted Learning, 19,* 308–319.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17–64). New York, NY: Praeger.

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher, 39*(8), 591–598.

Leinhardt, G., Crowley, K., & Knutson, K. (2002). *Learning conversations in museums*. Mahwah, NJ: Lawrence Erlbaum.

Marder, M., Walkington, C., Abraham, L., Allen, K., Arora, P., Daniels, M., Dickinson, G., Ekberg, D., Gordon, J., Ihorn, S., & Walker, M. (2010). *The UTeach Observation Protocol (UTOP) Training Guide* (adapted for video observation ratings). UTeach Natural Sciences, University of Texas Austin.

National Research Council. (2009). *Learning science in informal environments: People, places, and pursuits*. Washington, DC: The National Academies Press.

Noam, G., Larson, J, Dahlmer, C. (2008). *Dimensions of success observation tool*. Belmont, MA: Program in Education, Afterschool & Resiliency (PEAR), Harvard University.

Piburn, M., Sawada, D., Turley, J., Falconer, K., Benford, R., Bloom, I., & Judson, E. (2000). *Reformed teaching observation protocol (RTOP): Reference manual* (Technical Report No. IN00-3). Tempe, AZ: Arizona Collaborative for Excellence in the Preparation of Teachers.

Serrell, B. (2011). Paying more attention to "Paying attention." *Center for Advancement of Informal Science Education, 236*(3), R221–R224. Retrieved from http://ajpregu.physiology.org/cgi/content/abstract/236/3/R221

Weiss, I. R., Pasley, J. D., Smith, P. S., Banilower, E. R., Heck, D. J. (2003). *Looking inside the classroom: A study of K–12 mathematics and science education in the United States*. Chapel Hill, NC: Horizon Research, Inc.

Yohalem, N., Wilson-Ahlstrom, A., Fischer, S., & Shinn, M. (2007). *Measuring youth program quality: A guide to assessment tools* (1st ed.). Washington, DC: The Forum for Youth Investment, Impact Strategies, Inc.

Appendix A

Brief Synopsis of the Reformed Teaching Observation Protocol (RTOP)

RTOP still consists of 25 items divided into three subsets: Lesson Design and Implementation (5); Content (10); and Classroom Culture (10). The second and third subsets are each divided into two smaller groups of five items. The first subset was designed to capture what had become the Arizona Collaborative for Excellence in the Preparation of Teachers (ACEPT) model for reformed teaching. It describes a lesson that begins with recognition of students' prior knowledge and preconceptions, that attempts to engage students as members of a learning community, that values a variety of solutions to problems, and that often takes its direction from ideas generated by students. The second subset was directed at content and was divided into two parts. The first assessed the quality of the content of the lesson, and the second attempted to capture the ACEPT understanding of the process of inquiry. The final subset, consisting of ten items, was directed at the climate of the classroom. It was the authors' intention to capture the full range of ACEPT reformed teaching with these 25 items.

Example items from each of the three domains in RTOP:

III. LESSON DESIGN AND IMPLEMENTATION

	Never Occurred Very Descriptive
1) The instructional strategies and activities respected students' prior knowledge and the preconceptions inherent therein.	0 1 2 3 4
2) The lesson was designed to engage students as members of a learning community.	0 1 2 3 4

IV. CONTENT

Propositional knowledge	Never Occurred Very Descriptive
6) The lesson involved fundamental concepts of the subject.	0 1 2 3 4

Procedural knowledge	Never Occurred Very Descriptive
11) Students used a variety of means (models, drawings, graphs, concrete materials, manipulatives, etc.) to represent phenomena.	0 1 2 3 4

V. CLASSROOM CULTURE

Communicative Interactions	Never Occurred Very Descriptive
16) Students were involved in the communication of their ideas to others using a variety of means and media.	0 1 2 3 4

Student/Teacher Relationships	Never Occurred Very Descriptive
21) Active participation of students was encouraged and valued.	0 1 2 3 4

Brief Synopsis of the Horizon Research Inc. Observation Protocol

This protocol was developed as part of a large research study of science and mathematics classrooms and is part of a more comprehensive data collection that included multiple instructional artifacts, curriculum materials, and interviews with the teacher. Only the observation protocol is described in this section.

Researchers observed lessons in each of four component areas: the lesson design; its implementation; the mathematics/science content; and the classroom culture. In each case, the researcher first rated the extent to which the lesson exhibited each of a number of characteristics of high quality instruction. For example, in the case of mathematics/science content, the observer rated the extent to which the content was significant and worthwhile and the extent to which teacher-presented information was accurate; among other indicators.

After rating the individual indicators in a component area, the researcher was asked to provide a "synthesis rating" on a five-point scale, where 5 indicated the lesson was extremely reflective of current standards for mathematics/science education. The researcher was then asked to provide a brief description of the nature and quality of that particular component of the lesson, and to provide the rationale for the synthesis rating and evidence to support it, including examples/quotes illustrating the ratings of particular "focus indicators."

Example items include:

I. Design

A. Ratings of Key Indicators

	Not at all				To a great extent	Don't know	N/A
1. The design of the lesson incorporated tasks, roles, and interactions consistent with investigative mathematics/science.	1	2	3	4	5	6	7
2. The design of the lesson reflected careful planning and organization.	1	2	3	4	5	6	7

II. Implementation

A. Ratings of Key Indicators

	Not at all				To a great extent	Don't know	N/A
1. The instructional strategies were consistent with investigative mathematics/science.	1	2	3	4	5	6	7
2. The teacher appeared confident in his/her ability to teacher mathematics/science.	1	2	3	4	5	6	7

III. Mathematics/Science Content

A. Ratings of Key Indicators

	Not at all				To a great extent	Don't know	N/A
1. The mathematics/science content was significant and worthwhile.	1	2	3	4	5	6	7
2. The mathematics/science content was appropriate for the developmental leves of the students in this class.	1	2	3	4	5	6	7

IV. Classroom Culture

A. Ratings of Key Indicators

	Not at all				To a great extent	Don't know	N/A
1. Active participation of all was encouraged and valued.	1	2	3	4	5	6	7
2. There was a climate of respect for students' ideas, questions, and contributions.	1	2	3	4	5	6	7