Measuring Participant Engagement in Short-Duration Citizen Science Events

Bill Zoellick*[1], Hannah Webber[1], Karen James[2], Abraham Miller-Rushing[3], Michael Marion[2]

Collecting and analyzing observations of participants in citizen science programs is an important part of understanding and improving participant experiences and outcomes. This article describes an approach to designing and analyzing observation frameworks that has the potential to result in more useful, interval-scale estimates of the magnitude of theoretical constructs such as participant engagement, rather than simple counts of different kinds of behaviors. Specifically, the article explores use of a Rasch partial credit model in estimating participant engagement. Applying the model to two different sets of participant observations, it suggests that Rasch analysis has the potential to provide researchers and evaluators not only with statistically useful estimates, but also with insights into the performance of observation frameworks that can lead to framework improvement. The exploratory work described here suggests that larger scale testing and application of Rasch models to observations of citizen scientists is likely to be useful. It closes by outlining a program for the development and refinement of citizen science observation frameworks.

This paper explores a methodological problem. The problem emerged in preparing to investigate a conjecture about citizen science programs that are of relatively short duration and that take place in recreational settings.

The methodological problem arises as we seek to measure the degree to which participants are "engaged" in the citizen science work. As we explain later in more detail, collecting data about engagement in short duration events presents challenges that necessitate increased reliance on observations of participant behaviors, as opposed to collecting data through interviews and surveys.

Using observation data to construct inferences about the degree of participant engagement requires researchers to answer two measurement questions: (1) “Which behaviors should be recorded?” and (2) “What values should be assigned to these behaviors, corresponding to more or less engagement?” Taken together, the answers of these two questions comprise an "observation framework" that is used as a measurement tool in collecting data about engagement.

The methodological problem revolves around answering these two measurement questions in a way that considers them jointly and iteratively, using sets of recorded behaviors and initial assumptions about assigning values to behaviors to improve the measurement tool over time. A related problem is that of constructing the observation framework so that it provides interval scale measurements. Interval scale measurements are like those that we get from a thermometer, where the difference between 2 degrees and 4 degrees is the same amount of change as the difference between 4 degrees and 6 degrees. It would be very useful to have a way to measure engagement that produced interval scale values so that they can be added, averaged, and used in statistical tests to compare different program designs.

Describing the exploration of a methodological problem is not as straightforward as presenting the results of research using established methods. Rather than proceeding from research question to design to results and conclusions, this kind of paper is more tentative and, well … exploratory. Consequently, a brief overview of the structure of what follows will be useful for many readers.

We begin by describing why addressing this methodological question is potentially useful to any citizen science research that uses participant observation as an important source of data. We also provide a brief description of the research context in which this question emerged.

[1] Schoodic Institute at Acadia National Park; [2] MDI Biological Laboratory; [3] National Park Service

* Contact for corresponding author: bzoellick@schoodicinstitute.org

Zoellick, B., H. Webber, K. James, A. Miller-Rushing, M. Marion. (2015). *Measuring Participant Engagement in Short-Duration Citizen Science Events*. Paper presented at the Citizen Science Association Conference, San Jose, CA, Feb. 11-12, 2015.

Next, we look at an example of a simple observation framework that we adapted from research into the behaviors that people exhibit as they interact with exhibits in a science museum, which is another setting in which it is useful to understand participant engagement over relatively short time periods.

We then borrow an approach to measurement from psychometrics and education research known as the Rasch model in order to transform the results of the observations into interval scale data that we can graph and study. We describe the Rasch Model in enough detail to enable readers who are not familiar with it to follow as we move to the next step, in which we apply the Rasch model to two sets of observations of citizen science participants collected in different contexts.

We conclude with some thoughts about the potential value of what we learned through this exploration, along with the limitation of what we present here. Our goal is to describe what we learned with enough clarity that others might try the ideas presented here and either improve them or propose something better.

Motivation for the Work

The work that we describe here is part of research undertaken by MDI Biological Laboratory, Schoodic Institute, and the National Park Service that seeks better understanding of new citizen science opportunities that might be enabled by DNA barcoding, a technique that uses short DNA sequences to aid in species identification. The goal of this pilot-stage research is to identify, develop, and test techniques that might enable engagement in biodiversity-related citizen science by large numbers of park visitors and other people in similar settings. These pilot studies engaged participants in a variety of different activities, including sample collection, sorting of species in a lab, and work with computers to clean up and look up DNA sequences in online repositories such as the Barcode of Life Data System (BOLD).

The broader study explores several conjectures that reach beyond the work reported in this paper. One is that DNA barcoding can accurately identify invertebrate species of interest to biodiversity studies in the environs of Acadia National Park. A second conjecture is that having routine access to reliable species identification techniques will enable a broader range of citizen science activities, including activities that only require a few hours of a participant's time as they collect or sort samples. A third conjecture is that these shorter duration citizen science activities will enable engagement by participants who do not think of themselves as being deeply interested in science, but who are open to trying something new that provides a useful service.

It is this last conjecture that focused our attention on the problem of making sense of observations of participants in short-duration citizen science programs, which is the subject of this paper.

The Short Duration Problem

The approaches available to researchers describing participant engagement in citizen science depend on the nature of the citizen science activity. If a project engages participants over time periods spanning from days to weeks to months, interviews with participants are an excellent and often a preferred means to gain insight into participant experiences. If there are large numbers of participants engaged over such time periods, it often makes sense to use survey instruments to augment data collected through interviews.

However, not all citizen science work extends over periods of days, weeks, and months. Activities such as biodiversity inventories and other kinds of collection and sampling activities might involve participants for only a few hours. These kinds of sampling, data collection, and science support activities can be particularly attractive and useful in settings such as national parks, nature preserves, and other settings where people visit for short time periods and may be open to engagement in activities that they know are useful to the place and that offer opportunities for learning in addition to service.

Short duration events present a number of challenges to research or evaluation efforts. Since participants may only be involved for a few hours, asking them to complete a survey or interview can seem like a significant intrusion. Further, since participants in these free-choice activities are often on vacation or partic-

ipating in recreational activities, being asked to complete surveys and interviews may be inconsistent with their motivation to participate in the activity.

The research and evaluation challenges that emerge in short-duration citizen science programs are not unique to citizen science. Researchers working in museum settings have confronted such problems for years. For example, Palmquist and Crowley (2007), in their article describing research on how parents and children interact with each other as they visit dinosaur exhibits, begin with a funny and sometimes painful account of the challenges encountered in presenting parents with an informed consent form to read and sign when they are holding the hand of a six-year old who is only 50 feet away from seeing the dinosaurs. In addressing such challenges, researchers and evaluators studying visitor engagement with museum exhibits often rely heavily on observing visitors, rather than interacting in more direct and intrusive ways.

Observation Frameworks

As Yogi Berra said, "You can observe a lot by watching." In the context of learning about participant engagement, observing a lot by watching implies having some way to organize what one sees as participants engage with a program or museum exhibit. Stated more formally, one needs an observation framework that directs attention to specific behaviors and that specifies how these behaviors are to be counted, aggregated or otherwise converted from simple counts into an assertion about what was happening over the course of the participation.

Observation frameworks can be loosely classified as either "high-inference" or "low-inference." High-inference frameworks provide guidance to researchers or evaluators with regard to evidence to be collected and considered, but analysis of the evidence and judgments about quality depend on the expertise of the observer. Horizon Research's *Inside the Classroom Observation and Analytic Protocol* (Horizon Research, 2000) is an example of a widely-used high-inference observation framework. It contains scales that the evaluator uses to rate the quality of lesson design, lesson implementation, lesson content, classroom culture, and a number of other characteristics. The final step in completing the protocol requires the evaluator to provide an overall "capsule rating" of the lesson, drawing on his or her experience in observing many lessons to render that judgment. High-inference frameworks can work well in a mature research area where there is substantial agreement about what constitutes quality.

Low-inference frameworks focus more on collecting data that are more directly connected to observation. They typically require the observer to record the presence or absence of certain behaviors repeatedly, at some fixed time interval. This kind of data collection still depends on inferences, but the inferences are implicit and embodied in the protocol designer's selection of behaviors that are assumed to be important and worth recording. The inference-making is done by the framework's designer, rather than by the observer who is on scene.

Because the low-inference observations collect data at a lower level of abstraction, recording behaviors rather than judgments about behaviors, they can be more useful and flexible in research that is trying to understand how things work, as opposed to making quality judgments about practices that are well-understood. This makes low-inference observation frameworks potentially attractive in relatively young research areas such as inquiry into the structure and effect of citizen science activities.

Adapting an Observation Framework for Use with Short-Duration Citizen Science

Chantal Barriault, working with David Pearson (Barriault & Pearson, 2010), developed an observation framework that serves as a good example of a low-inference framework as well as a starting point for the observation work described here. Working from many hours of observation as visitors engaged in activities in a science center, she theorized that visitors exhibit a sequence of three categories of behaviors as they interact with an exhibit. They begin with what she called Initiation activities consisting of behaviors such as looking at the exhibit, watching others interacting with the exhibit, and trying out the exhibit

themselves. If the visitors are interested enough to continue interacting with the exhibit, they move to Transition behaviors that might include repeating the activity or expressing positive emotions in reaction to engagement. In some instances visitors move beyond the Transition behaviors to Breakthrough behaviors that might include referring to past experiences while they engage in the activity, seeking more information about the activity, sharing information with others, and engaging in inquisitive exploration about the activity through experimenting with it and trying out different kinds of actions. This pattern of engagement development is illustrated in Figure 1. Barriault and Pearson note that in their experience with using their framework in observing many visitors, not all visits fall into this pattern, but most do.

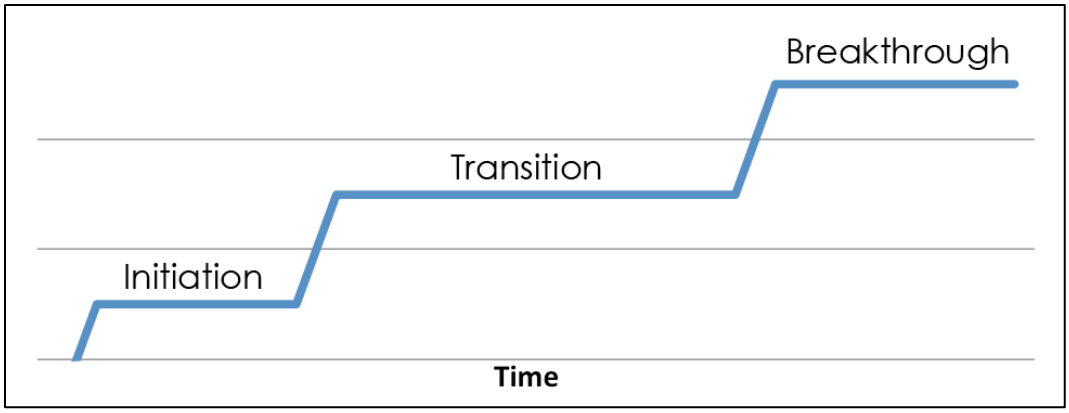

Figure 1. Expected pattern of behaviors in observation framework developed by Barriault and Pearson (2010).

Barriault and Pearson developed this framework in order to make decisions about exhibit design. Consequently, they were primarily interested in ensuring that an exhibit elicited a sufficient number of behaviors in the Transition or Breakthrough behavior categories. If observations revealed that visitors were only engaging in Initiation behaviors when interacting with an exhibit, that would suggest that the exhibit needed to be revised in some way.

We recognized that Barriault and Pearson's observation framework might be adapted to address the problem of making sense of observations of citizen science participants. It seemed reasonable to conjecture that participants would begin by focusing on mastery of the protocols for collecting, sorting, processing, or citizen science activities, and then, having mastered the fundamentals, would move on to higher levels of personal engagement with the activity.

Initial Shaping of An Observation Framework: What Behaviors Do We Record?

One of the key requirements for an observation framework is that it takes account of all the behaviors that seem meaningful and important. This is not just analytical work; it is also deeply empirical. If we are measuring engagement, one has to see what visitors or participants actually do, making sure that behaviors that appear to be important are recorded and included in the final assessment of engagement. Barriault (Barriault, 2014) provided a good example of this kind of work as she described the steps she took to adapt her initial observation framework, developed for use in science centers, for use in zoos and aquariums, where visitors were interacting with living animals rather than with more predictable exhibits that are designed by museum staff.

We used close observation of six citizen scientists who volunteered to assist in specimen sorting to begin the task of identifying the behaviors to include in an observation framework for the kinds of citizen science lab activities that might be part of biodiversity research programs. These six individuals provided informed consent to participate in human subject research, which meant that we could collect a variety of kinds of information, including interviews, that enabled us to connect the behaviors that we were observing with the participant's own perceptions of what they were doing. The laboratory tasks that the volun-

teers undertook included using dissecting microscopes to scan eelgrass blades to find and remove invertebrates living on the eelgrass. Then, working with the assistance of a taxonomist and from photographs of common invertebrates, the citizen scientists attempted to identify the animals that they found and placed them in containers for subsequent DNA analysis. We made observations at five-minute intervals over a period of 90 minutes.

One realization that emerged quickly was that rather than the simple progression from Initiation through Transition to Breakthrough illustrated in Figure 1, we were witnessing something more complicated. We did generally see evidence of Initiation, but the initiation was followed not by a Transition stage, but instead by a lot of "Doing" of the assigned task. Over the course of the Doing the participants periodically engaged in Breakthrough behaviors, but went back from Breakthrough to the ongoing "Doing" that was the primary focus of the lab work. Figure 2 illustrates this pattern.

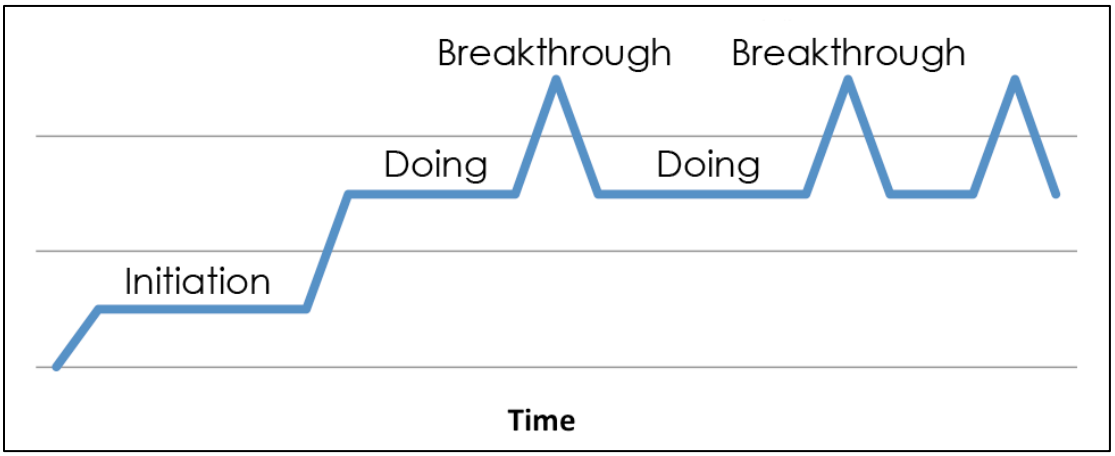

Figure 2. Pattern of behaviors observed in citizen science laboratory work.

Rather than a progression to successively higher levels of engagement in which the middle level is a Transition to something else, the middle level of engagement in this laboratory work appeared to be a kind of "base" state. Participant engagement might periodically rise above this base state, but generally returned to it within the five-minute observation interval. In retrospect, this finding makes intuitive sense: museum visits are primarily free choice activities in which visitors engage with an exhibit as long as it holds their attention, but then move on to something else. Citizen science activities, on the other hand, invite participant commitment to doing some work in support of research over a period of time that might extend for an hour, a day, or years. A participant's level of interest might wax or wane over this time period, but the work goes on.

Conversations with participants as they engaged in searching for invertebrates, sorting them, and preserving them contributed to a categorization of different behaviors as being either Initiation, Doing, or Breakthrough behaviors. For example, we were interested to note that participants took out smart phones to take pictures through their microscopes of what they were seeing. After talking with them about what they were doing and why, we decided to categorize this particular behavior as Breakthrough because it provided evidence of deeper participant engagement beyond the routine Doing of looking, sorting, and preserving. Table 1 provides a complete list of the behaviors that we recorded, together with the scheme for categorizing them as Initiation, Doing, or Breakthrough behaviors.

Table 1. Categorization of behaviors observed during citizen science laboratory activities

Initiation Behaviors	
Ask	Participant is asking questions about how to do things, about process, about protocols – seeking how to begin work
Try	Participant is interacting with equipment, tools, procedures in a tentative way to learn how things work
Watch	Participant is watching others use equipment, tools, procedures in a tentative way to learn how things work
Doing Behaviors	
Doing	Participant is engaged in intended task
Commenting	Participant is commenting on the process or task as he or she learns it. (e.g., "the focus knob is very delicate," "these things won't hold still")
Waiting	Participant is waiting for a resource, assistance, identification, etc.
Seeking Help	Participant is asking for assistance, clarification, additional information to improve at task
Breakthrough Behaviors	
Providing Help	Participant is providing assistance or information to another participant
Sharing	Participant is sharing information / insights / excitement with others ("Come look at this!") NOTE: "Sharing" is not the same as "chatter" that happens in the course of doing the activity. Sharing is related to something perceived, discovered, noticed, seen, and so on ... related to the work. Chatter and small talk in the course of doing the activity is just "Doing."
Pictures	Participant is taking pictures related to the task or subject matter (e.g., pictures of things collected, pictures of other participants, pictures of setting) NOTE: try to describe what is being photographed
Reference to past work	Participant is relating this experience to something he or she has done before. (This is a more specific kind of sharing … record as "R" rather than "S")
Inquiry	Participant is involved in manipulation of equipment, comparing of samples, and so on to achieve better understanding of the topic at hand.

Interpreting the Patterns

As Figure 1 and Figure 2 suggest, one approach to making sense of these observation data is to look at how the character of the behaviors changes over time. Figure 3 shows how the frequency of the different categories of behaviors changed over the hour and a half of work in the lab as participants inspected eelgrass for invertebrates. Initiation activities decreased over the course of the first half hour (up to interval T6). During this time the participants became familiar with using microscopes, sought more specific instruction, watched other participants, and waited for help. After interval T6, most of the participants were engaged in the intended task. After about an hour of work (at interval T12), one of the participants found a hydrozoan and another found a ribbon worm. This resulted in Breakthrough behaviors including sharing and taking pictures of the animals. These behaviors persisted for several more time intervals.

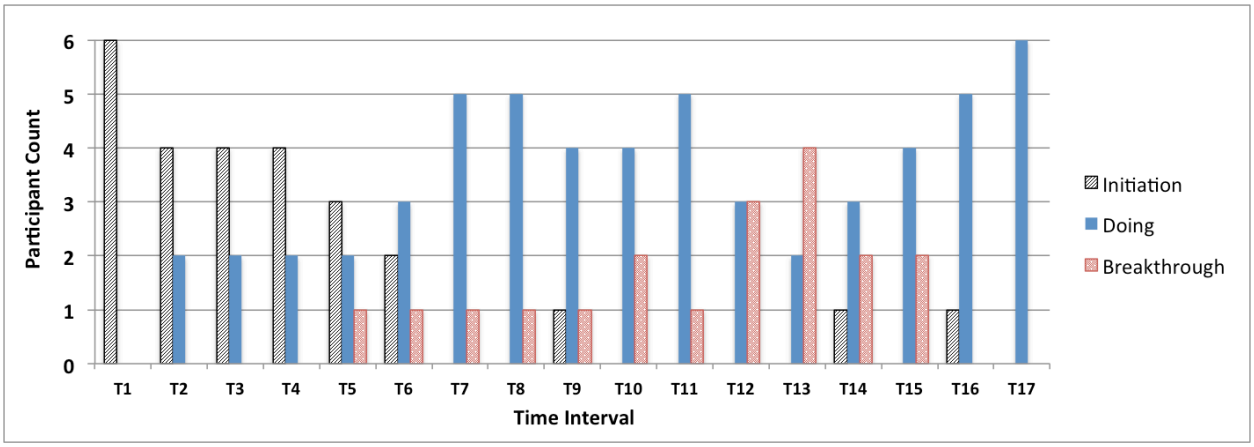

Figure 3. Frequency counts of different categories of behaviors during eelgrass lab activities.

The Rasch Model as a Way to Move Beyond Frequency Counts

As Figure 3 illustrates, looking at changes in frequency counts for different behavior categories over time helps make sense of observations. For studies where there are only a small number of participants, looking at patterns of frequency counts may be the most powerful analysis that the data can support. But studies that collect data about 50-100 participants, or more, have access to more powerful options.

The problem with frequency counts of different categories of behaviors is that they are counts of different kinds of things that are not related in any way that is obviously measurable. For example, in the simple observation framework presented above, we have defined Doing and Breakthrough behaviors as categorically different, but have no easy way to use the data we have collected to refine this definition. Perhaps some of the behaviors now labeled as Breakthrough should really be considered as Doing, or *vice versa*. It would be useful to have some way to put these different categories on a single scale that represented different degrees of engagement. Having such a scale would allow us to use statistical methods to explore such questions as, "Do the participants who are generally exhibiting higher levels of overall engagement consistently perform behaviors that we have placed within the Breakthrough category?" If the answer to such a question was "No," we might then consider revising our observation framework. Access to methods that would support iterative improvement of an observation framework as it was used with hundreds of participants across multiple studies would be beneficial to citizen science participant research as a whole, in addition to being useful within individual studies.

An additional benefit of being able to think of engagement as a scale of values, as opposed to counts of different categories of behaviors, is that it would support research questions that involved using inferential statistics to make comparisons between different sets of observations related to different program designs or different groups of participants.

The Rasch Model

Over the past four decades, psychometric research has made increasing use a model developed by Georg Rasch (Rasch, 1980) to address this kind of scaling problem. The Rasch Model treats the probability of responding correctly to a test item as a function of both the item difficulty and the ability of respondents who answer it correctly. Rather than assigning the same value to all items in computing a score (e.g,, a value of 1 for a correct answer and 0 for a wrong answer), it computes a different difficulty level for each item. Computation of this difficulty metric reflects the overall scores of those who answer the item. So, a more difficult item is one that more capable respondents can usually answer and that is challenging for less capable respondents. Clearly, there is some circularity in this definition of difficulty, since we identify the more capable respondents on the basis of their scores, and those scores depend on the item difficul-

ties, which, in turn, depend on knowing who is more capable. Consequently, producing a list of item difficulties is an optimization problem that is solved iteratively on a computer, over a series of successive approximations. But the modeling process converges for sufficiently large numbers of items and respondents. One of the important and useful features of the Rasch item difficulty levels is that they are interval scale measurements. In other words, the amount of ability, engagement, or whatever else we are measuring to move from a level of 1 to 2 is the same as the amount required to move from 2 to 3.

Rasch's original model worked only when answers were simply right or wrong, but it has now been extended to work with "partial credit models" in which the answer to an item can have a range of values that increase as answers demonstrate deeper understanding or, in our case, deeper engagement. Bond and Fox (2012) provide a good overview of Rasch models, including partial credit models, for readers seeking more detail.

The Rasch partial credit model is potentially applicable to the problem of assigning more meaningful and useful values to the behaviors observed during citizen science programs. We use the word "potentially" because the Rasch model requires adherence to some assumptions about the thing being measured. Specifically, it assumes that performance and ability are related along a single dimension. This assumption would be violated if we were measuring two different things at once, for instance speed and strength. Some highly capable respondents would do well on some difficult probes, but poorly on others. Rasch modeling would not fit that kind of measurement tool.

For the measurement problem explored here, meeting Rasch model requirements would mean that engagement, as defined by the behaviors in the observation framework, should emerge as a single construct. If we find that we can order participants in a consistent way from least engaged to most engaged and if we find that this ordering correlates with the degree to which engagement is more challenging at different times, we will have a measure for which Rasch modeling might be useful. On the other hand, if we uncover a more complicated relationship between participants and engagement, the Rasch model will not fit and—importantly—the Rasch analysis will tell us that. In short, trying to fit the Rasch model to our observation data can contribute to a better understanding of the construct that we are trying to measure.

Applying the Rasch Model to Samples of Citizen Science Observations

The observations of the six participants as they found and sorted invertebrates living on eelgrass gave us the first draft of an observation framework that, with refinement, might be useful in measuring engagement. We were now in a position to try the framework out on some larger groups of participants to see if we could use the Rasch model to begin a process of iteratively refining the framework. We collected two sets of observation data in two different settings.

Observing and Measuring Engagement During Computer Lab Work

One set of observation data was collected during a computer lab workshop in which participants worked to clean up digital DNA barcode records produced from local lab work. Once they had a clean record that could be used in a search against the BOLD barcode repository, they looked for a match on the record they had cleaned up. The workshop leader made it clear to participants that she was primarily interested in getting feedback from participants about how to organize this kind of activity so that it might be done remotely by other citizen science participants.

The workshop involved 24 participants for an hour. Behavior was observed every five minutes using the observation framework in Table 1. Figure 4 summarizes the frequency distribution of different classes of behavior over the course of the hour.

The Rasch model assumes that a measurement instrument consists of a set of items, or probes, that are each scored separately. In applying the model to this set of observations, we treated each time interval as a separate probe and looked at the distribution of behaviors for each interval. Rasch analysis software, by

convention, uses a value of 0 for the lowest category of response. Consequently, we mapped the different categories of observed behaviors, Initiation, Doing, and Breakthrough, to category numbers of 0, 1, and 2.

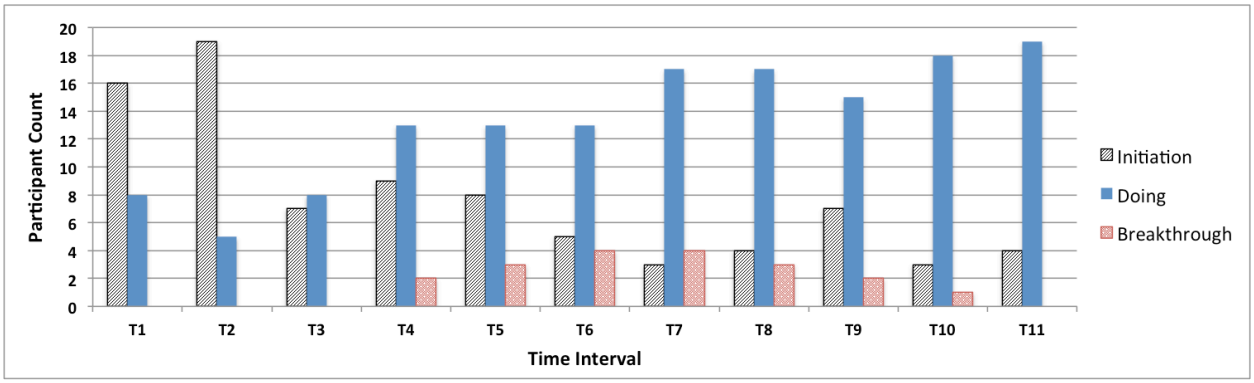

Figure 4. Frequency counts of different levels of behavior during DNA barcode clean up and look up.

The actual Rasch analysis of the observation data used the partial credit estimation in the 'eRm' package (Mair, et al., 2015), which is implemented in the R statistical programming language. Partial credit estimation works on the assumption that a transition from one category of response to another, say, from category 0 to category 1, is evidence that the respondent has moved beyond some threshold on the scale of values that is being constructed through the Rasch Model. In this case, that scale of values is an estimate of the amount of participant engagement. Figure 5 illustrates how this works. Working from the observation data, the analysis computes the probability of observing behaviors corresponding to the three categories at each time interval. In this figure we look just at intervals T4 and T6.

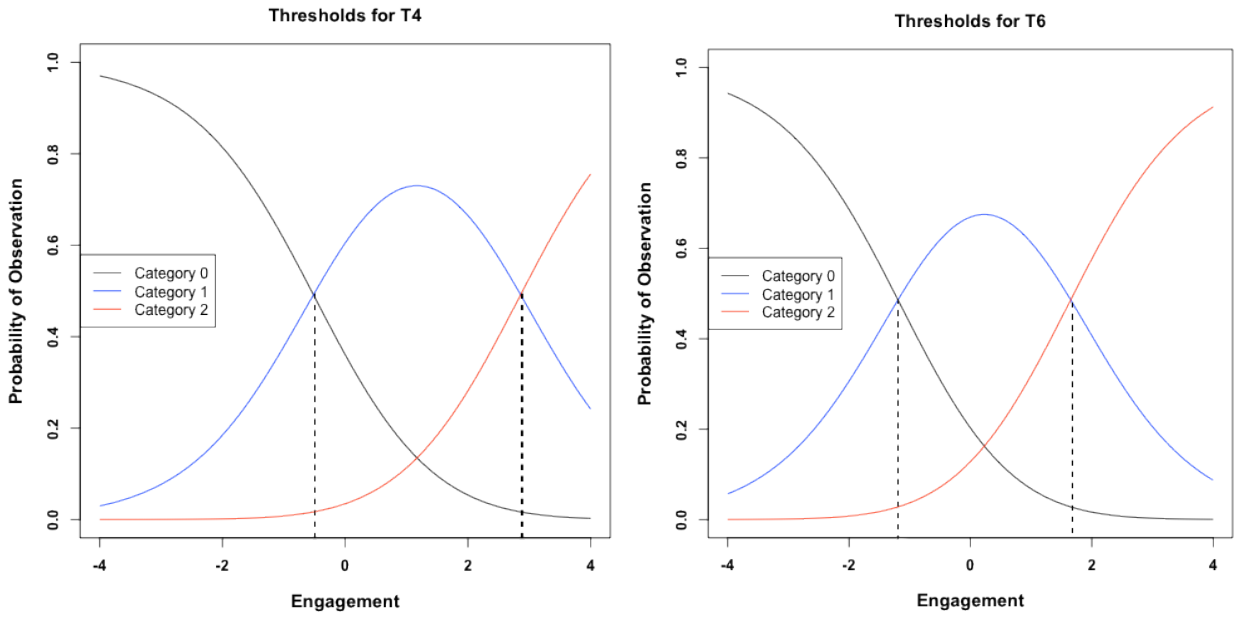

Figure 5. Converting categories of observations to levels of engagement.

The analysis computes the size of the intervals on the horizontal axis and the range of values that are displayed—in this case a scale that we are asserting to represent the degree of engagement—by considering the distribution of engagement across all the participants over the full duration of observations. The model is set up so that participants with an average level of engagement have a 0.5 probability of displaying an engagement level of zero. Note that this means that a negative value on this scale does not mean "non-

engagement." The scale resulting from Rasch analysis is like the Fahrenheit or Celsius temperature scales, where the placement of zero is according to convention or design, rather than an indication of an absolute value.

The dotted lines that drop from the points of intersection between successive categories to the X-axis show how the model converts the probability of observing a particular category of behavior (a probability that is related to the frequency counts shown in Figure 4) into a measurement of engagement. This translation moves from something that can actually be observed (categories of behaviors) to a theoretical construct (in this case, engagement) that can only be inferred from the observations.

Comparison of the engagement values at the two time intervals shows that, according to this model, the level of engagement required to make the transition from Initiation (Category 0) to Doing (Category 1) at interval T4 was about -0.5. Ten minutes later, at interval T6, the dotted line drops down to the X-axis at a value of approximately -1.2. So, moving from Initiation to Doing required 0.7 less engagement at T6 than it did at T4. What is even more interesting is that the level of engagement required to transition from Doing (Category 1) to Breakthrough (Category 2) showed an even greater decrease over the same ten minutes. An engagement level of approximately 2.9 was required to move to Breakthrough at interval T4, but this decreased by 1.2 to a value of approximately 1.7 at interval T6. Because the Rasch model produces an interval scale of values, it is possible to compare these two amounts of change. We can say that it appears that the engagement required to move to Breakthrough decreased about twice as much as the level required to move from Initiation to Doing over this ten minute period. This is the kind of analysis that having an interval scale enables. Clearly, we would want to look at many more observations before making strong assertions. The point here is simply that such comparisons of values are not even possible when researchers can only look at frequency counts, rather than at an actual measurement scale.

Table 2. Engagement transition thresholds at different time intervals during DNA barcode work.

Interval	Threshold 0-1	Threshold 1-2	Interval	Threshold 0-1	Threshold 1-2
T1	0.62	not present	**T7**	-2.07	2.16
T2	1.43	not present	**T8**	-1.96	2.71
T3	0.16	not present	**T9**	-1.31	3.04
T4	-0.52	2.86	**T10**	-2.22	5.89
T5	-0.93	2.22	**T11**	-2.32	not present
T6	-1.19	1.65			

Table 2 summarizes the transition thresholds for all of the observation intervals during the DNA barcode clean up and lookup program. A scan of the table's columns reveals that the transition from Initiation to Doing became relatively less demanding over time. This is what we would expect as participants become more familiar with how to do the required work in a citizen science program. The pattern of thresholds for transitions from Doing to Breakthrough is more interesting. The estimated engagement effort required to transition to Breakthrough activities decreased for a few time periods, but then began to increase. Figure 6 shows this pattern graphically.

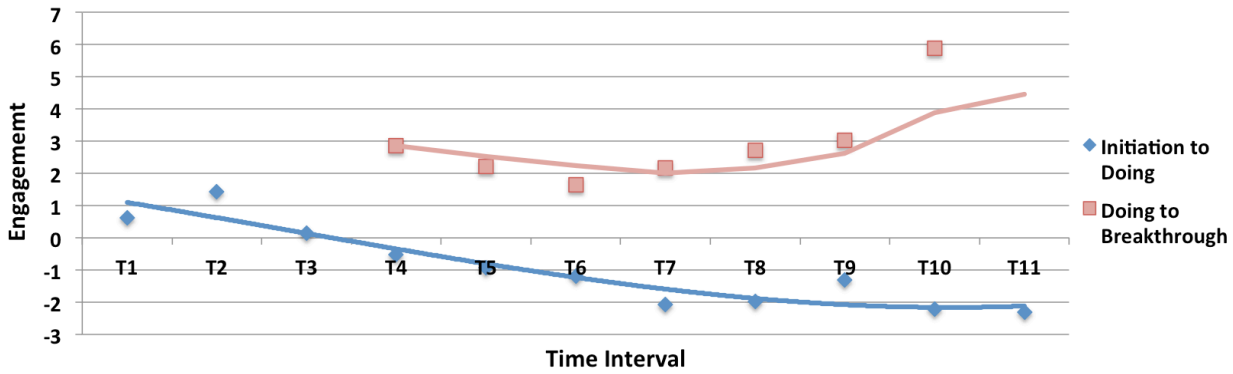

Figure 6. Amount of engagement required to transition between behaviors over time during DNA barcode work.

Applying the Rasch partial credit model to these observation data does not, of course, explain this initial decrease and subsequent increase in engagement associated with Breakthrough behaviors, it only assists in bringing the pattern into focus. We can usefully distinguish between two different kinds of conjectures about possible causes for the pattern. There are conjectures that assume that the pattern is correct, and look to program design, context, and participant motivation for explanations. An example of such a conjecture is that, perhaps, as the novelty of the activity wore off, participants were less engaged. Such conjectures are warranted when we have confidence in the measurement instrument. Making and testing such conjectures is the reason to have good instruments.

When we are using an observation framework that is still at such an early stage of development, there is good reason to question whether the measurement instrument is producing good estimates of engagement. This leads to a second kind of conjecture that considers the possibility that the observation framework needs refinement. For example, a review of the actual hand-coded observation record for the data presented above reveals that a substantial proportion of the behaviors that were classified as Breakthrough behaviors during this event were instances where a more knowledgeable participant was providing assistance to another participant. What difference would it make if we categorized these behaviors as instances of Doing rather than Breakthrough? Figure 7 shows the answer to this question.

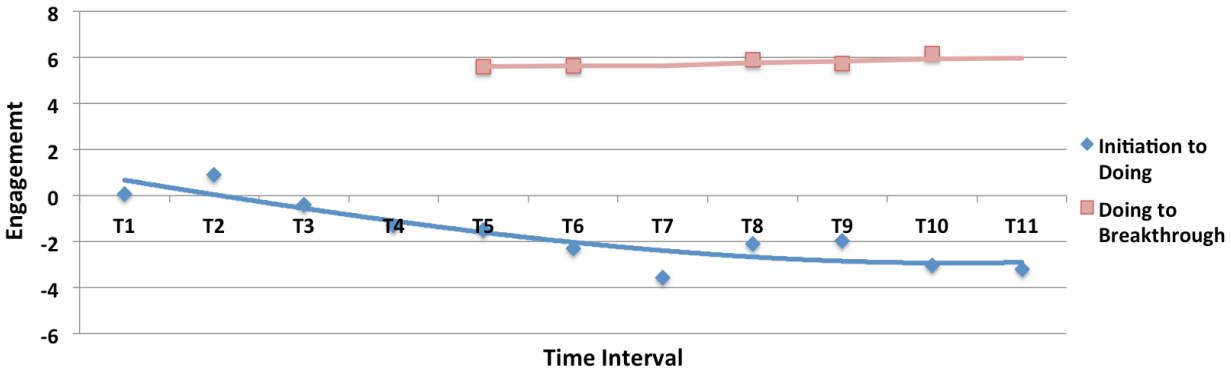

Figure 7. Change in engagement required for Breakthrough after adjusting observation framework: DNA barcode work.

The important point that this example illustrates is that by giving us a way to create a scale for measuring engagement, the Rasch analysis provides us with an easy way to see the effects of making a change in the observation framework. If we think of constructing an observation instrument for measuring engagement in the same way that one would approach constructing other psychometric measures, we would expect instrument construction to be iterative over many trials and to be concerned with validity: what is the chain of reasoning that connects evidence about the instrument's performance with the claim that it is

measuring the theoretical construct of interest? Having the ability to convert counts of observations to an interval scale facilitates such inquiry. In this particular instance, the unexpected shape of Breakthrough engagement plot in Figure 6 raised the question of whether providing assistance to others is more properly just part of the job, instead of evidence of deeper engagement. This kind of question stimulates important thinking about just what we mean by "Breakthrough" and about the nature of engagement. Pursuing that line of thinking may require interviews with participants who were providing assistance to ask them questions about their experiences. The Rasch partial credit model provides quantitative input to a dialogue that is at once quantitative, qualitative, and epistemic and that is an essential part of instrument design and construction.

A Second Test Case

We used the same observation framework to collect data in a different setting with different kinds of participants. Since our focus is on addressing a measurement problem related to participant observation, there is value in applying the measurement tool and techniques in different observation contexts.

We collected observation data during a bioblitz event as participants sorted beetles that had been collected earlier that day. Observations took place in the evening immediately after participants had finished supper and had settled back to work. We observed 17 participants at five-minute intervals for an hour. The group included amateur and professional entomologists who were generally comfortable with identifying and preserving samples as well as people with less experience with insects. Figure 8 summarizes the distribution of observed behaviors.

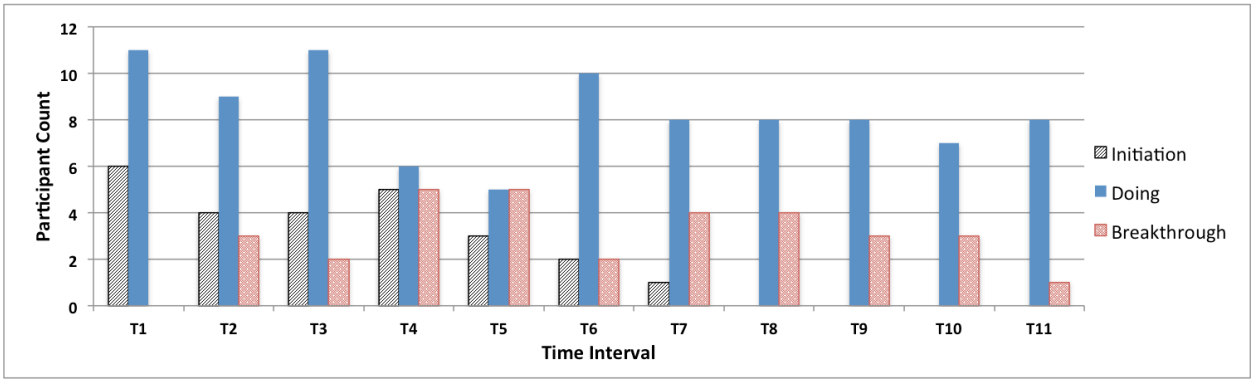

Figure 8. Frequency counts of different categories of behavior during biolblitz specimen sorting.

As one might expect, this group was substantially involved in Doing right from the outset. Participants who were not as familiar with the sorting and pinning process tended to begin by watching more experienced workers until they had a better idea of what was going on. In some cases, some "participants" never moved to the Doing stage, but just watched for a while and left.

Applying the Rasch partial credit model to these data, using the same tools and procedures as before, produced threshold data summarized in Table 3.

Table 3. Engagement transition thresholds at different time intervals during bioblitz sorting.

Interval	Threshold 0-1	Threshold 1-2	Interval	Threshold 0-1	Threshold 1-2
T1	-0.76	not present	T7	-3.54	2.37
T2	-1.82	2.46	T8	not present	1.08
T3	-1.87	3.71	T9	not present	3.72
T4	-0.96	1.22	T10	not present	3.01
T5	-1.59	0.78	T11	not present	4.24
T6	-2.73	3.19			

Scanning the values in this table reveals that Initiation behaviors disappeared after 40 minutes (starting at interval T8) and did not reappear. Figure 9 shows the overall pattern of changes in the level of engagement required to transition between categories of behavior over the hour of observed activity.

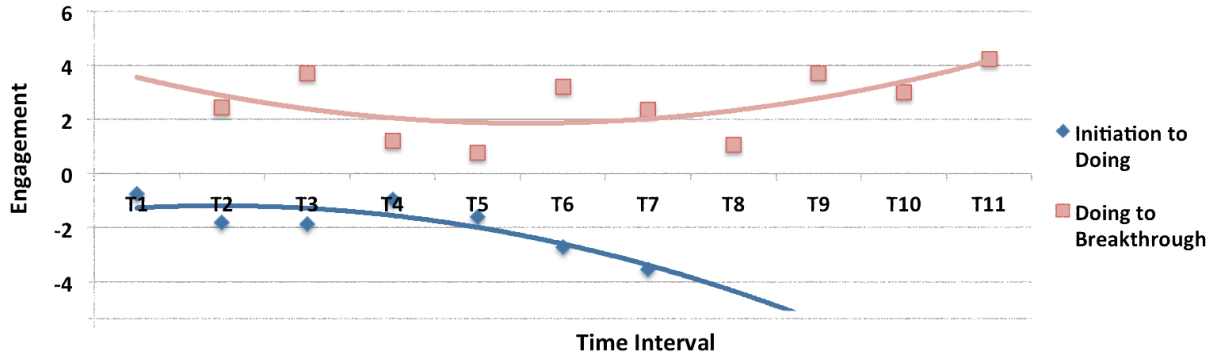

Figure 9. Amount of engagement required to transition between behaviors over time during bioblitz specimen sorting.

As was true for the DNA barcode transition thresholds in Figure 6, transitions from Initiation to Doing became easier over time. In fact, the level of engagement required to begin Doing decreases even more quickly than for the work with DNA barcodes and computers This might be due to the heterogeneous nature of the group: the novices who were not sure about transitioning to Doing left, while the others eventually learned enough by watching or through help by others to settle down to work.

Figure 6 and Figure 9 also show the same pattern of changes in engagement required for transition to Breakthrough behaviors; the amount of engagement required decreased over the first half of the observation period and increased toward the end. As was true for the observations during the DNA barcode work, some of the behaviors categorized as Breakthrough during the bioblitz sorting involved more expert participants providing help to others. We wondered what would happen if we once again re-categorized such behaviors as instances of Doing. Would the line once again flatten out? Figure 10 summarizes the result.

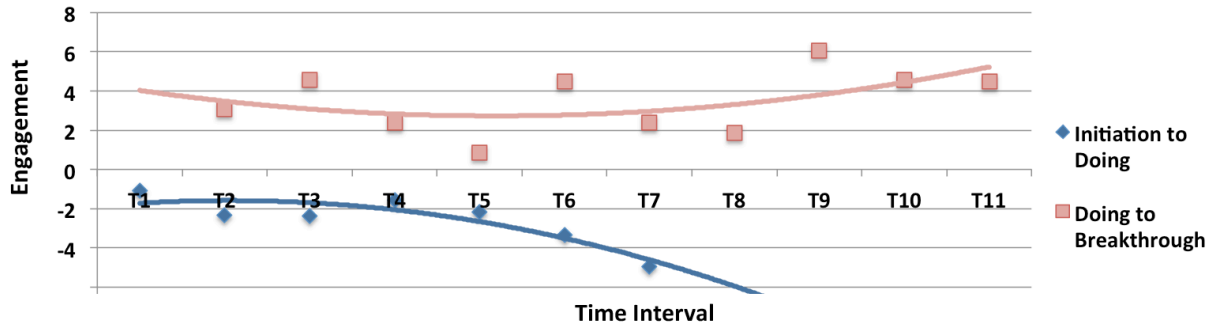

Figure 10. Change in engagement required for Breakthrough after adjusting observation framework: bioblitz sorting.

Adjusting the observation framework had the effect of flattening the curve for the amount of engagement required for Breakthrough to some extent, but the upward trend is still evident. Once again, one can generate conjectures about what is happening here. Perhaps as less experienced participants either settled down to work or left, there was more doing and less sharing. Or, perhaps we should think about other adjustments to the observation framework. The point of this example, in terms of the present paper, is that using the Rasch partial credit model to create a measurement scale for engagement assists in stimulating such thinking.

Discussion, Contribution, Limitations

As noted its opening paragraphs, the present paper is an exploration of a methodological problem: we have been unsatisfied with our efforts to develop and refine participant observation frameworks when the outputs from those frameworks are conceived only as counts of different categories of behavior. We were familiar with use of Rasch models from our work in formal education settings and so wondered whether it might be useful to apply a Rasch model to assist in refining an instrument that uses observation data to make estimates of participant engagement.

Using two small data sets and an observation framework that is still in the early stages of development, this paper demonstrates that Rasch analysis using a partial credit model produces results that are generally consistent with simpler, less quantitatively rich frequency count data. This is an important result. As we set out on this exploration, it was not clear to us that the Rasch model would fit the kinds of observation data that we were collecting. The fact that the Rasch model did fit reasonably well with these small datasets is encouraging. Rasch analysis is increasingly useful and reliable as the number of participants becomes larger, encompassing hundreds of participants over multiple observation sessions. The fact that we are seeing coherent and potentially useful results in these pilot-study datasets suggest that using the Rasch partial credit model will be useful with larger datasets.

Contribution to Citizen Science Research

In this paper we used Rasch partial credit models to assist us in gaining insight into the performance of an observation protocol focused on participant engagement. The approach outlined here is not limited to measures of engagement. Research into the practice of citizen science might rely on observation data in investigating a variety of important theoretical constructs such as persistence, curiosity, capability, and others. In exploring each of these theoretical constructs, having a way to transform counts of observations into an interval scale that represents the amount or degree of the construct has the potential to assist in instrument design and in refining the definition of the construct.

As we noted in the first paragraphs of this paper, researchers or evaluators who are designing observation frameworks to measure engagement, persistence, curiosity, capability, or something else must address two questions: (1) “Which behaviors should be recorded?” and (2) “What values should be assigned to these behaviors, corresponding to more or less of the construct that we are measuring?”

These two questions—one focused on matters of definition and the other focused on scaling—are connected. The exploratory work reported here shows how addressing the scaling question can inform the definitional question. Attention to scaling can assist in framing and gathering data to answer questions such as:

- Is the instrument measuring one thing or several things?
- Are the behaviors that are grouped together really of a piece, or should they be categorized in some other way?
- Are the assumptions that we make about categories of behaviors that we believe are associated with the construct supported by the data? For example, is category 1 consistently less demanding than category 2?

Limitations

As already noted, this paper describes exploration of a methodological problem that arose in the course of pilot work focused on broader questions. It reflects the limitations that are often inherent in exploratory and pilot work. In this case those limitations include use of small samples of participants and the use of tools, such as the engagement observation framework, that are still under development.

Because sample sizes are small, the work presented here should not be used to make claims about the DNA barcoding and bioblitz activities that we used in these analyses. This paper focuses on exploration of a measurement problem, rather than on making generalizations about the programs from which we drew the data used for exploration.

Suggestions For Future Work

We propose two different areas for future exploration and development on the basis of the work presented here. The first focuses on designing observation frameworks. The second extends more broadly to collaborative exploration within the community of researchers and evaluators seeking to understand the potential and the effects of citizen science programs.

Observation Framework Design

The exploration presented here suggests potentially fruitful additional research that might be pursued to improve the sensitivity of the observation framework and to tie it more firmly to a clear conception of engagement. The additional research would involve increasing the number of categories of behaviors and narrowing their focus. The observation framework presented here combines a number of different behaviors into the three broad categories of Initiation, Doing, and Breakthrough. As described above and illustrated in Figure 1 and Figure 2, these categories represent *a priori* conceptions of engagement.

Given the increased confidence emerging from this exploratory work that the Rasch partial credit model can be usefully applied to observation data, it would be useful to experiment with moving from grouping behaviors into categories toward making direct use of different observed behaviors (for example, sharing and picture taking) in the analysis. This would result in estimates of the engagement associated with each behavior, rather than an estimate of engagement for a category that is an amalgam of different behaviors. Doing this would require a great many more observations than we had available during this pilot work, since it would take many more observations to provide a rich picture of nine or ten individual behaviors, as opposed to just three general categories that group behaviors together. It is possible that such a more fine-grained investigation would lead back to some kind of grouping of behaviors, but it would have the advantage of grouping behaviors on the basis of what was learned from observation data, rather than grouping them on an *a priori* basis.

Broader Collaboration

The community of researchers and evaluators who study the design and outcomes of citizen science projects is already engaged in a number of collaborative endeavors that enable different researchers and evaluators to share and build on each other's work. For example, the DEVISE project has assembled and developed a suite of instruments to measure constructs such as interest, scientific self-efficacy, and skills (Phillips, et al., 2014), with the aim of stimulating use of measures that might be compared across studies.

We see the potential for a similar, or perhaps a related effort that could grow out of the exploratory work presented here. Such collaborative work could unfold in two stages.

In the first stage different researchers could collaborate in pooling observation data to create better observation frameworks capable of providing interval scale measures of constructs such as engagement, persistence, and so on. As noted above, the power and utility of the measurement tools explored in this paper increases substantially as they are applied to larger numbers of participants. Specifically, with access to many more observations of citizen scientists, it would be possible to explore observation frameworks that make the kinds of more nuanced distinctions described in the preceding section. Developing this kind of more powerful, reliable, and valid measure of different constructs would proceed much more quickly if the citizen science research community pooled its observation data.

During this first, design and construction stage the goal would be to produce a small number of frameworks aimed at use with different participant populations (for example, expert citizen scientists and more casual citizen scientists) or in different settings (for example, sample collection in the field as opposed to laboratory work). Working iteratively over time and using data from many studies and using analytical insights provided through Rasch modeling, these observation frameworks might develop into a suite of measurement instruments with well-understood properties and applications.

The second stage would focus on using, as opposed to refining, the observation frameworks emerging from the first stage of work. The aim of this effort would be similar to a second aim of the current DEVISE project: create a set of tools that facilitates comparison and learning across different programs, contexts, and participant populations. Having access to observation frameworks that resulted in interval scale measures, as opposed to frequency counts, would assist in making statistical inferences about gains, changes, and differences across studies.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. DRL-1223210. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Barriault, C. (2014). Assessing Visitor Learning in Zoos and Aquaria: A Revised Framework. Presented at the NARST Annual International Conference, Pittsburgh, PA, March 30-April 2, 2014.

Barriault, C., & Pearson, D. (2010). Assessing Exhibits for Learning in Science Centers: A Practical Tool. *Visitor Studies*, *13*(1), 90–106. doi:10.1080/10645571003618824

Bond, T. G., & Fox, C. M. (2012). *Applying the Rasch Model* (2nd ed.). New York: Routledge.

Horizon Research. (2000). Inside the Classroom Observation and Analytic Protocol. Chapel Hill, NC: Horizon Research Inc. Retrieved from http://www.horizon-research.com/inside-the-classroom-observation-and-analytic-protocol/

Mair, P., Hatzinger, R., Maier, M. J., & Rusch, T. (2015). Package "eRm." CRAN. Retrieved from http://cran.r-project.org/web/packages/eRm/eRm.pdf

Palmquist, S. D., & Crowley, K. (2007). Studying dinosaur learning on an island of expertise. In R. Goldman, R. Pea, B. Barron, & S. J. Derry, *Video Research in the Learning Sciences* (pp. 271–286). Mahway, NJ: Routledge.

Phillips, T. B., Ferguson, M., Minarcheck, M., Porticella, N., & Bonney, R. (2014). *User's Guide for Evaluating Learning Outcomes in Citizen Science*. Ithaca, NY: Cornell Lab of Ornithology.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.