CITIZEN SCIENCE:
THEORY AND PRACTICE

Using Shared Embedded Assessment Tools to Understand Participant Skills: Processes and Lessons Learned

METHOD

RACHEL BECKER-KLEIN
CATHLYN DAVIS
TINA B. PHILLIPS
VERONICA DEL BIANCO
AMY GRACK NELSON
EVELYN CHRISTIAN RONNING

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

This paper describes the collaborative process for how a group of citizen science project leaders, evaluators, and researchers worked together to develop, validate, and test embedded assessments of two different volunteer science inquiry skills. The development process for creating these embedded assessments (activities integrated into the learning experience, allowing learners to demonstrate competencies) is articulated, as well as challenges encountered in assessing two science inquiry skills common in citizen science projects: *notice relevant features* and *record standard observations*. The authors investigate the extent to which the assessments were successful at achieving four criteria identified as ideal for shared embedded assessments of volunteers' skills, namely: broadly applicable, authentic, performance-based, and integrated.

CORRESPONDING AUTHOR:
Rachel Becker-Klein
Two Roads Consulting, US
rachel@consulttworoads.com

KEYWORDS:
citizen science; volunteers; embedded assessment; science inquiry skills

TO CITE THIS ARTICLE:
Becker-Klein, R, Davis, C, Phillips, TB, Del Bianco, V, Grack Nelson, A and Ronning, EC. 2023. Using Shared Embedded Assessment Tools to Understand Participant Skills: Processes and Lessons Learned. *Citizen Science: Theory and Practice*, 8(1): 20, pp. 1–12. DOI: https://doi.org/10.5334/cstp.487

INTRODUCTION

A persistent challenge for measuring science outcomes in informal science learning (ISL) settings and especially in citizen science (CS) is the development of a method that is both rigorous and appropriate for the context (Allen and Peterman 2019; Fu, Kannan, and Shavelsen 2019). Self-report surveys and interviews remain the most common methods for measuring science outcomes in ISL evaluations, providing researchers a valuable glimpse into the inner thoughts of participants (Fu et al. 2016), and may be most appropriate for measuring latent variables such as interest and attitudes. Since skill-based outcomes are essential to project activities to ensure skill proficiency for following protocols and enhancing data quality, (Burgess et al. 2017; Stylinksi et al. 2020), developing complementary (i.e., non-traditional) approaches for measuring science inquiry skills may be of particular importance for CS projects

Embedded assessments (EAs) comprise innovative "opportunities to assess participant progress and performance that are integrated into instructional materials and virtually indistinguishable from day-to-day [program] activities" (Wilson and Sloane 2000, p. 84). As such, EAs allow learners to demonstrate their science competencies through tasks that are integrated seamlessly into the learning experience itself, and thus offer potential to determine participants' skills in authentic and unobtrusive ways (Becker-Klein et al. 2016). Although EAs offer many opportunities as an assessment tool, they also pose some significant challenges that may have prevented widespread adoption to date. One such obstacle is the difficulty of creating a comprehensive EA for science inquiry skills that would be broadly applicable and also meet other criteria of assessment that we find ideal for CS: authentic, performance-based, and integrated into project activities. To date, there has been a dearth of EA tools published, especially EAs that are relevant across projects (rather than customized to only one particular project) (Hussar et al. 2008; Kim et al. 2021).

Our team for this project (funded by the National Science Foundation, DRL# 1713424) had the opportunity to investigate and publish a series of papers on the volunteer skill assessment processes and impacts within several different citizen science (CS) projects. All of the papers in this series focused on the embedded assessment of CS volunteer skills (Davis et al. 2022; Peterman et al. 2022; Stylinski et al. 2020). These publications have made the case for an increase in CS projects that assess their volunteers' targeted science skills. We have also called for innovative approaches to measuring volunteer skills that can complement existing self-report surveys of skills (Becker-Klein et al. 2016). In this paper, we discuss the collaborative process used to develop and validate two new EAs to assess volunteer science inquiry skills that are broadly applicable, authentic, performance-based, and integrated to a CS project experience.

CRITERIA FOR COMPREHENSIVE EMBEDDED ASSESSMENTS

BROADLY APPLICABLE SHARED MEASURES

Shared measures are defined as "rigorous measures that can be shared, or applied, across programs that are addressing the same construct or outcome" (Grack Nelson et al. 2019, p. 60). The literature on the importance of shared measures provides a strong rationale for the need for assessments to be broadly applicable, especially in informal learning environments (Grack Nelson et al. 2019; Hussar et al. 2008). Shared measures can take many forms, such as surveys, tests or quizzes, observation protocols, etc. Recently, research and evaluation teams have pioneered the development of survey-type instruments specifically for use in assessing volunteer outcomes within cross-project evaluations of citizen science (e.g., the Developing, Validating and Implementing Situated Evaluation Instruments [DEVISE] scales; Phillips et al. 2014); science interest and science classroom practices (Noam et al. 2017); "activation" of science learning that can bridge formal and informal contexts (Learning Activation Lab 2018); and outcomes for scientists who participate in public engagement activities (Peterman et al. 2017). Grounded in theory and developed using a process to ensure that the tools measure what they are supposed to measure across contexts, these shared measures have tremendous potential to propel evaluation and research about informal learning outcomes among participants in citizen science projects.

AUTHENTIC

For an EA to be authentic, it must directly examine participant performance on related tasks, rather than relying on indirect or proxy activities (Wiggins 1990). The learning environment must be considered and mirrored as closely as possible (Ashford-Rowe et al. 2014). For instance, if a project is about observation of bees, then the assessment should be about observing bees and not other species or other types of data collection. Other project considerations should also be taken into consideration, such as the type of observation conducted. For example, a project that asks volunteers to examine photographs should have an assessment involving sample photographs.

PERFORMANCE-BASED

A performance-based EA for volunteer science inquiry skills asks volunteers to use or apply a skill rather than reporting on their own abilities or on their self-efficacy with said skills (Fu et al. 2019). Both formal and informal science education have seen recent calls encouraging researchers and evaluators to begin using performance as a key metric of skill (National Research Council 2009; National Research Council 2010; National Research Council 2015). Participants need to demonstrate that they can perform a skill by engaging in a task or set of tasks that requires them to apply the identified skill (Ashford-Rowe et al. 2014).

INTEGRATED

Finally, integrated tasks are incorporated into a project's specific activities. This could be incorporated into a project's training or could be part of regular data collection or submission of data. Integrating assessments into the curriculum or instruction process is a key component of embedding assessments (Wilson and Sloane 2000), and it is widely recognized to be challenging to accomplish (Sloane et al. 1996). Yet, there has been a call in the field for "more direct and less obtrusive measures," (Fu, Kannan, and Shavelson 2019); this is one of the primary advantages of EAs—they can be integrated into the programmatic experience without adding on to the assessment burden for volunteers.

Here, we describe the processes used, and opportunities and challenges encountered, in creating shared embedded assessments of CS volunteer skills (within ten different CS projects) that aim to meet the criteria of broadly applicable, authentic, performance-based, and integrated. This work was part of a larger study funded by the National Science Foundation (DRL #1713424).

To further understand the success of this development process for creating new methods to assess volunteer observation skills, we developed the following primary question to guide the inquiry:

To what extent do these embedded assessment tools meet our criteria for shared embedded assessments of skills: (1) broadly applicable, (2) performance-based, (3) authentic, and (4) integrated?

METHODS

SHARED EMBEDDED ASSESSMENT DEVELOPMENT PROCESS APPLIED TO 10 CITIZEN SCIENCE PROJECTS

Staff from ten CS projects (see Table 1) were partners in developing, implementing, and revising the shared EAs developed over the course of three years. Two of these project leaders were identified through our initial work on a prior grant (Stylinski et al. 2020) in which we created a set of customized EAs. We continued to recruit project leaders through additional inquiries with known relevant CS investigations, as well as through a snowball sampling method. We intentionally limited our search to focus on a single discipline within CS (environmental science) for which nature observations are an important skill. We believe that this focus on environmental science CS projects represents a substantial portion of the CS field—a portion broad enough to capture important diversity, yet narrow enough to allow for the development and testing of new EA tools customized to a subset of science observation skills. We interviewed several project leaders in the fall of 2018, asking questions about what a typical volunteer does in their project, what science inquiry skills volunteers learn and practice, and whether the project provides training for

PROJECT NAME	RESEARCH TOPIC	IDENTIFICATION FOCUS
Natural North Carolina	Tracking biodiversity and species distribution in North Carolina	Various plants and animals
Nature's Notebook	Seasonal changes in plants and animals	Plant and animal phenophases
FrogWatch USA	Amphibian presence and behavior	Frogs and toads
eMammal	Wildlife presence through camera trapping	Various mammals
Chesapeake Bay Parasite Project	Presence of parasites on marine invertebrates	Mud crabs
Chestnut MegaTransect	Document American chestnuts along the Appalachian Trail	American chestnut trees
Biosphere2 Agrivoltaics	Co-production of food and solar electricity	Plant phenology and fruiting
BeeSpotter	Collect baseline data on honey bees and bumblebees	Honey bees, bumblebees
Michigan Butterfly Network	Assess population of butterfly species in MI	Butterfly species
Project FeederWatch	Count of birds that visit backyards and/or supplementary feeding stations	Feeder birds

Table 1 Citizen science projects involved in developing shared embedded assessments (EAs).

those skills. On the basis of the interview results, project leader interest, and selection for heterogeneity in several areas (i.e., different species/phenomena of interest, online and in person, and different research topics), we ended up with 10 projects. The ten CS projects chosen intentionally included a diverse array of research topics and identification foci, as illustrated in Table 1.

These 10 CS projects worked with our team over the course of three years to participate in and provide feedback to create, test, and implement two shared EAs. The current study is based off of our previous work on a National Science Foundation grant (DRL #1422099) to test and articulate an EA development process that was customized to specific projects (described in detail in Peterman et al. 2017). In contrast, this current study focuses on a Shared EA Development Process in which the leadership team (consisting of the authors on this paper) collaborated with project leaders to guide them through the process of creating EAs to be used across multiple projects (Figure 1). The work was carried out in three distinct phases.

Stage 1 focused on working with the projects to determine what skill to measure, and consisted of (1) collaboratively identifying common science inquiry skills, and (2) articulating each project's goals and activities that align with those skills. An in-person meeting initiated the collaborative process with project leaders, followed by a series of group conversations identifying both the skills to focus on and existing activities that featured those skills and potentially demonstrated skill proficiency or development.

The leadership team took the information from this meeting and follow-up conversations to draft two potential EAs: one for the skill of *notice relevant features* and one for the skill of *record standard observations*, both of which are a part of the larger task of scientific observation and thus, science inquiry. Our work builds on that of Eberbach and Crowley (2009), who defined characteristics that distinguish everyday observations from scientific observations. In the case of the EA for *notice relevant features*, an observer must be able to match what they see (e.g., floral structure) with their content knowledge (e.g., plant families). For this EA, volunteers are asked to identify organisms in different photos (relevant to a specific project), and are prompted to "show their work" with the question, "what features of the animal did you use in your identification?" The EA for *record standard observations* is related to what was described in Eberbach and Crowley's (2009, p. 56) framework as "Record observations using established disciplinary procedures and representations." In general, we refer to "standard observations" as those that provide a consistent or uniform set of measurements to describe a phenomenon or event. For this EA, project leaders created a three- to five-minute video to simulate the perspective of the data collector, and then volunteers were asked to watch the video and fill in a modified data sheet as if recording the data observed in the video.

Each project leader chose the EA for the skill most relevant to their project's activities, resulting in five projects choosing each EA. Project leaders iteratively provided input

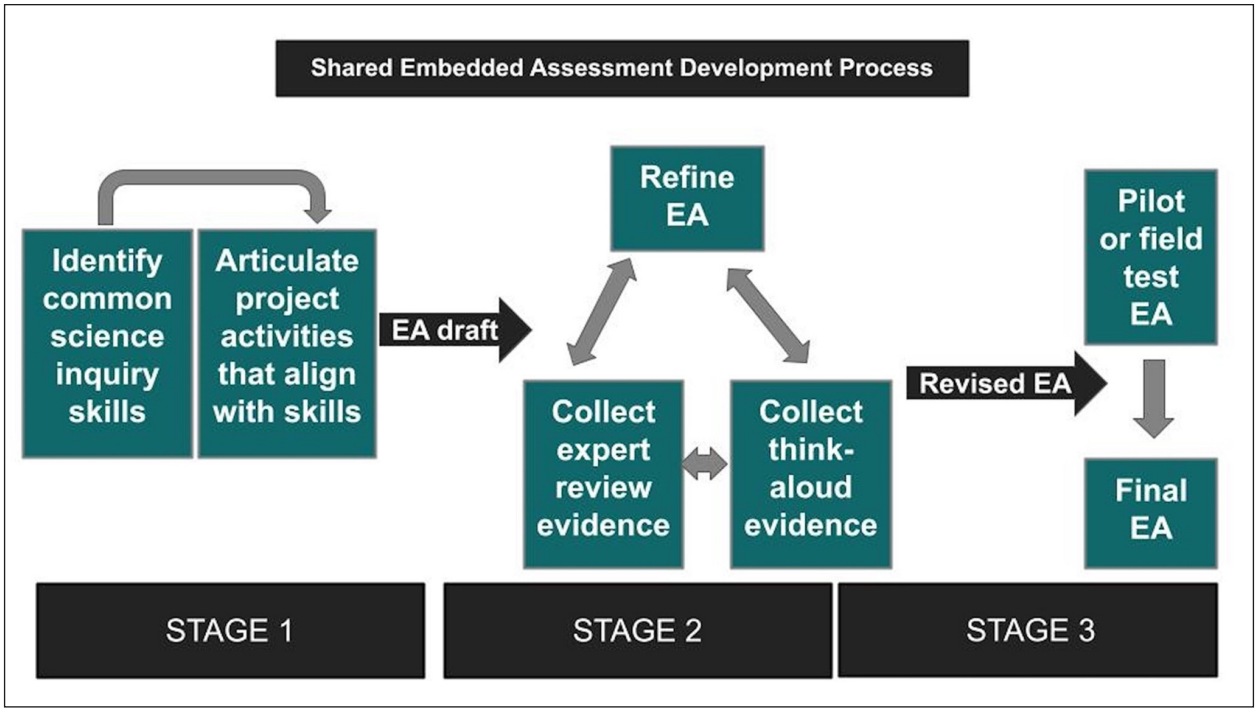

Figure 1 Shared embedded assessment development process.

and feedback on the draft EA, getting it ready for Stage 2 of the process.

Stage 2 focused on gathering evidence and making revisions to ensure the draft EAs were collecting trustworthy data about the skills they hoped to be measuring. There are various methods to check that measures are gathering reliable and valid data, but this project focused on two methods that would be most useful for the EAs developed: an expert review process and think-aloud interviews (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014).

Expert reviewers examined whether the tasks on each EA were sufficiently addressing the skill area of interest. The two draft EA measures were reviewed by four advisors with expertise in measurement, evaluation, scientific inquiry, and CS, using an adapted version of Newman, Lim, and Pineda's (2013) expert review process. For each EA task, experts were asked to provide feedback on how well that task aligned with the project's definition of that skill area, and if needed, provide suggestions for how the task might be revised to better align with the skill. The expert review results were then used to revise the EA tasks so they more thoroughly and accurately addressed the skill areas.

Think-aloud interviews were then conducted with volunteers from two different citizen science projects (one project for each draft EA). To recruit participants, project leaders emailed volunteers inviting them to participate in a think-aloud interview. For the *notice relevant features* EA, sixteen volunteers signed up, and from these, 10 were selected to ensure variability across gender identity, age, and self-reported skill level of identifying animals. Eight volunteers expressed interest in testing the *record standard observations* EA, all of whom were included in the think-aloud sample. The purpose of the think-aloud interviews was to help uncover any confusing or misinterpreted tasks. During the think-aloud interviews, CS volunteers described their thinking processes out loud as they responded to each task in the assessment (Beatty and Willis 2007). This process helped to uncover if EA tasks were interpreted, and thus completed, as intended; if not, further revisions were made in preparation for pilot testing.

In Stage 3 of the process, the projects were asked to pilot the revised EA tools with a few participants first, and then later implement a broader field test within their projects. The leadership team worked with CS project leaders to customize and finalize the EA for their project and to determine how to integrate it into project activities (for both skills, EAs ended up integrated into the project training for most projects). At the end of this stage, the leadership team facilitated another meeting to examine findings for each project and to reflect on the process.

In the first round of applying the shared EA development process, five different CS projects worked together in small groups to each create an EA for a particular skill. These two small groups, consisting of five projects each with different content areas and formats, were each able to develop an EA for *notice relevant features* and for *record standard observations*. All 10 CS projects piloted some version of at least one of the EAs.

Once the first round of EAs were developed, the leadership team facilitated a second round of EAs, and three projects that had not done so before chose to try out the EA developed for *notice relevant features*. In the scope of this study, the EA for *record standard observations* was not tried in the second round by any of the CS projects.

EXAMPLE OF NOTICE RELEVANT FEATURES EMBEDDED ASSESSMENT

One of the partner projects that co-developed the EA for *notice relevant features* asks volunteers to identify animals seen in photos. The identified features include picking out the animal from its background as well as recognizing particular features of a species, such as the color of the ears (i.e., for a red fox) or a stripe on the back (i.e., for a grey fox). The EA includes different photos taken from a camera trap, and participants are asked to determine if there are indeed animals in the photo, and then identify the animal(s) as specifically as possible. For instance, one photo is a red fox (see Figure 2), which volunteers would be able to identify if they knew the distinguishing characteristics of black-tipped ears, white-tipped tail, and black "boots." Via an online questionnaire, volunteers answered the questions, "Please identify this animal as specifically as possible," and "what features of the animal did you use for your identification?"

EXAMPLE OF RECORD STANDARD OBSERVATIONS EMBEDDED ASSESSMENT

One of the projects that co-developed the EA for *record standard observations* produced a video simulating the data collection process for identification of butterflies. The video, which is just under six minutes, offers a first-person perspective, and starts with a sweeping view of a verdant meadow full of flowers, cut through by a path. The person taking the video walks through the meadow, panning from left to right and back again as they "look around." When the species of interest is spotted, the camera focuses on it. The intent of the video is that the project volunteer watching should feel as if they themselves are walking through the meadow; the volunteer records any species of interest that they see on a datasheet provided, which is then scored by project leaders to determine whether the volunteer correctly identified the family or species shown in the video. In one part of the video, a distractor (an example

Figure 2 Photo of a red fox taken from a camera trap project.

of the protocol not being followed correctly) is introduced as the camera pans behind the person walking (volunteers are only supposed to record data they see in front and to the sides of them).

DATA COLLECTION

To understand the efficacy of the EAs during testing and implementation, we interviewed the project leaders at two different points. The first interview was conducted with each project in December of 2020 (immediately after Stage 2 of the Shared EA Development process) to gather perspectives about the implementation of their EAs. Specifically, we asked project leaders to consider which parts of the EA development process they considered to be most valuable, and which were most challenging, and whether they had any specific difficulties or "aha moments." We also asked the leaders to consider our ideal EA and which components (i.e., broadly applicable, authentic, performance-based, and integrated) were met adequately by the two EAs developed. A second interview was conducted in spring 2021 (at the conclusion of Stage 3 of the Shared EA Development process) to see how project leaders used the findings from the EAs, but data from this interview were not relevant for this study. All interviews lasted between 30 and 60 minutes and were recorded and later transcribed verbatim.

Notes were kept from meetings that occurred in Stages 1, 2, and 3 for any relevant project feedback. For instance, in the final meeting (after Stage 3), project leaders were asked to rate how well they thought the EAs had met each of the criteria (i.e., authentic, performance-based, and integrated), and discussion ensued.

This research was approved by the University of Maryland Institutional Review Board (IRB #1072528). Informed consent was provided by all volunteers who participated in interviews and meetings.

DATA ANALYSIS

Coding schemes to analyze the project leader interview questions were developed by a team of three researchers (see Supplemental File 1: Sample Interview Questions and Codebook), using six steps of collaborative qualitative analysis (Richards and Hemphill 2018). First, individual researchers reviewed different interview transcripts to conduct open coding and to create memos to guide the development of codes. The team met regularly throughout this process to discuss individual findings, and to work together to develop a preliminary codebook. The two codes relevant for the project leader interviews were about the process of developing EAs (i.e., mention of the value of meeting with other project leaders) and the EA components (i.e., discussion of the four components of an ideal EA). The team then tried to apply the preliminary codes to separate transcripts and met again to review and refine the codes. Two coders then pilot tested the codebook, independently coding the same transcripts, and final adjustments were made to the codebook on the basis of those experiences. Once the codebook was finalized, consensus coding was

conducted on all data records. Two researchers coded each interview independently, and then compared codes; all disagreements were discussed, and the final code was agreed upon. Interview data were coded and analyzed using NVivo 12.

RESULTS

The findings below address the EA development process question investigating the extent to which participants found the EAs developed to be: (1) broadly applicable; (2) authentic; (3) performance-based; and (4) integrated.

TO WHAT EXTENT ARE THESE TOOLS CONSIDERED BROADLY APPLICABLE?

One way to determine whether an EA is broadly applicable is to consider the application of the tool to a project for which it was not originally developed. In this case, each of the three projects that used the EA *notice relevant features* in the second round of EA development found it relatively straightforward to take the instructions developed initially and create their own version of the EA customized for their particular project. One project was able to copy the survey developed in the first round from a different CS project, and change the photos and species question slightly to use the EA for their own project. So, instead of asking to identify "which animal" a volunteer saw, they were asked "which butterfly" they saw, but the rest of the survey remained the same.

When asked how they felt about creating EAs that could apply across projects, most project leaders (80%) thought that the two EAs developed, and especially the format or process of these assessments, were indeed applicable to projects beyond their own, as exemplified by the following two quotes.

> *I think the video was broadly applicable, the technique…obviously [the particular species] are very specific, but I think the video technique can be really workable for many projects.*

> *I think the notice relevant features assessment is broadly applicable. I feel like [projects could] easily change out photos, change out the answers for the features and then it would apply to a bunch of different programs.*

Some project leaders noted challenges in creating an assessment in that in order to apply across projects, there had to be customization of EAs to individual projects. One said,

> *…it seemed like things could be developed that worked in really interesting and helpful ways for individual projects, I'm less convinced that this is something that can be done across projects… there could be some styles or approaches that are generalizable, but I think they're going to have to be pretty heavily customized for each project.*

TO WHAT EXTENT ARE THESE TOOLS CONSIDERED AUTHENTIC?

Challenges in the authenticity of the EAs were raised in the validation process, and project leaders agreed that it was quite a challenge to use videos as a format to mimic real-life data collection, questioning the authenticity of the EA. Ensuring that videos showed enough detail and paused on the identified species long enough for someone to be able to adequately see the details of the species or phenomenon was an obstacle. These considerations led project leaders to retake their videos to pay attention to how a viewer might see the images.

Some project leaders noted other ways that the video did not mimic data collection in the real world, such as noting that the view seen through the lens of a camera is not the same view as the eye can see, with its peripheral vision capabilities. One said,

> *I loved the video, but it was not like looking directly at the [species] because it is a much more narrow view.*

Project leaders also noted that when participants collect data, they can turn their heads and bodies in response to cues, but when watching a video, viewers are constrained by the scope and direction of the videographer's lens.There were even a few complaints of jarring or dizzying scenery due to video camera instability in one of the videos that involves walking while collecting data.

Even with the challenges identified above, almost all (80%) of project leaders stated that they thought the EAs developed were indeed authentic to their projects. Some project leaders mentioned,

> *I think they are authentic since we used actual photos that people had submitted…I don't know if you could get more authentic than asking them **while** they are submitting [their data].*

> *I know these assessments are very authentic to my project, because they really fit in seamlessly with what I wanted people to do, so I would consider that very successful.*

> *It certainly felt authentic. [The assessment] was dealing with a real problem and using real people and real data, and so the authenticity was there.*

The projects that felt the assessments were less authentic to their project talked about how either the level of analysis was not quite right or the assessment may have not been quite authentic to project activities. For instance, two people said,

> *At points, it felt like we were getting down to a level of detail that almost seemed unrelated [to the project's training]…I think [the assessment] gets to participant learning rather than just from the content side [of what our trainers do with participants].*

> *I thought we did a pretty good job on all of the [EA components], but there was some question as to the authenticity of the mode.*

TO WHAT EXTENT WERE THESE TOOLS CONSIDERED PERFORMANCE BASED?

Project leaders were in consensus (100%) that the embedded assessments were performance based. This finding was intriguing to the research team leaders, because there was some question about whether the *notice relevant features* EA, which was an online questionnaire, could be classified as performance based. However, project leaders felt that since participants were being asked to do a task, rather than self-report on their ability to do that task, it did indeed qualify as a performance-based assessment. For instance, some people said,

> *[Participants] were being asked to identify features and demonstrate their skills as far as being able to figure out what are relevant features, so I think [the assessments] seem clearly performance based to me.*

> *I think they're performance based, yeah, [participants are being asked to] follow instructions and identify stuff.*

In addition, some project leaders appreciated that these assessments could be used in their training as a way both to gauge skill level and to give participants a chance to practice the skill. One person asserted,

> *…I think what's nice is that this is something that I feel like we could refresh and continue to utilize both as a training reinforcement as well as an opportunity for us to assess [their skills] on the back end.*

TO WHAT EXTENT WERE TOOLS CONSIDERED INTEGRATED?

The component that was least endorsed as being successfully achieved by project leaders (50%) was integration. Several reasons for lower ratings on the integrated component were cited, such as: the assessments were not placed comprehensively into project activities and platforms (they were primarily used as a training exercise), the assessments were not used for all project participants, and the assessments generally missed the mark. Some of the interviewees asserted,

> *Sending people a video doesn't seem like bad advice, it's just like testing them, which is fine, but I always struggled with this notion of it being integrated, that was always the tricky part for me.*

> *I hesitate to say we missed the mark, I just don't know that there's another good way to do that…but I think that unless we're going to do virtual reality or something like that, a video [is not that integrated].*

> *Part of [our project] is this fairly complex, integrated platform of software, and I didn't have a way to integrate the assessments into that platform.*

Yet, other project leaders felt that including the assessment into the training was a way of integrating it seamlessly into project activities, as exemplified in the following quote:

> *You didn't feel like you were asking them to do something extra, but you could make it make sense as part of the training…there was a natural place, it didn't feel forced.*

DISCUSSION

In this paper, we investigate the process of co-developing, validating, and implementing shared EAs with CS project leaders to reveal opportunities and challenges for other citizen science projects, and potentially for the broader field of informal science education. The collaborative process did indeed lead to two shared EAs, one each for the skills of *notice relevant features* and *record standard observations.* In our study, most project leaders agreed that the shared EAs developed were applicable across projects, and were performance based and authentic to their projects, providing evidence that though time and resource intensive, shared EAs can be developed to more directly measure skill in different settings.

There is a legitimate concern within informal science education that standardizing measures could "undermine the ecological validity" of those instruments by not adequately addressing the nuanced nature of program settings (Allen and Peterman 2019, p. 22). However, there are notable examples in the field of informal science education that have successfully developed and used shared and standardized measures such as the Common Instrument Suite (Noam et al 2017) and the standardization of tracking and timing methods in museums (Serrell 1998). In our study, the EAs developed provided additional evidence that standardized measures can be created that are ecologically valid in that they can be both shared (i.e., broadly applicable) and considered authentic to specific CS projects.

When measuring skills, particularly in informal science learning settings, direct assessments have been found to be valuable, especially assessment tasks that mirror the situations that occur in the project setting (Fu et al. 2016; Shavelson et al. 2018). Although the "directness" of an assessment could be thought of as a continuum, Fu, Kannan, and Shavelson (2019) offer an example of less and more direct measures of visitors' understanding about invasive species, in which "a *less direct* measure... might ask individuals to rate their level of knowledge, a *more direct* measure might ask individuals to answer test questions about invasive species; and an even more direct measure might ask individuals to walk through a garden and play a game in which they identify various native and non-native plant species" (p. 38). The EAs developed in this project seemed to make progress in moving toward more direct measures of skills. This is particularly relevant for citizen science, in which "the 'real-life' feature of many citizen science projects facilitates a space in which participants are able to immerse themselves directly in the use of project-specific tools" (National Academies of Sciences, Engineering, and Medicine 2018). The importance of the assessment mirroring the task for such projects is paramount.

We recognize the potential benefits of more direct complementary assessments (such as EAs) that may not be subject to the same validity concerns as self-reports (e.g., social desirability bias). However, EAs are subject to other challenges and validity concerns themselves, some of which we encountered in this study. Instrument validity is determined through a complex series of gathering evidence related to how an instrument is interpreted and what conclusions can be drawn from data gathered by that measure (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education 2014), including both content validity (i.e., how well a measure represents a particular construct) and response process validity (i.e., what processes participants use when completing a task). The two EAs developed in this study went through testing from expert reviewers to check for content validity and through think-aloud interviews with potential participants to check for response process validity. These two validity tests uncovered some challenges which were addressed and others that we were not able to address, which in turn impacted the extent to which the tools were considered to be authentic to the projects. For instance, video or simulation-type EAs should be tested to ensure the EA is focused solely on someone's ability to do the tasks the EA is assessing, rather than unknowingly measuring other variables such as an individual's ability to use the features of an online video player.

CONCLUSION

After spending three years with this group of CS projects and leaders, we believe that there is value to the process of co-developing rigorous shared measures such as EAs, and acknowledge the several challenges we faced. The challenge of creating shared EA measures that are broadly applicable across projects was amplified in our study by the selection of a widely diverse pool of CS projects. The research team intentionally chose a range of CS projects (all under the umbrella of environmental observation) to investigate the extent to which co-developed EAs could be applicable to a wide array of projects with different species and foci. The leadership team believes that testing the shared EA development process with such a heterogeneous group of projects allowed investigation of opportunities and challenges of the process in a way that can inform the field more broadly. Although each EA tool was customized to each particular project, the fact that the EA tools could be used across different projects with a shared process indicates that the tools were generalizable enough to be used for multiple projects, while customizable enough to be relevant to particular CS projects. However, future attempts may benefit from selecting a less heterogeneous group of projects (e.g., water quality monitoring projects) to co-develop an EA that might be applicable across projects with more consistent activities and/or goals.

The integrated component was considered to be the most challenging part to get right, particularly since a measure could be worked into a training, but if the assessment required new vocabulary or activities, it was more of a struggle to seamlessly put it into the project's activities. One important note is that to integrate an EA

beyond the training often required changes to the project's data entry platform in a meaningful way, and most project leaders (all except one) did not have control of the data infrastructure of their project. Future studies may select projects that have more direct control of the data platform, so that findings from the assessments can inform relevant changes to the way data are collected, in addition to the training itself.

There are numerous benefits to developing and using shared measures within informal science education. Grack Nelson et al. (2019) identified several of these benefits, such as increasing the quality of evaluation conducted, enabling cross-project analyses, and saving time and resources. These benefits seem to make the effort of the shared EA development process worth pursuing. In particular, future studies could make use of the data collected through these EAs to investigate cross-project analyses of volunteer skills.

Our work developing EAs focused on scientific observation skills represents one small step forward toward creating shared measurement systems, and there is much work to be done to assess and evaluate not just skills but a broader range of outcomes, both in citizen science and in contexts beyond. Since we developed EAs aligned with specific scientific inquiry skills, we expect the assessments will be applicable to other environmental citizen science efforts, but they likely cannot be applied to skills beyond targeted nature observation skills. However, we expect that the process for developing a shared EA could be used across a range of other CS projects that have goals similar to each other. For instance, CS projects focused on proteomics and genomics (such as Foldit and Phyllo) are two examples of online puzzle-based projects that rely on pattern recognition and spatial reasoning skills. If these projects worked together to apply the EA development process, it is possible that an EA for pattern recognition could be developed and tested. In this way, we believe that this process could be adapted for many other CS fields beyond environmental science, but this process must be tested in other fields and arenas.

Indeed, the immense challenges associated with measuring skills are not unique to CS. Both formal and informal science education have seen recent calls encouraging researchers and evaluators to begin using performance as a key metric of skill (Bell et al. 2009; Fenichel and Schweingruber 2010; American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014; National Academies of Sciences, Engineering, and Medicine 2018). The EAs developed in this study are a step in this direction, and could make a significant contribution in this area. However, these EAs must be tested more broadly beyond the CS projects included in this study.

DATA ACCESSIBILITY STATEMENTS

To maintain anonymity and confidentiality of study participants, data from this study cannot be made publicly accessible.

SUPPLEMENTARY FILE

The supplementary file for this article can be found as follows:

- **Supplemental File 1.** Sample Interview Questions and Codebook. DOI: https://doi.org/10.5334/cstp.487.s1

ETHICS AND CONSENT

This research was approved by the University of Maryland Institutional Review Board (IRB #1072528). Informed consent was obtained from all participants before being interviewed.

ACKNOWLEDGEMENTS

We thank the citizen science practitioners who contributed to this work and brought many important insights: Alison Cawood, Ashley Cole-Wick, Christine Goforth, Emma Grieg, Gillian Cannataro, Jen Meilinger, Kevin Bonine, LoriAnne Barnett Warren, Michael McKelvey, Robert Costello, Sara Fitzsimmons, and Shelly Grow. We appreciate the support from additional team members (Karen Peterman, Andrea Grover, and Jenna Linhart). We also thank our advisors for their helpful suggestions throughout this research (Anne Bowser, Mac Cannady, Joe Heimlich, and Martin Storksdieck).

FUNDING INFORMATION

This material is based upon work supported by the National Science Foundation under Grant No. DRL-171342.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTION

All authors were involved in the design and data collection of this study. RBK led manuscript writing with VDB, CD, TP, AGN, and ECR reviewing and providing extensive edits.

AUTHOR AFFILIATIONS

Rachel Becker-Klein orcid.org/0000-0002-1456-3491
Two Roads Consulting, US

Cathlyn Davis orcid.org/0000-0002-0968-4336
University of Maryland Center for Environmental Science, US

Tina B. Phillips orcid.org/0000-0002-5010-6052
Cornell Lab of Ornithology, US

Veronica Del Bianco orcid.org/0000-0001-9088-9488
University of Maryland Center for Environmental Science, US

Amy Grack Nelson orcid.org/0000-0003-0620-4621
Science Museum of Minnesota, US

Evelyn Christian Ronning orcid.org/0000-0002-2917-021X
Science Museum of Minnesota, US

REFERENCES

Allen, S and **Peterman, K.** 2019. Evaluating informal STEM education: Issues and challenges in context. *New Directions for Evaluation*, 2019(161): 17–33. DOI: https://doi.org/10.1002/ev.20354

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Ashford-Rowe, K, Herrington, J and **Brown, C.** 2014. Establishing the critical elements that determine authentic assessment. *Assessment and Evaluation in Higher Education*, 39(2): 205–222. DOI: https://doi.org/10.1080/02602938.2013.819566

Beatty, PC and **Willis, GB.** 2007. Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71(2): 287–311. DOI: https://doi.org/10.1093/poq/nfm006

Becker-Klein, R, Peterman, K and **Stylinski, C.** 2016. Embedded assessment as an essential method for understanding public engagement in citizen science. *Citizen Science Theory and Practice*, 1(1): 8. DOI: https://doi.org/10.5334/cstp.15

Bell, P, Lewenstein, B, Shouse, AW and **Feder, MA.** 2009. *Learning Science in Informal Environments: People, Places, and Pursuits*, 140. Washington, DC: National Academies Press

Burgess, HK, DeBey, LB, Froehlich, HE, Schmidt, N, Theobald, EJ, Ettingger, AK, HilleRisLambers, J, Tewksbury, J and **Parrish, JK.** 2017. The science of citizen science: Exploring barriers to use as a primary research tool. *Biological Conservation*, 208: 113–120. DOI: https://doi.org/10.1016/j.biocon.2016.05.014

Davis, C, Del Bianco, V, Peterman, K, Grover, A, Phillips, T and **Becker-Klein, R.** 2022. Diverse and important ways evaluation can support and advance citizen science. *Citizen Science Theory and Practice*, 7(1): 30. DOI: https://doi.org/10.5334/cstp.482

Eberbach, C and **Crowley, K.** 2009. From everyday to scientific observation: How children learn to observe the biologist's world. *Review of educational research*, 79(1): 39–68. DOI: https://doi.org/10.3102/0034654308325899

Fenichel, M and **Schweingruber, HA.** 2010. Surrounded by Science: Learning Science in Informal Environments. National Academies Press.

Fu, AC, Kannan, A and **Shavelson, RJ.** 2019. Direct and unobtrusive measures of informal STEM education outcomes: Direct and unobtrusive measures of informal STEM education outcomes. *New directions for evaluation*, 2019(161): 35–57. DOI: https://doi.org/10.1002/ev.20348

Fu, AC, Kannan, A, Shavelson, RJ, Peterson, L and **Kurpius, A.** 2016. Room for rigor: Designs and methods in informal science education evaluation. *Visitor Studies*, 19(1): 12–38. DOI: https://doi.org/10.1080/10645578.2016.1144025

Grack Nelson, A, Goeke, M, Auster, R, Peterman, K and **Lussenhop, A.** 2019. Shared measures for evaluating common outcomes of informal STEM education experiences: Shared measures for evaluating common outcomes. *New Directions for Evaluation*, 2019(161): 59–86. DOI: https://doi.org/10.1002/ev.20353

Hussar, K, Schwartz, S, Bioselle, E and **Noam, GG.** 2008. *Toward a systematic evidence-base for science in out-of-school time*. Available at: http://ncil.spacescience.org/images/stem-in-libraries/evaluation/Toward-Systematic-EvidenceBase-Science.pdf.

Kim, YJ. (yj), Murai, Y and **Chang, S.** 2021. Implementation of embedded assessment in maker classrooms: challenges and opportunities. *Information and learning science*, 122(3/4): 292–314. DOI: https://doi.org/10.1108/ILS-08-2020-0188

Learning Activation Lab. 2018. *Tools: Measures and data collection instruments*. Available at: http://activationlab.org/tools/ (Accessed: December 5, 2022).

National Academies of Sciences, Engineering, and Medicine. 2018. *Learning Through Citizen Science: Enhancing opportunities by design.* Washington, DC: The National Academies Press. DOI: https://doi.org/10.17226/25183

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

National Research Council. 2010. *Surrounded by science: Learning science in informal environments*. Washington, DC: National Academies Press.

National Research Council. 2015. *Identifying and supporting productive STEM programs in out-of-school settings*. Washington, DC: The National Academies Press.

Newman, I, Lim, J and **Pineda, F.** 2013. Content validity using a mixed methods approach: Its application and development through the use of a table of Specifications methodology, *Journal of Mixed Methods Research*, 7(3): 243–260. DOI: https://doi.org/10.1177/1558689813476922

Noam, GG, Allen, PJ, Shah, AM and **Triggs, B.** 2017. Innovative use of data as game changer for OST programs. *The growing out-ofschool time field: Past, present, and future*, 161–176.

Peterman, K, Becker-Klein, R, Stylinski, C and **Grack Nelson, A.** 2017. Exploring embedded assessment to document scientific inquiry skills within citizen science. In *Citizen Inquiry*, 63–82. Routledge. DOI: https://doi.org/10.4324/9781315458618-5

Peterman, K, Del Bianco, V, Grover, A, Davis, C and **Rosser, H.** 2022. Hiding in Plain Sight: Secondary Analysis of Data Records as a Method for Learning about Citizen Science Projects and Volunteers' Skills. *Citizen Science: Theory and Practice*, 7(1). DOI: https://doi.org/10.5334/cstp.476

Phillips, T, Ferguson, M, Minarchek, M, Porticella, N, Bonney, R, Tessaglia-Hymes, D, Nguyen, L, Shirk, J, Garibay, C, Haley-Goldman, K, Heimlich, J, Lewenstein, B and **Ellenbogen, K.** 2014. *User's guide for evaluating learning outcomes from citizen science*. Ithaca, NY: Cornell Laboratory of Ornithology.

Richards, KAR and **Hemphill, MA.** 2018. A practical guide to collaborative qualitative data analysis. *Journal of Teaching in Physical education*, 37(2): 225–231. DOI: https://doi.org/10.1123/jtpe.2017-0084

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

Shavelson, RJ, Zlatkin-Troitschanskaia, O and **Mariño, JP.** 2018. International performance assessment of learning in higher education (iPAL): Research and development. In *Assessment of Learning Outcomes in Higher Education*, 193–214. Cham: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-74338-7_10

Sloane, K, Wilson, M and **Samson, S.** 1996. *Designing an embedded assessment system: From principles to practice*. Berkeley, CA: Univesity of California.

Stylinski, CD, Peterman, K, Phillips, T, Linhart, J and **Becker-Klein, R.** 2020. Assessing science inquiry skills of citizen science volunteers: a snapshot of the field. *International Journal of Science Education Part B*, 10(1): 77–92. DOI: https://doi.org/10.1080/21548455.2020.1719288

Wiggins, G. 1990. The case for authentic assessment. *Practical Assessment, Research, and Evaluation*, 2(1): 2.

Wilson, M and **Sloane, K.** 2000. From principles to practice: An embedded assessment system. *Applied Measurement in Education*, 13(2): 181–208. DOI: https://doi.org/10.1207/S15324818AME1302_4

TO CITE THIS ARTICLE:
Becker-Klein, R, Davis, C, Phillips, TB, Del Bianco, V, Grack Nelson, A and Ronning, EC. 2023. Using Shared Embedded Assessment Tools to Understand Participant Skills: Processes and Lessons Learned. *Citizen Science: Theory and Practice*, 8(1): 20, pp. 1–12. DOI: https://doi.org/10.5334/cstp.487

Submitted: 20 December 2021 **Accepted:** 10 January 2023 **Published:** 26 April 2023