DEVISE
Developing, Validating, and Implementing Situated Evaluation Instruments

Tina Phillips and Rick Bonney
Cornell Lab of Ornithology, Ithaca NY

Project Goals & Description

DEVISE was conceived to address the need for improved evaluation quality and capacity across the field of citizen science. We envisioned five major goals:

- Inventory extant tools and instruments to measure science and environmental learning
- Develop contextually relevant instruments to measure learning in citizen science
- Implement evaluation strategies with case studies
- Provide professional development opportunities
- Build a community of practice for evaluations of citizen science projects

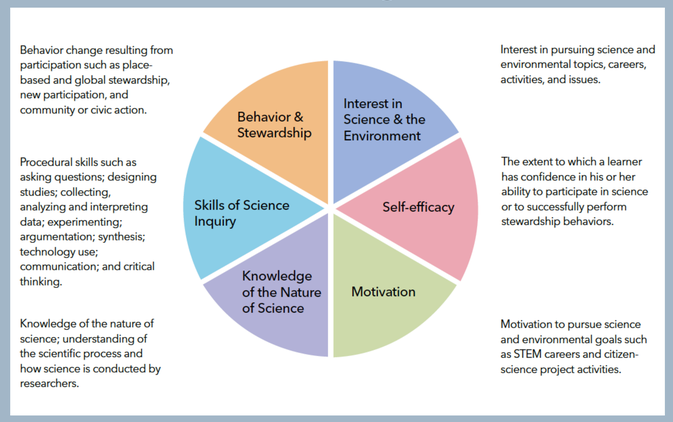
DEVISE has assessed the state of evaluation in citizen science and determined common goals, objectives, and indicators across projects. We inventoried existing instruments, aligned them with the conceptual framework seen at right, and developed and/or modified new and existing evaluation tools. Much of the work of DEVISE has focused on testing and refining these tools with more than 15,000 citizen scientists. We have now entered the professional development phase in which we are actively disseminating these products and building a community of practice for administering these tools. Ultimately, with widespread adoption of these tools, we will be able to conduct cross-programmatic comparisons to determine field-wide outcomes from citizen science participation.

Scale Construction & Validation

1. Clearly define what is to be measured
2. Draft initial items
3. Expert rating of individual items, revise as necessary
4. Pilot test draft scale to 8-10 people similar to target audience via "think alouds," revise as necessary
5. Field test to larger community
6. Construct Validity - Statistical tests
   - Reliability (internal, test/retest, split half)
   - Factor analysis (factor reduction)
   - Item Response Theory (IRT)
7. Criterion-Related Validity Checks
   - Convergent: Test whether the scale aligns with other similar constructs.
   - Concurrent: Test whether scale can discriminate between two populations that should be different.
   - Predictive: Test the scale's ability to predict something it should theoretically be able to predict.
   - Discriminant: Test whether the scale construct is not similar to something that theoretically it should not be similar to.
8. Revise as necessary

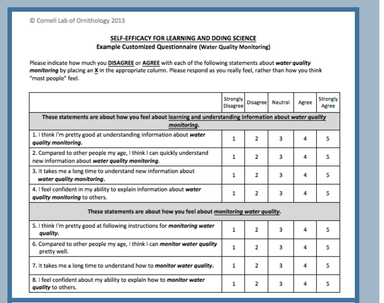
Products

Framework for Evaluating Individual Learning Outcomes



Custom & Generic Scales

Example of custom scale

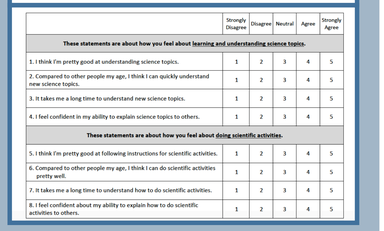

© Cornell Lab of Ornithology 2013

SELF-EFFICACY FOR LEARNING AND DOING SCIENCE
Example Customized Questionnaire (Water Quality Monitoring)

Please indicate how much you **DISAGREE** or **AGREE** with each of the following statements about ***water quality monitoring*** by placing an **X** in the appropriate column. Please respond as you really feel, rather than how you think "most people" feel.

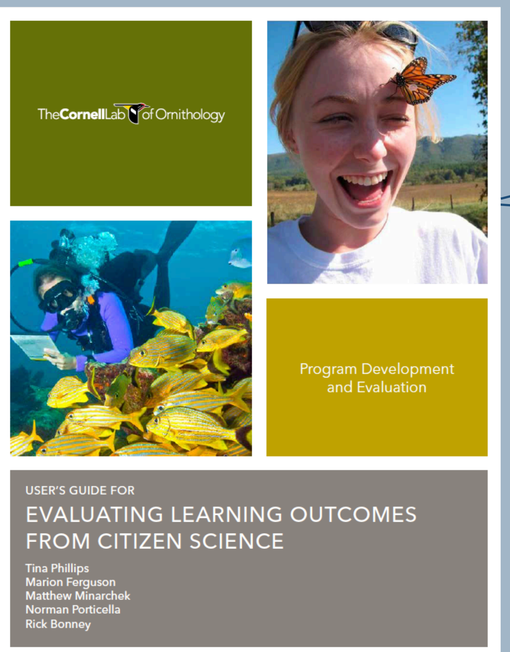
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
These statements are about how you feel about learning and understanding information about *water quality monitoring*.					
1. I think I'm pretty good at understanding information about ***water quality monitoring***.	1	2	3	4	5
2. Compared to other people my age, I think I can quickly understand new information about ***water quality monitoring***.	1	2	3	4	5
3. It takes me a long time to understand new information about ***water quality monitoring***.	1	2	3	4	5
4. I feel confident in my ability to explain information about ***water quality monitoring*** to others.	1	2	3	4	5
These statements are about how you feel about *monitoring water quality*.					
5. I think I'm pretty good at following instructions for ***monitoring water quality***.	1	2	3	4	5
6. Compared to other people my age, I think I can ***monitor water quality*** pretty well.	1	2	3	4	5
7. It takes me a long time to understand how to ***monitor water quality***.	1	2	3	4	5
8. I feel confident about my ability to explain how to ***monitor water quality*** to others.	1	2	3	4	5

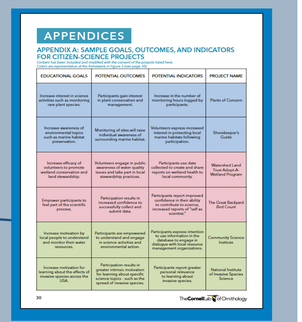
Example of generic scale

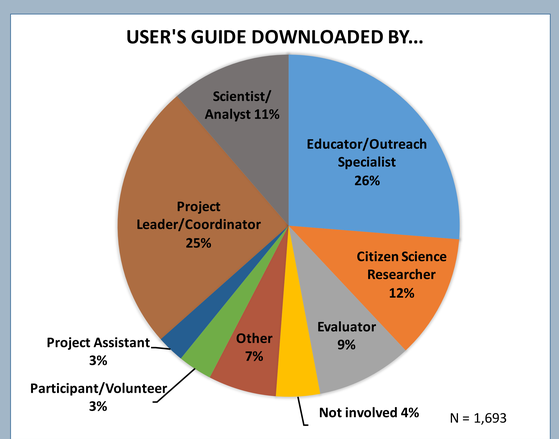


	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
These statements are about how you feel about learning and understanding science topics.					
1. I think I'm pretty good at understanding science topics.	1	2	3	4	5
2. Compared to other people my age, I think I can quickly understand new science topics.	1	2	3	4	5
3. It takes me a long time to understand new science topics.	1	2	3	4	5
4. I feel confident in my ability to explain science topics to others.	1	2	3	4	5
These statements are about how you feel about doing scientific activities.					
5. I think I'm pretty good at following instructions for scientific activities.	1	2	3	4	5
6. Compared to other people my age, I think I can do scientific activities pretty well.	1	2	3	4	5
7. It takes me a long time to understand how to do scientific activities.	1	2	3	4	5
8. I feel confident about my ability to explain how to do scientific activities to others.	1	2	3	4	5

Free Downloadable User's Guide







All products available for free download at: Citizenscience.org/evaluation

Acknowledgments: Funding support provided by the National Science Foundation (DRL # 1010744) and the Noyce Foundation. We greatly appreciate the support of Kirsten Ellenbogen and Candie Wilderman (Co-PIs), Joe Heimlich (COV Chair), Norman Porticella, Amy Grack Nelson, Marion Ferguson, and the rest of the DEVISE team. Special thanks to the thousands of participants involved in our research.

Audience

This work was originally intended to provide citizen science practitioners and ISE researchers with easy to use tools that, in combination with other tools, can facilitate high-quality evaluations. The tools have since been downloaded and used by a variety of professionals and disciplines beyond citizen science.

Results

Scale Name	Type	Psychometrics	Custom Version?	Youth Version?
Interest in Science	12-items, Likert-type 5 pt.	Internal Reliability = .93; EFA: unidimensional, all items load at >.30;	✘	✔
Self-Efficacy for Learning and Doing Science	8 items, Likert-type 5 pt.	Internal Reliability = .92; EFA: unidimensional, all items load at >.70; Test-Retest: all Pearson's r's > .30, all p's < .05	✔	✔*
Self-Efficacy for Environmental Action	8 items, Likert-type 5 pt.	Internal Reliability = .89; EFA: unidimensional, all items load at >.70; Test-Retest: all Pearson's r's > .49, all p's < .001	✔	✔*
Motivation for Learning and Doing Science	16 items, Likert-type 5 pt.	Internal Reliability =.81/.85; EFA: 2 Factors (Internal/External Motives) all items load at >.50; Test-Retest Reliability: all Pearson's r's > .33, all p's < .05	✔	✔*
Motivation for Environmental Action	16 items, Likert-type 5 pt.	Internal Reliability =.84/.75; EFA: 2 Factors (Internal/External Motives) all items load at >.40; Test-Retest Reliability: all (Internal) Pearson r's > .29, all p's < .01; all (External) r's > .39, all p's < .001	✔	✔*
Skills of Science Inquiry*	12 items, Likert-type 5 pt.	Internal Reliability =.89; EFA: 2 factors, all items load at >.40; IRT analysis: discriminant scores between .479 and .70 for all	✔	✔*
Data Interpretation Skills*	9 multiple choice questions	Internal Reliability between .399-.445 for three groups of questions; IRT: low discrimination; EFA: poor factor loadings	✘	✘
Environmental Stewardship Scale*	24 items, 7 pt. responses	Internal Reliability = 881; CFA: 5 factor solution, 22/24 load >.40;	✘	✘

**Denotes scales still in development or testing. Psychometric results provided for adult versions of scales only.*

Challenges

- Creating "generalized" STEM tools that are sensitive enough to detect change and capture long-term effects of participation in informal settings.
- The time and resources needed to successfully conduct psychometric testing to develop valid and reliable instruments.
- Creating a quantitative scale to measure the knowledge of Nature of Science.
- Tracking usage and behavior of the scales after dissemination.