CUR Quarterly
COUNCIL ON UNDERGRADUATE RESEARCH

CUR Focus

Sandra L. Laursen,
University of Colorado Boulder

Assessing Undergraduate Research in the Sciences: The Next Generation

Inherent in the conceptualization of the apprentice model of undergraduate research (UR) is a fundamental tension between the educational goals of UR and its foundation in faculty scholarship (Laursen, Seymour and Hunter 2012). This tension in goals leads to challenges for faculty in guiding undergraduate researchers in their daily work, as well as in positioning faculty's own UR work within the institutionally bifurcated domains of teaching and research.

In interviews, faculty UR advisors explain how the task of crafting and supervising developmentally appropriate projects is "very much a teaching thing," as one faculty interviewee put it, and faculty recognize and make use of the many rich learning opportunities for students that are embedded in real research problems. The teaching aspect of UR is also a source of great pleasure, pride, and learning for faculty as they see their student researchers' progress and become independent. But faculty also describe challenges: They must consider carefully how to build and support their own overall scholarly trajectory given the constraints of students' slower pace, variable progress, and the need to cobble together the contributions of multiple "short-term helpers in a long-term enterprise." While faculty members navigate these tensions in their everyday work as UR advisors, institutional descriptions often over-simplify UR as either education-focused or as scholarly work.

This tension between the educational and scholarly purposes of UR also generates challenges when it comes to measuring the outcomes of UR. Traditionally, at least in the sciences where UR is most established, institutions have counted their successes in terms of student researchers' scholarly contributions—such as numbers of student-coauthored publications and presentations—and research-oriented career choices, especially the number of students who go on to pursue graduate degrees in a similar field. Such measures help to identify the value of students' contributions to new knowledge; they call out the importance of maintaining the scholarly engagement of faculty (especially at primarily undergraduate institutions) and of developing the skilled research workforce in scientific disciplines. Yet these measures may overstate the role of scholarly publication, which is valued by faculty and institutions but which research to date has not linked to the quality or extent of students' educational outcomes. Indeed, deep learning from authentic undergraduate research experiences requires that students have opportunities to try out their own ideas, make mistakes, and try again. At the same time, giving students such independence may also slow faculty's acquisition of publishable results (Laursen, Hunter, Seymour, Thiry and Melton 2010). And, by counting only those students who go on to graduate school, we undervalue other contributions to the nation's workforce and electorate, such as developing research-literate technicians, science teachers, physicians, parents, and citizens.

Gaya Shivega working in the lab at Concordia College. (Photo credit: Paige Borst)

In this article, I analyze these and other reasons that assessing the outcomes of the apprentice model in undergraduate research is an inherently difficult proposition. These challenges mean that development of new approaches to assessing undergraduate research is itself a topic for research. I note strengths and limitations of the assessment approaches tried to date and suggest other questions and approaches for individual investigators and the community at large to consider, with the goal of prompting thinking and inspiring experimentation that will lead to the next generation of UR assessment tools.

My focus here is on the traditional, intensive model of undergraduate research in which students pursue a multi-week, open-ended scientific project outside class and under the guidance of a faculty member and other, more experienced, researchers in the research group. Many students engage in apprentice-model UR as an immersive summer project, but they may also participate during the academic year. However, I note useful lessons that may be learned from assessing outcomes of research-based courses. In this article, I focus on scientific disciplines in which the apprentice model is common and in which most previous research and evaluation of student outcomes has been carried out. Finally, I use the term "assessment" to refer to any measurement of UR student outcomes, which may use a variety of methods such as surveys, tests, interviews, or review of students' research products, and which may be carried out for purposes of research, evaluation, or monitoring.

Measuring Outcomes: A Worthy Challenge

Assessing outcomes of undergraduate research is of keen interest to faculty, administrators, funders, and policymak-

ers. Indeed, faculty who work with undergraduate researchers have long known and valued the learning they observe among their students, and more recently research has begun to document these outcomes (see reviews in Laursen et al. 2010; Sadler, Burgin, McKinney and Ponjuan 2010; Crowe and Brakke 2008). Large-scale studies identify UR as one of several "high-impact" practices that foster deep learning and persistence in college, especially for students from underrepresented minority groups (Eagan et al. 2013; Hurtado, Cabrera, Lin, Arellano and Espinosa 2009; Kuh 2008; see also studies reviewed in Laursen et al. 2010). In this era of accountability, funders of UR seek good information about the value of their investment; departments and institutions are interested in pinpointing the educational contributions of these important out-of-class experiences to their overall program; and faculty have a stake in seeing that their UR work is indeed recognized for the educational value it delivers.

At the same time, assessing UR outcomes in uniform ways is challenging. Both students and their research advisors experience UR differently, depending on their discipline and its intellectual and pragmatic ways of working. Even within the same research group, each student's outcomes will differ depending on the nature and stage of her individual project and its relation to ongoing work in the laboratory, as well as her own characteristics and background. Moreover, many of the most valued outcomes of research activity are not only highly contextual but also inherently difficult to define and measure, such as understanding the nature of science or of scientific inquiry (Lederman 1992; Lederman et al. 2014). There are also measurement and sampling challenges: In any department, program, or institution, the number of participants is often small, and both institutional and self-selection influence which students have the chance to participate, making comparisons of UR participants with non-participants problematic. Alas, probing the outcomes of undergraduate research is not as simple as sticking a probe in students' ears or scanning them with a Starfleet tricorder.

Strengths and Weaknesses of Current Studies

The first generation of UR assessments has been derived from education research that has documented students' personal and professional learning as a result of UR, including their acquisition of new skills and conceptual understandings of their field and of disciplinary inquiry, growth in confidence and responsibility, and development of a scientific identity (Laursen et al. 2010, and studies cited within). This body of work helps to balance prior emphasis on students' scholarly contributions by recognizing the strong educational role of UR. Interview studies, in particular, reveal the types and depth of student learning from UR and find commonalities in outcomes across multiple disciplines and research settings. Often based on long-established and well-designed examples of apprentice-model UR, such studies help by identifying student outcomes that result from best-case scenarios; these outcomes then guide what can be searched for when examining programs with other designs or durations. Interview data also capture the language students themselves use to express their nascent understanding of complex ideas about the scientific process and the nature of the knowledge it generates, or use to describe their developing identities. Use of a semi-structured interview protocol enables interviewers to identify and probe emergent issues—whether benefits not anticipated or difficulties not perceived by program designers. Thus interviews and focus groups remain a useful tool for program evaluation, but the time commitment and cost of data analysis are barriers to their routine use.

Survey instruments based on these interview-derived findings seek to capture these gains in a holistic manner from students' perspective, asking students to self-report their gains across multiple domains. Compared to interviews and focus groups, instruments such as the Undergraduate Research Student Self-Assessment (URSSA, Hunter, Weston, Laursen and Thiry 2009) and the Survey of Undergraduate Research Experiences (SURE, Lopatto 2004) are inexpensive and easy to use, and thus complement other sources of information. Such instruments have advantages in that the set of items covers many learning domains already identified through qualitative research; responses can be compared over time or across programs and linked to particular experiences and activities that students also report. Self-report is an obvious way to probe student gains that are personal, internal, and not easily tested, such as changes in students' career plans or the growth in confidence that is so important for student researchers. A recent validity study of URSSA based on more than 3,600 student responses supports the reliability and validity of the four main categories of student gains captured by that survey, but that study also points to possible improvements that can be made to improve the sensitivity and discrimination of this instrument (Weston and Laursen 2014).

To measure other gains, such as research competencies and skills, self-assessment may be less satisfactory; the literature does not show a predictable relationship of self-report to criterion-referenced measures such as tests of knowledge or experts' rating of student skills (Boud and Falchikov 1989). Poor survey design, such as when students do not understand the questions or do not have the relevant knowledge to answer them, can compromise survey items. Careful development using approaches such as think-aloud interviews is needed to craft relevant items and to determine whether the intended audience can answer them (Hunter et al. 2009). And self-reports of competencies work best when people have received

feedback on their progress and abilities; indeed, some level of skill in a domain is required in order to evaluate one's own competence in that domain (Kruger and Dunning 1999). The availability of such feedback may be variable (and certainly unsystematic) for research students, especially in domains new to them.

Different risks to the reliability of self-report arise when consequences such as grades, money, or advancement depend upon students' responses to survey questions (Albanese et al. 2006). For UR, candid responses are more likely when students can respond anonymously and have no stake in how the results are used. Social-desirability bias can arise when students feel pressure from their research advisor or program director to answer a certain way. We commonly observe pitfalls of these types, especially those that compromise student anonymity, when surveys are used in evaluating UR programs. For example, small numbers of participants mean that individuals may be easily identified from demographics; consequences such as stipend payments may be tied to completion of the survey; programs want to link survey responses to long-term outcomes so program administrators do not wish to give surveys anonymously, yet non-anonymous survey answers may be biased because students continue to depend on their advisors for recommendations and support. Thus those who lead UR programs express a need for other assessment tools to augment students' self-reports.

In addition to the measurement problems inherent in survey approaches, issues of research design often surface in existing studies of student outcomes from undergraduate research. Eagan and coauthors (2013) note that such studies too often generalize from small, non-representative samples, draw on retrospective reflections (e.g., of program alumni) rather than real-time probes, and do not properly account for selection bias both in who chooses to apply and who is admitted to UR programs. In their own study examining the relationship between UR and students' intention to enroll in STEM graduate programs, Eagan et al. (2013) mitigate some of these problems through the use of data on students' degree aspirations from an anonymous, nationwide survey administered in students' first year and again in their senior year of college.

While aspiring to an advanced degree is not the same as earning one, the authors argue that aspiring to that degree is a necessary first step, citing research showing that intention to pursue a graduate degree is the strongest predictor of eventual enrollment in a graduate or professional program. With more than 4,000 students in their sample, these authors were able to apply sophisticated statistical methods, using propensity scoring and hierarchical modeling with student- and institution-level covariates to statistically control for student self-selection into UR programs. They conclude that UR participation had a significant positive effect on undergraduate STEM majors' intent to pursue STEM-related postgraduate study, increasing this likelihood by between 14 and 17 percent compared with non-participants.

Steps Forward: Tighter Focus in Research Studies of UR

The study by Eagan and coauthors cited above is a useful example because it points the way toward one promising strand of future work on assessment of UR. This carefully designed study focuses on a single outcome, in this case intent to pursue graduate study. By probing at two distinct times students' self-reported intentions to pursue graduate study, the authors could identify changes in students' thinking in a less biased manner than would be possible from retrospective self-report, and by carefully controlling for influences other than UR, they could attribute changes in students' plans for STEM graduate study to UR experience. Other studies in this vein might similarly target specific outcomes or narrow domains, drawing on prior work to develop and validate surveys or tests for domains already known to be related to UR experiences, such as:

- formation of a scientific identify (e.g., Estrada, Woodcock, Hernandez and Schultz 2011),
- project ownership (e.g., Hanauer and Dolan 2014),
- understandings of the nature of scientific inquiry (e.g., Lederman et al. 2014),
- student beliefs about science (e.g., Adams et al. 2006), and
- experimental design (e.g., Dasgupta, Anderson and Pelaez 2014).

Other important constructs might include creativity, persistence, and invention. Because students develop in these domains as a result of many types of experiences, not just UR, good data about students' prior experiences and background are also needed if the goal is to establish a causal relationship between a particular outcome and the UR experience.

Initial work to develop or adapt and test such targeted assessments might be carried out in venues other than apprentice-model research environments, especially in inquiry-driven or research-based courses (Auchincloss et al. 2014; Gasper, Minchella, Weaver, Csonka and Gardner 2012; Dasgupta, Anderson and Pelaez 2014). Research-based courses offer certain advantages for exploring the domain and for testing measurement approaches, such as larger sample sizes and faster iteration times, and, at least within a given course, a more standardized and less context-based intervention. Because some of these tools have been developed to study science learners who are less experienced than the typical UR student (Adams et al. 2006; Lederman et al. 2014), additional up-front work will be required to determine whether and how the instruments and methods are useful in detect-

ing outcomes of apprentice-model UR, and whether they apply across varied scientific approaches and disciplines (see Schwartz and Lederman 2008). In some domains, further research to identify and generalize (if possible) elements of advanced learning in the domain may be needed.

For example, Dasgupta, Anderson, and Pelaez (2014) review existing literature on experimental design to categorize student difficulties in the ability to design experiments, drawing on studies of middle-school, high-school, and college students. They apply this categorization together with some existing assessments to diagnose student difficulties with experimental design and to measure changes in this ability as the result of a college course. This study offers a model of careful thinking about a specific domain—in this case, experimental design—and validation of a measurement approach, while also raising the question: Are there other difficulties or understandings of experimental design that would be identified in a sample of more experienced researchers? As the authors point out, this type of assessment could also help to identify the processes or experiences through which students learn to design experiments and could lead to development of good interventions to teach that skill. Thus, while I argue that this type of targeted study offers one promising strand to follow in preparing the next generation of UR assessments, it is likely to require work by experienced educational researchers, and would not be easily carried out within single UR programs by the science faculty who run them.

Steps Forward: Approaches to Program Evaluation

A research agenda that turns attention to specific outcomes will ultimately produce results useful to practitioners. But in the meantime, practitioners will continue to need assessment tools for evaluating local programs. These tools must yield data of a depth and quality that help practitioners to monitor, improve, and justify their programs but need not presume a standardized approach across units or institutions. Assessment for program evaluation should focus on questions about what is good and what can be improved about the local UR program, not on comparing or trying to generalize student outcomes. Program directors are likely to prefer holistic or broad approaches that do not focus on one outcome to the neglect of others. Familiarity with the research literature may be helpful in identifying which outcomes are most likely and what program elements give rise to them; it also helps in making "golden spike" arguments that connect local evaluation of practice to evidence from research (Urban and Trochim, 2009). Measurement issues, such as the small samples, student selection, and self-report biases discussed above, may limit the claims that can be made relative to other programs, but such problems do not invalidate the worth of knowing, rather than assuming, what happens in one's own program and why.

To prompt creativity in this type of local program evaluation, I suggest some other sources of information that are amenable to local use and that examine multiple or broad domains. Some of these are straightforward, and some are better suited for those ready to explore more deeply. Sources of information may include:

1. Reflections from students, such as exit interviews, a facilitated group discussion, or a personal reflective essay. These activities offer intriguing opportunities for spurring student metacognition about their research experiences, at the same time that they document student perspectives and help students to recognize gains as they develop graduate school or job applications. Singer and Zimmerman (2012) note such metacognitive benefits from the repeated use of a student self-evaluation rubric in concert with faculty ratings of students on the same rubric.

2. Faculty-developed rubrics applied to students' research work or research products, such as abstracts (perhaps both a technical abstract and a general-audience summary), posters, or talks. In speaking with faculty UR advisors at liberal arts colleges (Laursen et al. 2010), our research team found that faculty could offer sophisticated judgments of students' research skills and capacities, for example when they wrote letters of recommendation, but that faculty colleagues did not generally have ways to standardize these so that they could compare research skills among students who worked with different advisors. Thus the process of coming to consensus on this rubric could itself be a valuable exercise for some departments. Dahm, Newell, and Newell (2003) describe how developing a rubric for a semester-long team engineering "clinic" course provided greater clarity to students and faculty alike about course goals and student achievement of these goals.

3. Broad tests of integrated content knowledge. Content knowledge is not often the focus of UR assessment, but students commonly report growth in their depth of understanding of disciplinary concepts and in their ability to connect concepts across disciplines or subdisciplines. Could that growth be detected by appropriate instruments? One interesting disciplinary example is a test offered by the American Chemical Society Exams Institute, the Diagnostic of Undergraduate Chemistry Knowledge (DUCK) (http://chemexams.chem.iastate.edu/exam-details?id=41783).

4. Oral exams by outside experts (Wright et al., 1998). In this study, oral exams designed and given by outside faculty examiners were used to judge the competence of students who had taken one of two versions of a course, one more collaborative and project-oriented

and the other lecture-based. The examiners developed their own questions and did not know which course each student had taken, but reported the greatest difference in student skills when they chose to focus their assessment on students' problem-solving abilities. The study shows, the authors argue, that "it is possible to measure in an unbiased and quantitative way the extent to which the goal of increasing student competence can be achieved." Could this approach be applied to settings such as UR to assess general competency in research thinking or problem-solving?

5. A normed test of transferable critical thinking. For example, the Critical Thinking Assessment Test (CAT) examines a set of critical-thinking skills valued by faculty across disciplines (Stein and Haynes 2011; https://www.tntech.edu/cat/). This carefully developed test is said to be sensitive both to course-level changes in student skills and to changes across the college career. Could this test offer useful data to a department in considering how its majors develop these skills over time—not only through research but also through other experiences in the major?

6. Systematic tracking of participants. Records of student involvement in presenting and publishing research and documentation of students' graduate and career outcomes are likely to continue to hold value locally. Departments can improve their practice by being systematic in tracking and proudly reporting outcomes for all their majors—not just those who go on to graduate school.

As an example, imagine a chemistry department that seeks evidence to offer an accrediting body that its summer undergraduate research activity is an important and meritorious part of its educational activities. The faculty members decide to administer the URSSA anonymously to measure student gains and the learning experiences that give rise to them. They also ask students to write an individual one-page reflection that prepares them for an end-of-summer group debriefing session facilitated by a colleague from the campus teaching center. These combined approaches provide the faculty with a picture that is both broad and deep; they offer additional benefit by inviting students' reflection and metacognition about their research experience. The department members gather to review the outcomes data and identify what is good for students' growth as chemists, as communicators and team members, as future professionals, and as science-literate world citizens. Later on, the data help the department to refine its criteria for how student research work is weighed in considering departmental honors.

Our imaginary chemists also decide to keep track more systematically of their graduates' career paths, not just rely on graduates to update their faculty mentors of their own accord. In examining data on students' experiences in their UR program, the faculty members learn what can be improved and identify small ways in which their summer activities can be refined to build community among summer research students, increase students' communication skills, and adjust the requirements to smooth the path for students who wish to submit their UR work for departmental honors. Their goal is not to demonstrate that their UR program is better than the one in the math department or than the one down the road at a neighboring institution, but rather to make explicit the value that UR adds to their own major. Yet the simplicity and success of their approach leads to a feature about the chemistry department on an institutional web site about engaged teaching and learning. Ideas like these for assessing and improving a local program are no means novel, but this scenario shows how they might augment current practices on many campuses.

These approaches are likely to work best when integrated into an overall approach to assessment that looks broadly at student outcomes in the major. Yet because it offers both educational and scholarly contributions to the organizational mission, undergraduate research may be a good place to begin such conversations about assessment. Institutions and funding agencies should recognize that there is value in experimenting with different approaches to outcomes assessment and developing assessment "habits of mind," even when sample size and other constraints prevent the resulting data from meeting publication-level standards of research design. As they assess and communicate about their UR programs with students, colleagues, and administrators, program directors and scientists who work with undergraduate researchers can value both scholarly achievements and educational outcomes, and thus honor the special types of teaching and learning processes by which these are achieved.

Acknowledgments

Joanne Stewart and Tim Weston provided helpful comments on this essay. The author also thanks Heather Thiry, Chuck Hayward, Janet Branchaw, Julio Soto, Christine Jones, Erin Dolan, and the CUREnet assessment working group for helpful conversations that informed the ideas presented here. This work was supported in part by the National Science Foundation, Division of Chemistry, DUE, Biological Sciences Directorate, and the Office of Multidisciplinary Affairs, under grant #CHE-0548488, and by the Division of Biological Infrastructure, Directorate for Biological Sciences, under grant #DBI-1052683.

References

Adams, Wendy K., Katherine K. Perkins, Noah S. Podolefsky, Michael Dubson, Noah D. Finkelstein, and Carl E. Wieman. 2006. "New Instrument for Measuring Student Beliefs about Physics and Learning Physics: The Colorado Learning Attitudes about Science Survey." *Physical Review Special Topics-Physics Education Research* 2(1): 010101-1—010101-14.

Albanese, Mark, Susan Dottl, George Mejicano, Laura Zakowski, Christine Seibert, Selma VanEyck, and Carolyn Prucha. 2006. "Distorted Perceptions of

Competence and Incompetence are More than Regression Effects." *Advances in Health Sciences Education* 11: 267-278.

Auchincloss, Lisa C., Sandra L. Laursen, Janet L. Branchaw, Kevin Eagan, Mark Graham, David I. Hanauer, Gwendolyn Lawrie, Collen M. McLinn, Nancy Pelaez, Susan Rowland, Marcy Towns, Nancy M. Trautmann, Pratibha Varma-Nelson, Timothy J. Weston, and Erin L. Dolan. 2014. "Assessment of Course-Based Undergraduate Research Experiences—A Meeting Report." *CBE-Life Sciences Education* 13(1): 29-40. doi: 10.1187/cbe.14-01-0004

Boud, David, and Nancy Falchikov. 1989. "Quantitative Studies in Student Self-Assessment in Higher Education: A Critical Analysis." *Higher Education* 18: 529-549.

Crowe, Mary, and David Brakke. 2008. "Assessing the Impact of Undergraduate-Research Experiences on Students: An Overview of Current Literature." *CUR Quarterly* 28(1): 43-50.

Dahm, Kevin D., James A. Newell, and Heidi L. Newell. 2003. "Rubric Development for Assessment of Undergraduate Research: Evaluating Multidisciplinary Team Projects." *Proceedings, 2003 American Society for Engineering Education Conference.*

Dasgupta, Annwesa P., Trevor R. Anderson, and Nancy Pelaez. 2014. "Development and Validation of a Rubric for Diagnosing Students' Experimental Design Knowledge and Difficulties." *CBE-Life Sciences Education* 13(2): 265-284.

Eagan, M. Kevin, Jr., Sylvia Hurtado, Mitchell J. Chang, Gina A. Garcia, Felisha A. Herrera, and Juan C. Garibay. 2013. "Making a Difference in Science Education: The Impact of Undergraduate Research Programs." *American Educational Research Journal* 50: 463-713.

Estrada, Mica, Anna Woodcock, Paul R. Hernandez, and Wesley P. Schultz. 2011. "Toward a Model of Social Influence that Explains Minority Student Integration into the Scientific Community." *Journal of Educational Psychology* 103(1): 206-222.

Gasper, Brittany J., Dennis J. Minchella, Gabriela C. Weaver, Lastio N. Csonka, and Stephanie M. Gardner. 2012. "Adapting to Osmotic Stress and the Process of Science." *Science* 335(6076): 1590-1591.

Hanauer, David I., and Erin L. Dolan. 2014. "The Project Ownership Survey: Measuring Differences in Scientific Inquiry Experiences." *CBE-Life Sciences Education* 13(1):149-158.

Hunter, Anne-Barrie, Timothy J. Weston, Sandra L. Laursen, and Heather Thiry. 2009. "URSSA: Evaluating Student Gains from Undergraduate Research in Science Education." *Council on Undergraduate Research Quarterly* 29(3): 15-19.

Hurtado, Sylvia, Nolan L. Cabrera, Monica H. Lin, Lucy Arellano, and Lorelle L. Espinosa. 2009. "Diversifying Science: Underrepresented Student Experiences in Structured Research Programs." *Research in Higher Education* 50(2): 189-214.

Kruger, Justin, and David Dunning. 1999. "Unskilled and Unaware of It: How Difficulties in Recognizing one's Own Incompetence Lead to Inflated Self-Assessments." *Journal of Personality and Social Psychology* 77(6): 1121-1134.

Kuh, George D. 2008. *High-impact Educational Practices: What They Are, Who has Access to Them, and Why They Matter.* Washington, D.C.: AAC&U.

Laursen, Sandra, Anne-Barrie Hunter, Elaine Seymour, Heather Thiry, and Ginger Melton. 2010. *Undergraduate Research in the Sciences: Engaging Students in Real Science.* San Francisco: Jossey-Bass.

Laursen, Sandra, Elaine Seymour, and Anne-Barrie Hunter. 2012. "Learning, Teaching and Scholarship: Fundamental Tensions of Undergraduate Research." *Change: The Magazine of Higher Learning* 44(2): 30-37.

Lederman, Norman G. 1992. "Students' and Teachers' Conceptions of the Nature of Science: A Review of the Research." *Journal of Research in Science Teaching* 29(4): 331-359.

Lederman, Judith S., Norman G. Lederman, Stephen A. Bartos, Selina L. Bartels, Allison A. Meyer, and Renee S. Schwartz. 2014. "Meaningful Assessment of Learners' Understandings about Scientific Inquiry—The Views About Scientific Inquiry (VASI) Questionnaire." *Journal of Research in Science Teaching* 51(1): 65-83.

Lopatto, David. 2004. "Survey of Undergraduate Research Experiences (SURE): First Findings." *Cell Biology Education* 3(4): 270-277.

Sadler, Troy D., Stephen Burgin, Lyle McKinney, and Luis Ponjuan. 2010. "Learning Science Through Research Apprenticeships: A Critical Review of the Literature." *Journal of Research in Science Teaching* 47(3): 235-256.

Schwartz, Renee, and Norman Lederman. 2008. "What Scientists Say: Scientists' Views of Nature of Science and Relation to Science Context." *International Journal of Science Education* 30(6): 727-771.

Singer, Jill, and Bridget Zimmerman. 2012. "Evaluating a Summer Undergraduate Research Program: Measuring Student Outcomes and Program Impact." 32(3): 40-47.

Stein, Barry, and Ada Haynes. 2011. "Engaging Faculty in the Assessment and Improvement of Students' Critical Thinking using the Critical Thinking Assessment Test." *Change: The Magazine of Higher Learning* 43(2): 44-49.

Urban, Jennifer Brown, and William Trochim. 2009. "The Role of Evaluation in Research—Practice Integration Working Toward the 'Golden Spike'." *American Journal of Evaluation* 30(4): 538-553.

Weston, Timothy J., and Sandra L. Laursen. 2014. "The Undergraduate Research Student Self-Assessment (URSSA): Development, Use and Validation." Manuscript in preparation.

Wright, John C., Susan B. Millar, Steve A. Kosciuk, Debra L. Penberthy, Paul H. Williams, and Bruce E. Wampold. 1998. "A Novel Strategy for Assessing the Effects of Curriculum Reform on Student Competence." *Journal of Chemical Education* 75(8): 986-992.

Sandra Laursen

University of Colorado Boulder, Sandra.Laursen@colorado.edu

Sandra Laursen is senior research associate and co-director of ethnography & evaluation research at the University of Colorado Boulder where she leads research and evaluation studies focusing on education and career paths in science, technology, engineering, and mathematics (STEM) fields. Her particular research interests include the underrepresentation of women and people of color in the sciences, the professional socialization and career development of scientists, teacher professional development, and organizational change in higher education. She is also interested in inquiry-based teaching and learning, and the challenges of improving STEM education in and out of the classroom and across organizations. Laursen is primary author of Undergraduate Research in the Sciences: Engaging Students in Real Science *(Jossey-Bass 2010), describing a large research study of faculty and student participants in undergraduate research, and she is a coauthor of the Undergraduate Research Student Self-Assessment (URSSA) instrument. Laursen earned a PhD in chemistry from the University of California, Berkeley.*