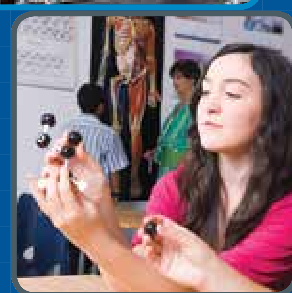
DESCRIBING & MEASURING UNDERGRADUATE STEM TEACHING PRACTICES



A Report from a National Meeting on the Measurement of Undergraduate Science, Technology, Engineering and Mathematics (STEM) Teaching

17–19 DECEMBER 2012

NSF

AAAS
ADVANCING SCIENCE, SERVING SOCIETY

About AAAS

The American Association for the Advancement of Science (AAAS) is the world's largest general scientific society, and publisher of the journal, *Science* (www.sciencemag.org) as well as *Science Translational Medicine* (www.sciencetranslationalmedicine.org) and *Science Signaling* (www.sciencesignaling.org). AAAS was founded in 1848, and includes some 261 affiliated societies and academies of science, serving 10 million individuals. *Science* has the largest paid circulation of any peer-reviewed general science journal in the world, with an estimated total readership of 1 million. The non-profit AAAS (www.aaas.org) is open to all and fulfills its mission to "advance science and serve society" through initiatives in science policy; international programs; science education; and more. For the latest research news, log onto EurekAlert!, (www.eurekalert.org), the premier science-news Web site, a service of AAAS.

NSF

AAAS
ADVANCING SCIENCE, SERVING SOCIETY

DESCRIBING & MEASURING UNDERGRADUATE STEM TEACHING PRACTICES

A Report from a National Meeting on the Measurement of Undergraduate Science, Technology, Engineering, and Mathematics (STEM) Teaching
December 17–19, 2012

Hosted by the American Association for the Advancement of Science (AAAS) with support from the National Science Foundation (NSF)

This material is based upon work supported by the National Science Foundation under grant No. DUE 1252972.

ISBN #: 978-0-87168-756-2

MEETING ORGANIZERS

Charles Henderson
Physics Department and Mallinson Institute for Science Education
Western Michigan University

Yolanda George
Education and Human Resources Programs
American Association for the Advancement of Science

WHITE PAPER AUTHORS

Melissa H. Dancy
Physics Education Research Group
University of Colorado

Matthew T. Hora and Joseph J. Ferrare
Wisconsin Center for Education Research
University of Wisconsin–Madison

Ellen Iverson
Science Education Resource Center
Carleton College

Lisa R. Lattuca
Center for the Study of Higher and Postsecondary Education
University of Michigan

Jennifer Turns
Human Centered Design and Engineering
University of Washington

WORKING GROUP FACILITATORS

Scott Franklin
Physics Department
Rochester Institute of Technology

Scott Grissom
School of Computing and Information Systems
Grand Valley State University

Sandra Laursen
Ethnography Evaluation Research
University of Colorado–Boulder

Thomas Litzinger
Leonhard Center for Enhancement of Engineering Education
Penn State University

Karen Marrongelle
Chancellor's Office
Oregon University System

Publications Staff:
Diane Smith, writing and editing
Brian Baker, copy editing and proofreading, Write-With, Inc.
Gail Peck, layout and design, Peck Studios, Inc.
Janel Kiley, art direction, AAAS Office of Public Programs

Art collaboration by Jack Biesek, Gladys Brenner, Margaret Faye, Healther Merrifield, Kate Keating, Wendy Olmstead, Todd Pierce, Jamie Cowgill & Jim Bolek

ACKNOWLEDGEMENTS

With funding from the National Science Foundation (NSF) *Widening Implementation and Demonstration of Evidence-based Reforms (WIDER) Program*, AAAS was able to bring together some of the best minds to offer advice on measuring faculty teaching practices in undergraduate science, technology, engineering, and mathematics (STEM) education. We are grateful to NSF for providing this opportunity for STEM education experts to engage in discussions about the strengths and weaknesses of various approaches to measuring undergraduate teaching practices, including faculty and student surveys, faculty interviews, classroom observations, and portfolios and other artifacts.

This report represents the collective wisdom of 60 faculty members, evaluators, researchers, and administrators who participated in workshop planning activities, writing pre-conference papers, facilitating and writing reports for discussion groups, reviewing draft reports, and/or editing pre-print proofs. We are especially grateful for the leadership provided by the Chair of this workshop, Charles Henderson, Associate Professor of Physics and education researcher at the Mallinson Institute of Science Education, Western Michigan University. We also thank Maura Borrego, a Program Director at the NSF Division of Undergraduate Education (DUE), for her guidance.

We particularly appreciate the contributions of the White Paper Authors (listed on page ii) and Hilda Barko, workshop plenary speaker, for their provocative insights that helped to guide the work of the discussion groups. In addition, we appreciate the service of the facilitators (listed on page ii), who worked long into the night and after the workshop, capturing and documenting the ideas generated in their energetic working groups.

The report would not have been possible without the meticulous attention of the writer and editor, Diane Smith. Compiling the planning documents, papers, and working group reports into a coherent document and then writing and rewriting the report based on input from the leaders and participants in the workshop took months of work.

We would like to thank Brian Baker, the copy editor, who not only edited the report but also double checked the exhaustive list of works cited. Finally we would like to express our gratitude to Gail Peck, Peck Studios, Inc. (graphic design), Janel Kiley, AAAS publication staff, and AAAS proof readers (Betty Calinger, Brooke Clayton, Gerard Boulin and Derry Earle).

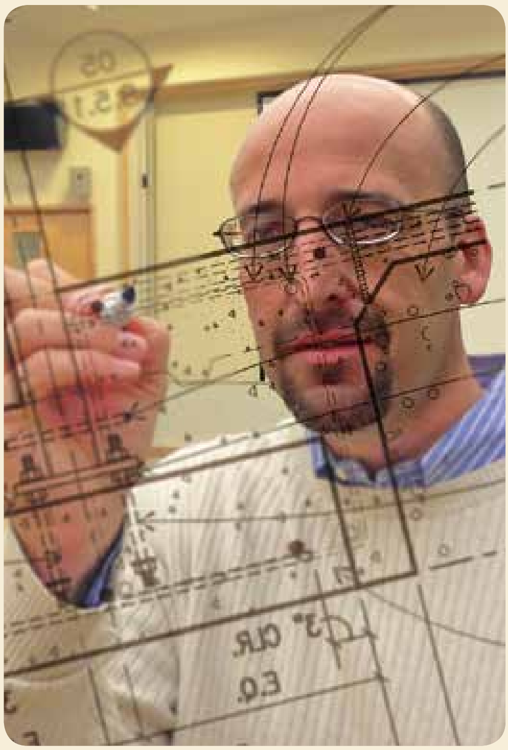
We sincerely hope that this report will provide STEM faculty, researchers, and evaluators with new ideas about how to describe and measure undergraduate STEM teaching practices. Better understanding of "what works" in undergraduate teaching practices is critical to increasing the retention and enhancing performance of STEM undergraduate students, indeed of all students, and thereby building the next generation of STEM professionals and science literate citizens.

Best Regards,

Yolanda S. George Shirley M. Malcom

Yolanda S. George and Shirley M. Malcom
AAAS Education and Human Resources Programs

DESCRIBING & MEASURING UNDERGRADUATE STEM TEACHING PRACTICES



INTRODUCTION

In recent years, there has been a growing emphasis on the importance of improving undergraduate education in science, technology, engineering, and mathematics (STEM) disciplines. Being able to describe teaching practices in undergraduate STEM learning environments is an important foundation for many of these improvement initiatives. Yet, there has been little systematic work to identify tools and techniques that can be used in such descriptions.

The improvement of STEM teaching is critical to the nation's future. High attrition rates of students from STEM majors—particularly students from underrepresented groups—a growing demand for STEM professionals, and the national need for a strong science background for all students make it urgent that the problem be addressed. This sense of urgency has been recognized by organizations such as the National Science Foundation (NSF), which has supported systemic efforts to improve teaching and learning in undergraduate STEM education. But even with widespread national investments, education researchers, administrators, and faculty do not yet have shared and accepted ways to describe and measure important aspects of teaching. Developing the language and tools necessary to describe teaching practices in undergraduate education is crucial to achieving productive discussions about improving those practices.

On December 17, 2012, the American Association for the Advancement of Science (AAAS), with support from NSF, convened a three-day meeting of leading experts from around the country to explore what is known about describing and measuring undergraduate STEM teaching. Participants included content experts drawn from STEM disciplines, higher education researchers, and faculty development specialists. The meeting highlighted approaches and perspectives that will allow the research community, administrators, and faculty to document teaching practices across STEM disciplines and to identify gaps that require additional research.

To help facilitate discussions, prior to the meeting the organizers commissioned five white papers, each focusing on a different measurement technique: faculty interviews, faculty portfolios, faculty surveys, student surveys or interviews, and observations in educational settings. The authors of these papers surveyed the relevant literature and identified key issues and ideas to ensure that meeting participants would base their discussions on current research. (Citations of the literature are included at the end of

this report.) Participants used the white papers as a basis for their three days of discussions.

Based on the white papers and follow-up discussions, this report serves as a resource for faculty, administrators, policymakers, researchers, and evaluators who wish to, or are being asked to, describe teaching for the purpose of improving teaching, document appropriate teaching, or do research on teaching. The report identifies four basic measurement techniques (surveys, interviews, observations, and portfolios), provides an overview of the strengths and weaknesses of each, identifies and summarizes specific protocols and measurement tools within each technique, and gives references for further details. An important conclusion is that the best descriptions of STEM teaching involve the use of multiple techniques.

Measuring STEM Teaching Practices

During the three-day meeting, the participating experts discussed and synthesized what is currently known about the measurement and documentation of STEM teaching practices. Because teaching practices have been measured more extensively in K–12 education, the organizers of the meeting invited Hilda Borko, an educational researcher at Stanford University, to give a plenary presentation on "Measuring Teaching Practice" from the perspective of K–12 education. Borko also attended breakout sessions throughout the meeting's second day, providing an additional source of expertise as participants discussed using effective measurement techniques in undergraduate classrooms.

Although not all meeting participants agreed on all points, they did concur that, to be successful, measuring and describing STEM teaching practices requires expertise, planning, and resources. Moreover, without the ability to measure teaching practice, the desired advances in STEM teaching and learning are difficult to substantiate, document, and promote. Participants also agreed that all stakeholders, from STEM faculty, administrators, and evaluators to funding agencies, professional societies, and policymakers, require a better understanding of how to document undergraduate teaching practices. Illuminating what is currently known about this critical aspect of improving STEM teaching and learning was the primary reason for convening the national meeting of experts.

Moving from Teaching to Learning

As many noted during the meeting, it is difficult to discuss teaching practices without discussing student learning. By focusing on describing teaching practices, this report does not mean to imply that student learning and its measurement are not important. Indeed, documenting both teaching practices and student learning is necessary in a well-functioning educational system. However, although much information is available that summarizes techniques for measuring student learning (see, e.g., the 2012 National Research Council report on discipline-based education research), such overviews do not exist for the documentation of teaching practices.

Meeting participants sought to encourage their colleagues around the country to develop, adopt, or adapt, in their own undergraduate classrooms and programs, the techniques and instruments currently available to describe STEM teaching. With more widespread measurements of STEM teaching, meeting participants hope to foster the collection and sharing of data and descriptive information across a wide spectrum of disciplines and institutional settings—an important first step in creating a common language to talk about STEM teaching. Further, careful descriptions of teaching can then be related to student outcomes in order to identify effective, evidence-based teaching practices. After researchers and STEM faculty all speak the same language, they can focus on identifying and improving student learning and effecting real change.

Resources Available

The sections that follow provide an overview of four techniques that individuals or organizations can use to measure STEM teaching: faculty and student surveys, interviews, classroom observations, and teaching portfolios. As with any type of research, before choosing a measurement technique or techniques, it is important to clarify the purpose and ultimate goal of the measurement because each technique has its own strengths and weaknesses.

In designing a measurement study, it is also important to assess the resources that are available to meet the study's goals, since most techniques for describing instructional practices require considerable time and expertise to implement. However, trade-offs can be made. For example, researchers might want to develop a measurement tool specifically suited to a particular purpose and measurement perspective. But they could find that developing a new instrument will be significantly more expensive than using an existing instrument, even though the existing instrument is not ideal for the situation. In such a case, the researchers have to make their decision in part on the resources available. Thinking about describing teaching practices sometimes entails looking to experts for assistance. Campus-based educational researchers or evaluators may possess the expertise necessary for implementing many of the techniques described in this document.

MAKING DECISIONS ABOUT MEASUREMENT

Faculty, administrators, faculty developers, and policymakers usually have at least one of three goals for the measurement of teaching practices: documenting classroom practice, improving teaching, and conducting research and evaluation. Although these goals are not mutually exclusive, care must be taken when more than one of them is identified. For example, providing confidential feedback to individual faculty members for the purpose of teaching improvement could conflict with the goal of providing data for use in tenure and promotion decisions.

Documenting Practice: The ability to document effective teaching is becoming increasingly important for higher education institutions. Faculty are called upon to document their teaching practices for the purposes of getting and retaining their jobs. Also, institutions are more frequently being asked to justify their teaching practices to a wide variety of stakeholders, including state legislatures and potential students.

Improving Teaching: Documenting an existing practice is often the first step in improving it. Documentation can be initiated by an individual teacher for the purpose of self-improvement or by a department or institution to better understand teaching practices at the institution. For example, many centers for teaching and learning document measurements of teaching practices as part of an individual or department-level needs assessment. Measurement for the purpose of improving teaching can also be initiated externally through reform initiatives or by professional societies.

Conducting Research: STEM education researchers often make claims about the impact of certain teaching practices on student learning. However, without clear articulation and documentation of teaching practices (and robust measures of student learning), making a credible claim is nearly impossible. Similarly, a researcher may notice that different instructors who claim to be using the same general teaching methods have significantly different student-learning outcomes, but without careful measurement of teaching practices, important differences in those practices that influence student learning could go unidentified.

An Important Note About Terminology

The term "measurement of teaching" is used throughout this report to refer to documenting teaching practices as accurately as possible. As noted in the accompanying sidebars below and on page 4, measurements may be used for either descriptive or evaluative purposes.

The term "teaching" is used to refer to activities done by people in the professional role of teacher with the intention of bringing about student learning. As the meeting participants emphasized, teaching is much broader than what happens in classrooms, lecture halls, laboratories, or the field: It includes the preparation that instructors invest in designing their courses, as well as informal interactions with students during office hours or in other one-on-one or small group consultations. Indeed, as many noted, some of the most important aspects of STEM teaching occur outside the four walls of the traditional classroom.

Measurement Perspective: Descriptive or Evaluative

Two basic perspectives can be taken on the measurement of teaching practices: descriptive and evaluative. It is important for the person or organization engaged in measuring teaching practices to clearly articulate the measurement perspective.

A descriptive approach documents teaching practices with as little value judgment as possible. Of course, any description has an implicit value judgment based on what the researcher chooses to describe, but this judgmental aspect can be minimized through the development of protocols. Some observations and surveys, such as the *Higher Education Research Institute* (HERI) survey discussed later in this report, use a descriptive approach to document existing practices. Alternatively, descriptive measurements might be used at the beginning of a change initiative to identify the starting point for future work.

In an evaluative approach, the goal is to compare the teaching against some model of good teaching. A department may wish to focus on encouraging faculty to use higher order questions on tests (as measured, e.g., by Bloom's taxonomy). Tests from each instructor could then be collected and each question rated according to the type of thinking skill required (e.g., remembering, applying, evaluating). Or the developer of an instructional reform might want to document the extent to which secondary adopters are using the essential features of the reform. In this case, the effective use of each feature would be seen as ideal. The results of such a measurement are also sometimes combined into a cumulative single-rating scale to compare the "quality" of the implementation across participants.

Also important is the fact that not all undergraduate STEM "instructors" are traditional tenure-track faculty. Many are adjunct faculty members or even graduate students. Unless otherwise specified, the terms "faculty" and "instructor" refer to those whose undergraduate teaching practice is being documented, regardless of their individual tenure or hiring status.

The report sometimes refers to those using the various measurement techniques as "researchers," "evaluators," or "investigators." These terms are used throughout the report because each of the techniques can be adapted for the purposes of research, evaluation, or faculty development. For example, a researcher or an investigator could be a faculty member, an administrator, a faculty developer, a consultant, or an education researcher who uses the surveys, interviews, portfolios, or observation protocols presented to measure undergraduate STEM teaching practices. (The different purposes for which the data obtained might be used, and specific considerations pertaining to uses of the data, are discussed later.)

Finally, the term "undergraduates" covers students at community colleges and four-year institutions who are majoring in any discipline. All students, regardless of their academic majors, career goals, or economic, ethnic, or social background, can benefit from improved teaching in undergraduate STEM courses. To this end, faculty, department heads, administrators, and others need a better understanding of how to identify, measure, and, with the resulting data, ultimately promote best teaching practices in both general education and major-specific courses.

WHERE TO START

A good starting point for those interested in measuring teaching practices, especially with the goal of improving teaching or documenting effective teaching, is Nancy Chism's *Peer Review of Teaching: A Sourcebook*. This accessible book describes philosophies behind the measurement of teaching practices, offers practical advice for implementing measurement procedures, and provides a wide variety of measurement criteria and sample measurement forms. The emphasis of the book is on faculty portfolios and faculty observations.

(See Chism, N. (2007). *Peer Review of Teaching: A Sourcebook* (2nd ed.). Bolton, MA: Anker.)

SURVEYING STEM FACULTY AND STUDENTS

Surveys can be one of the most efficient ways to obtain information about instructional practice, from the perspectives of both STEM faculty and their students. Surveys provide a scalable and relatively efficient approach for studying STEM teaching, allowing for responses from a widely dispersed population. Researchers can gather data to describe current teaching environments, general attitudes toward teaching, faculty and student demographics, use of the Internet and other technologies for teaching, and instructional behaviors and beliefs. They can also gather baseline data to document change over time. If the same survey instrument is used repeatedly with the same sample population, analysis can reveal trends or shifts in responses. Moreover, with large and representative samples, an investigator can generalize the results to a larger population.

Although most items on typical surveys force the respondent to make a choice, surveys can also include open-ended items that provide valuable information about the perspective of the respondent. Offering an opportunity to respond to more open-ended questions can be important because many commonly used words in education, such as “critical thinking” or “active learning,” are interpreted differently by different people. Asking respondents to write a short sentence in addition to, or even instead of, selecting checkboxes can be a useful approach at strategic locations in the survey. Of course, the addition of open-ended items increases the difficulty of analysis.

Those interested in using or developing a survey to measure STEM teaching practices should keep in mind the following strengths and challenges inherent in that technique:

Survey Strengths

- A survey represents a time-efficient and convenient method for collecting quantifiable data.
- A survey can reach large numbers of students or instructors.
- When both students and instructors are surveyed, the responses can provide data to compare student perceptions with instructor intent.
- Representative sampling can allow inferences to be made about a larger population.
- Survey software streamlines data collection and preparation, and data collected can be exported directly into spreadsheets for statistical analysis.

Survey Challenges

- Few validated STEM instructor survey instruments are currently available; thus, survey development may be necessary.
- Survey development requires careful consideration and pilot testing to create unambiguous questions and response options that are interpreted in a consistent manner by those who take the survey.
- Response rates can be low, in part because of “survey fatigue.”
- Self-reported data represent respondents’ perceptions, which may not align with other measurements.
- Student responses to surveys about instructional practices are grounded in their previous experiences with instruction, which need to be taken into account.

IMPLEMENTING SUCCESSFUL SURVEYS

As with any data-gathering instrument, investigators should define their goals clearly before designing a survey. Also, they should conduct a pilot study to ensure that survey questions are unambiguous and address the measurement goals. In particular, careful thought must be given to the construction and analysis of questions that require respondents to recall events in the past and to indicate their frequency. In addition, the use of words must be carefully considered, because even seemingly simple terminology may be unfamiliar or misunderstood. Clearly, the more care and analysis invested before widespread distribution of a survey, the more likely it is that the survey will deliver quality results.

Self-reported data can be biased. One of the more common biases arises when respondents answer questions with what they believe will be viewed most favorably rather than with the response that most accurately reflects their own beliefs or

practices. This kind of bias tends to occur when respondents face consequences—such as a demotion or promotion in their job—for their responses or when the behaviors probed are stigmatizing, embarrassing, or socially undesirable. For example, STEM faculty commonly report the use of active-learning techniques in their classrooms, yet follow-up questions often fail to document behaviors consistent with that approach to teaching. To minimize the impact of bias related to social undesirability, survey designers have developed techniques such as asking about important issues from multiple perspectives and in different parts of the survey.

Although technology has simplified the administration of online surveys, it also has led to a surplus of surveys and a corresponding decrease in response rates. To increase response rates, it helps to provide survey participants with a clear understanding of what they might gain from participating, of whether or not they will have access to the results of the survey, and of any altruistic or intrinsic value there may be to participating. Advance notice, endorsements by respected authorities, incentives, and a guarantee that responses will be anonymous also may improve response rates. Starting with a set of questions that respondents will most likely find motivating can be an effective strategy for keeping respondents engaged.

Getting responses from students can be particularly difficult at the end of the semester, because they are busy preparing for exams. If appropriate, incentives (e.g., raffle gift cards, extra credit) can be particularly useful in improving the response rate among students. Also, when possible, have students complete the survey during class time.

Online survey tools often allow respondents to store contact information in an address book. Such a book can facilitate custom invitations written to individuals assigned to different groups (e.g., departments, schools, the entire faculty) stored in the address book. Personal introductions (e.g., “Dear Prof. lastname”) can be created from variables pulled from address book fields. Finally, online survey tools can track those who have responded to the survey, permitting customized reminder e-mails to be sent selectively to nonrespondents.

The sections that follow describe many existing surveys. When adopting or adapting entire surveys or specific questions for use with a new population, it is important to verify that the respondents are interpreting and responding to the questions as expected. Many factors (such as the type of institution, the academic field surveyed, and the location where the survey is conducted) can influence how respondents answer specific questions.

COORDINATED SURVEYS OF FACULTY AND STUDENTS

Although faculty and students are often surveyed separately, significant strength may be achieved by surveying them together. One example of this approach is the Engineering Change (EC2000) study, in which surveys were carefully designed to describe several aspects of educational practices in U.S. engineering programs. The survey development process was an iterative, yearlong undertaking based on many interviews with engineering faculty, students, and program chairs on multiple campuses. In addition, formal pilot testing of the survey instruments was carried out with engineering faculty, administrators, students, and members of the study's national advisory board in order to refine survey items, eliminate faulty questions, and produce a set of multiquestion scales (i.e., scales in which different wording is used to measure the same outcome; single-question scales are less reliable). For example, faculty and department chairs were asked about changes in teaching methods, such as computer simulations, case studies, open-ended problems, design projects, and the use of groups in class and lectures, that took place since implementation of the EC2000 criteria. Similarly, students were asked about how often they experienced these teaching methods in their engineering courses. Comparisons of graduating seniors' reports with those of alumni from the same programs 10 years after graduation allowed researchers to conclude that current students perceived more active engagement in their own learning, more interaction with instructors, and more instructor feedback on their work. An important strength of this study is its use of a conceptual framework that logically connects potential EC2000-related program changes reported by faculty and administrators with both student reports of their experiences and outcome measures.

See Lattuca, L. R., Terenzini, P. T., & Volkwein, J. F. (2006). *Engineering Change: Findings from a Study of the Impact of EC2000, Final Report. Baltimore: ABET.* http://www.abet.org/uploadedFiles/Publications/Special_Reports/EngineeringChange-executive-summary.pdf. Surveys are available at http://hdl.handle.net/2027.42/97373.

FACULTY SURVEY INSTRUMENTS

A number of national STEM faculty surveys have been, and continue to be, conducted, typically as part of research studies. Each study reveals new information about how to survey faculty members about their teaching; however, none of the surveys has necessarily used or resulted in validated instruments. The studies cited next have in many cases been informed by each other, so they can be compared. The surveys are organized on the basis of the types of information about teaching practices that they were designed to collect.

Faculty Use of Specific Teaching Strategies

A number of studies in physics and engineering education have focused on the propagation of "named" research-based instructional approaches, inquiring into methods such as "peer instruction" and "service learning." As some of these studies have demonstrated, a weakness of focusing on named instructional approaches is that the names can be misinterpreted by respondents. Borrego, Froyd, and Hall (2010) asked engineering department chairs about pedagogies, such as service-learning projects, learning communities, and interdisciplinary capstone design projects. For example, one question was "approximately when did you first hear about [the pedagogy X]?" and one response option was "this is the first I have heard of it." The exact wording of other items is listed in the article.

Henderson and Dancy (2009) asked questions with similar response options: "I currently use all or part of it" (current user); "I have used all or part of it in the past" (former user); "I am familiar with it, but have never used it" (knowledgeable nonuser); "I've heard the name, but do not know much else about it" (little knowledge); and "I have never heard of it" (no knowledge). The 24 specific instructional strategies included Just-in-Time Teaching, Modeling Physics, Peer Instruction, Physlets, and Workshop Physics. The entire list is included in their publications (Henderson & Dancy, 2009; Henderson, Dancy, & Niewiadomska-Bugaj, 2012).

Later, Froyd, Borrego, Cutler, Henderson, and Prince (2013) adapted these questions to apply to faculty members in chemical, electrical, and computer engineering. On the basis of feedback from the physics study, they revised the response options to "I currently use it"; "I have used it in the past"; "I have used something like it but did not know [the] name"; "I am familiar with it but have never used it"; "I have heard [the] name but know little else about it"; and "I have never heard of it." They also adapted the list of instructional strategies on the basis of evidence from the literature that the strategies were being used in engineering. For example, Just-in-Time Teaching and Peer Instruction were retained from the physics list.

Focus on General Teaching Activities

Some surveys have described teaching in terms of more general teaching activities, instead of focusing on named teaching strategies. This approach helps reduce the chance that respondents will misinterpret the questions, but it can also make it more difficult for respondents to complete the survey and for the researcher to interpret the results. MacDonald, Manduca, Mogk, & Tewksbury (2005) asked geoscience faculty members about course activities (e.g., traditional lecture, lecture with demonstration, small-group discussion, fieldwork) and problem-solving activities (e.g., reading primary literature, working on online problem sets, engaging in structured collaborations). Their response options were “never,” “once or twice,” “several times,” “weekly,” and “for nearly every class.” Dancy and Henderson (2010) adopted a similar scale, asking about the use of conceptual questions, small-group discussions, and traditional lectures.

There are several variations on how to quantify the extent of use of these types of activities. The preceding examples emphasized frequency in terms of class periods. In that regard, Borrego et al. (2010) asked respondents to “indicate what percentage of time on average your students spent/spend on each of the activities below during class time.” The options were 0%, 1–25%, 26–50%, 51–75%, and 76–100%. For all activities except lecture, the majority of responses were in the 1–25% range. The Faculty Survey of Student Engagement (FSSE, 2012) includes a question specific to a course taught recently by the respondent: “In your selected course section, what percent of class time is spent on the following?” The options are 0%, 1–9%, 10–19%, 20–29%, 30–39%, 40–49%, 50–74%, and 75% or more. Ten activities include lectures, teacher-led discussions, small-group activities, student presentations, and experiential activities (laboratory experiments, fieldwork, art exhibits, etc.). This survey uniquely compares faculty and student responses related to a variety of in-class and out-of-class activities and includes many more items, about student workloads and levels of challenge at the institution.

The Higher Education Research Institute’s (HERI) faculty survey asks, “in how many of the courses that you teach do you use each of the following?” Response options are “all,” “most,” “some,” and “none.” The instructional techniques or methods asked about include class discussions, experiential learning/field studies, extensive lecturing, and electronic quizzes with immediate feedback in class (Hurtado, Eagan, Pryor, Whang, & Tran, 2012; the survey may be found at http://www.heri.ucla.edu). Borrego et al. (2010) asked department chairs to answer similar questions on behalf of all their faculty members and added items to estimate the percentage of faculty and engineering majors involved in the activity.

To better understand the relationships between specific activities and specific research-based instructional strategies, Borrego, Cutler, Prince, Henderson, & Froyd (2013) compared engineering faculty responses to questions about what students do in class with specific research-based instructional strategies currently being used. For example, among faculty who say they use "think–pair–share" in their engineering science courses, 94% have students "discuss a problem in pairs of groups" and 64% have students "report their group's findings to the entire class (formally or informally)." The authors examine how well 16 student activities reported by faculty map to 11 research-based instructional strategies described in the literature.

Finally, some surveys of faculty have emphasized assessment techniques and grading tools. A survey of geoscience faculty inquired about their use of a set of assessment techniques, including exams, quizzes, problem sets, rubrics, and concept maps (MacDonald et al., 2005). Similarly, the National Study of Postsecondary Faculty (NSOPF; the survey may be found at nces.ed.gov/surveys/nsopf) asked, "for the undergraduate classes you taught for credit during the 2003 Fall Term at [institution], did you use any of the following?" Options focused primarily on assessment methods (various types of midterm and other exams, "group and team projects producing a joint product," "student evaluations of each other's work," and "laboratory, shop, or studio assignments"). Response options were "used in all classes," "used in some classes," and "not used."

Learning Goals

Some surveys also ask about the learning goals that faculty have for their students. One of the questions on the 2010–2011 HERI faculty survey (Hurtado et al., 2012) asked respondents to indicate the importance to them of a number of education goals for undergraduate students. The response options were "essential," "very important," "somewhat important," and "not important." Education goals included "develop ability to think critically"; "prepare students for employment after college"; "prepare students for graduate or advanced education"; "develop moral character"; and "promote ability to write effectively." In another question on the same survey, faculty members were asked how often in their interactions with undergraduates did they encourage them to participate in each of the following activities: “support their opinions with a logical argument,” “seek alternative solutions to a problem,” and “look up scientific research articles and resources.” The response options were “frequently,” “occasionally,” and “not at all.”

Marbach-Ad et al. (2012) asked chemistry and biology instructors about a similar set of instructional goals. They used a five-point scale ranging from “not important” to “very important.” Instructional goals were phrased as skills and included “scientific writing”; “memorize some basic facts”; and “remember formulas, structures, and procedures.” The complete survey instrument can be found at http://cmns-tlc.umd.edu/tlcmeasurementtools.

Beliefs, Attitudes, and Values about Teaching and Learning

Trigwell and Prosser (2004) created a 16-item Approaches to Teaching Inventory based on a typology of teaching intention and teaching strategy resulting from interviews with STEM faculty members. Items include “I feel that the assessment in this subject should be an opportunity for students to reveal their changed conceptual understanding of the subject” and “I think an important reason for running teaching sessions in this subject is to give students a good set of notes.” Each item is rated by the instructor on a five-point Likert scale from “rarely” to “almost always.” The authors report that instructors’ approaches to teaching are correlated with their students’ approaches to learning. For example, when teachers adopted student-focused approaches to teaching, their students adopted a deeper approach to learning. By contrast, teacher-centered approaches were correlated with students taking a surface approach to studying (as measured on a similarly constructed survey of students’ approaches to learning).

STEM surveys often ask faculty members about barriers that might prevent them from using specific instructional strategies. Dancy and Henderson (2010) categorized physics faculty members’ open-ended responses into “time,” “lack of knowledge,” “weaknesses of the method,” and “lack of compatibility (personal or organizational).” Similarly, Borrego et al. (2010) coded department chairs’ responses into “resources (funding, technology, space or staff),” “student reactions,” and “faculty motivation, time, and reward systems.” These were later codified into multiple-choice options on engineering faculty surveys: “takes up too much class time to let me cover the syllabus”; “too much advanced preparation time required”; “lack of evidence to support the efficacy of this instructional strategy”; “students would not react positively”; “my department does not have the resources to support implementation”; and “my department and administration would not value it” (Froyd et al., 2013).

Faculty, Student, and Institutional Characteristics

Most of these surveys also ask for information about the settings in which faculty are working, in order to ensure representative results and understand any differences that might exist. Common items include the type of institution; faculty rank and tenure/part-time status; responsibilities in teaching and research; publication record as a measure of research activity; participation in faculty development related to teaching (distinguishing on-campus workshops from external offerings); gender; and years of teaching experience. None of the surveys except HERI (Hurtado et al., 2012) ask for race and ethnicity information, perhaps because the pool of STEM faculty is too small to preserve the anonymity of respondents. However, national studies have shown that there are systematic differences in the use of pedagogical strategies based on gender and ethnicity, so, in most situations, it makes sense to collect this information. If anonymity concerns arise, a decision could be made not to report any race or gender information collected.

Faculty surveys about instructional practices are most meaningful when respondents are asked to focus their answers on a specific course; information collected may include class size, level or year of the students, whether the student is enrolled as a major in the course, whether the course is required or elective, and additional discipline-specific characteristics. An extensive list of items is found in Henderson et al. (2012), and discipline-specific items are listed in MacDonald et al. (2005) and Froyd et al. (2013). These sources describe both the variables and the response options. Additional examples can be found on the FSSE, HERI, and NSOPF survey forms. These higher education (though not STEM-specific) surveys also include questions about online course resources; teaching facilities; teaching assistants; and the percentage of time faculty members spend on activities such as teaching, research, and service.

STUDENT SURVEY INSTRUMENTS

The most ubiquitous survey in higher education is the survey distributed to students at the end of a class, asking them to rate their instructor and the instruction they received. Although each institution tends to have its own instrument, one readily accessible example that some workshop participants felt was of good quality is the Virginia Tech Student Perception of Teacher (SPOT) Survey, online at http://www.undergraduate.vt.edu/faculty/ExampleSPOTquestionnaire.pdf.

End-of-course surveys generate a summative judgment of an instructor's teaching in a given course. There is a significant body of research about the validity and reliability of students' evaluations and how various factors (e.g., the student's expected grade, the gender of the instructor, class size, whether the course is required or elective) correlate with the resulting ratings. A review of this large body of literature is beyond the scope of this report, and many excellent reviews already exist (e.g., Berk, 2005; Marsh, 2007; Wachtel, 1998).

Student data from surveys are one source of information about teaching practices that can be especially strong when combined with information from other sources (e.g., course-specific questionnaires, peer observation of teaching, instructor self-assessment). Following is an overview of other types of representative student survey instruments available for use or adaptation. These instruments provide a starting point for researchers interested in using student surveys.

Teaching Behaviors Inventory (TBI)

Used in many college classrooms, the TBI is based on an observation protocol that has been adapted into a student survey (Murray, 1987). Focusing on teaching behaviors, such as instructor clarity, enthusiasm, organization, and rapport, the inventory is designed to capture objective reporting of behaviors rather than student judgments about teaching effectiveness. The 60 items are divided into eight categories, three of which are clarity ("points out practical applications of concepts"), organization ("explains how each topic fits into the course as a whole"), and disclosure ("tells students exactly what is expected of them on tests, essays or assignments"). The instrument is found at http://www.calvin.edu/admin/provost/documents/behaviors.pdf.

Student Assessment of their Learning Gains (SALG)

SALG asks students to report the extent to which they believe that specific aspects of a course have helped them learn. An example of an instrument that can be used at the course level and for formative feedback to individual instructors, SALG has also been used for program evaluation and research. A “wizard” allows users to create their own surveys from a bank of questions that can be customized. One of the basic questions is “How much did the following aspects of the course help you in your learning?” Customizable response options might include class activities, graded assignments, resources used, etc. SALG is found at http://www.salgsite.org.

National Survey of Student Engagement (NSSE) and Wabash National Study of Liberal Arts Education: Student Experiences Survey

The NSSE and the Wabash study are particularly well-known national student surveys, but their focus on the impact of college, writ large, may make them less useful for studying particular STEM courses. Prompts include “during the current school year, about how often have you done the following?” “During the current school year, to what extent have your instructors done the following?” and “During the current school year, how much has your coursework emphasized the following?” Some of these questions could be useful if adapted to a particular setting (course, instructor, or major department). Response options might be “faculty reviewed and summarized the material effectively”; “faculty gave assignments that helped in learning the course material”; “the presentation of material was well organized”; “faculty were well prepared for class”; “faculty interpreted abstract ideas and theories clearly”; “class time was used effectively”; and “course goals and requirements were explained clearly.” Information about the NSSE is found at http://nsse.iub.edu; the Wabash survey is at http://www.liberalarts.wabash.edu/storage/assessment-instruments/Student_Experiences_Survey.pdf.

Motivated Strategies for Learning Questionnaire (MSLQ)

The MSLQ probes cognitive and affective dimensions of learning and can be used to inform teaching decisions. The instrument examines several aspects of motivation related to learning, such as goal orientation and self-efficacy, and has been used at the college level. Among the response options are "compared with other students in this class I expect to do well" and "when I am studying a topic, I try to make everything fit together." The MLSQ is found at http://www.indiana.edu/~p540alex/MSLQ.pdf.

BOTTOM LINE

Although designing and implementing an effective survey to document undergraduate STEM teaching requires expertise and can be a daunting experience, a well-designed survey can support a deeper understanding of teaching practice. Collaborators with expertise in the design and analysis of surveys can be found in many higher education institutions through institutional research offices, offices of assessment or evaluation, and teaching and learning centers, as well as in education, sociology, and other social science departments. Results of surveys taken of teaching practice from both faculty and student perspectives also can help STEM instructors view their own teaching in a larger context.

INTERVIEWING STEM FACULTY AND STUDENTS

While surveys can reach a large number of STEM faculty and students, interviews provide an opportunity to explore teaching practices in more depth. Interviews can be conducted one-on-one or with a small focus group. Using an open-ended format, an interviewer can ask faculty or students directly about their recollections and perceptions of STEM teaching, as well as respond to interviewees with follow-up questions in real time. This approach allows for deeper exploration and the emergence of the unexpected. Alternatively, interviews can be highly scripted, with interviewers asking every respondent the same questions in the same order.

Interviews create a constructive opportunity for a researcher to interact with an individual and gather data about specific STEM teaching practices. Interviews can explore everything from teaching activities and beliefs to motivations and perceptions, and can be used to identify common barriers to, or misconceptions about, STEM teaching and other complex or poorly understood topics.

In addition, interviews allow the interviewer to explore causal mechanisms in ways that are difficult to accomplish with other research methods. By asking open-ended questions, an interviewer can discover what is salient to the respondent. Interviews can be used as pilots or case studies to help researchers develop quantitative instruments and observational protocols or to compare descriptions of teaching with other forms of data collection. Interviewing student focus groups is particularly useful when investigators would like students to interact with one another regarding a specific topic.

In deciding to use or develop interviews to measure teaching practices, the following strengths and challenges inherent in the technique should be kept in mind:

Interview Strengths

- Interviews allow for the identification and investigation of important new areas and topics that may not have been considered a priori by the interviewer.
- Interview protocols can be designed to collect both quantitative and qualitative data, and interviewers can pursue lines of questioning that would be difficult to ask in a survey.
- Interview data can help illuminate not only actions and beliefs, but also the reasons behind the actions and beliefs.
- The open-ended nature of the questions, combined with the ability of the interviewer to adapt as the interview progresses, allows for issues to emerge; this flexibility also allows investigators to follow up on interesting issues as they emerge.

Interview Challenges

- Interviews are time and labor intensive in all aspects of design, delivery, and data analysis. This consideration limits the number of respondents, sometimes leading to concerns about the validity of the results.
- Investigators require training or experience in order to collect, analyze, and interpret data and to report results.
- Self-reported data collected through interviews represent the perceptions of the respondents, and those perceptions may not align with measurements and observations garnered from other instruments.
- Because interviews may not be viewed as a rigorous methodology by some STEM personnel, researchers must be ready to justify their choice of method and explain its advantages.

IMPLEMENTING SUCCESSFUL INTERVIEWS

Putting together a qualified team is essential to defining the goals of the interview, designing the questions to be asked, and then analyzing the results. It is also important to consider who will conduct the interviews, because mismatches of status or power between the interviewer and interviewee may bias the results or make the interviewee feel at risk (e.g., a faculty member who is up for promotion or tenure). Interviewers must also be able to maintain strict confidentiality and present themselves as nonjudgmental throughout the interview.

As with all self-reported data, interviews measure personal perceptions, beliefs, and memories, with some interviewees forgetting or misremembering events. Thus, interviewers should ask participants to report particular behaviors in detail and to explain fully any viewpoints elicited. Follow-up and probing questions can be used to ensure rich, descriptive data that provide strong evidence for particular behaviors, attitudes, dispositions, and experiences. For example, many STEM faculty report teaching interactive classes. Follow-up questions to determine whether that is an accurate characterization might include "how often do students talk in class?" "To whom do the students talk?" and "what do the students talk about?" The responses can help researchers analyze the degree of actual interaction in a classroom more accurately.

The structure of interviews can vary with the ultimate goal of the interviews and the resources available. Interview protocols that are more structured are easier to implement and analyze, but less structured interviews with open-ended questions and opportunities to diverge from the strict line of questioning often provide deeper and richer data. One option is to structure an interview that uses both approaches, developing a set of questions all interviewees will respond to but also giving the interviewer the option to diverge as needed. Providing optional follow-up questions can help guide a discussion into other areas, even if the interviewee does not bring them up. Such questions also furnish the interviewer with follow-up ideas should an interviewee not be very talkative or forthcoming.

Interview Development and Data Analysis

Interviews are a qualitative data collection tool; thus, the most useful interviews seeking information about measurements of teaching practices need to follow rigorous qualitative research methods. Accordingly, careful planning is required to properly align the research or evaluation questions, the data collection tool, the selection of research participants, and the methods for data analysis. Creswell (2012) is a good basic reference for those interested in designing strong qualitative research studies.

Resources ultimately determine the number of interviews that can be conducted; however, the goal of the data collection and how the results will be used can also help shape a research or evaluation design. In addition, the goal influences the criteria for selecting participants and whether individual or small-group interviews are more appropriate. For example, if interviews are meant to explore ideas for future research, a smaller sample often provides valuable insights. By contrast, generalizing results to a larger population typically requires a larger sample. In some qualitative research traditions, the number of interviews is not identified in advance, but rather, new interviews are conducted until they fail to yield new ideas.

Although transcribing interviews is time intensive and can be expensive, the most difficult aspect of interviewing is analyzing the data. When analyzing interviews, researchers must be able to recognize nuances yet at the same time not overstep their interpretation. Different people, even those with similar backgrounds, will likely see different things in the interview data, with researchers disagreeing about what an interviewee meant or missing a particularly important point. Extensive discussion about the interpretation of data is a typical feature of strong research studies involving interviews. In addition, the effective interpretation of interview data requires a fairly sophisticated understanding of the research topic. When the necessary expertise is not located in a single researcher and when studies explore unfamiliar and complex topics, multiple researchers may be beneficial, as they bring different perspectives to the task of interview design, data analysis, and the interpretation of findings.

COORDINATED INTERVIEWS OF FACULTY AND STUDENTS

As a form of teaching that takes place outside the classroom, undergraduate research leads to significant student learning through the collaborative pursuit of mutual scholarly interests with faculty. A large interview study of apprentice-model undergraduate research points to the importance of comparing multiple data sources in order to understand both the student-learning outcomes and the teaching processes involved in such research. Interviews with research students were compared with interviews with students who did not participate in research and with interviews of faculty who acted as research advisors to the student researchers. The overall similarity of student and faculty observations corroborates many important aspects of the undergraduate research experience, while differences in particular aspects reflect meaningful differences in faculty and student perspectives. Comparisons of gains reported by participating students with those reported by nonparticipating students highlight areas in which research experiences offer tangible gains relative to other college learning experiences. Faculty spoke explicitly of their advisory role in undergraduate research work as a form of teaching and described how they used the opportunities inherent in authentic research projects as everyday teaching tools to accomplish their research goals while also meeting students' educational needs. For example, they articulated a variety of strategies for helping students to become independent problem solvers, for normalizing the inevitable messiness and risk of carrying out authentic science research, and for teaching students to use their peers as a sounding board to work out ideas and practice communication skills. Student data clearly reflect how these strategies contributed to the students' development as researchers. The study is significant in revealing just how faculty make use of the authentic learning context to accomplish explicit learning objectives for their research students.

See Laursen, S., Hunter, A.-B., Seymour, E., Thiry, H., & Melton, G. (2010). *Undergraduate Research in the Sciences: Engaging Students in Real Science*. San Francisco: Jossey-Bass. Appendix C contains the interview protocols.

FACULTY INTERVIEWS

Because they are so labor intensive, many faculty interviews are conducted as part of a research study. As discussed in the next section, observation instruments sometimes are accompanied by a short interview. But interviews can be used alone, to identify faculty teaching practices. A number of studies also document faculty beliefs and decision making about teaching and learning. Almost all of the studies make use of semistructured, open-ended interview protocols. A direct focus on faculty teaching practices was an uncommon goal for interviews with faculty; nonetheless, some important insights about teaching have been gained in previous studies of faculty interviews.

Another use of faculty interviews is for a needs assessment—that is, an attempt to gain understanding about the specific needs of the faculty member being interviewed. A needs assessment is typically done by a center for teaching and learning, and allows the center to target professional development activities on an individual basis. Good examples of this use of faculty interviews related to teaching practices are given at the University of Maryland Teaching and Learning Center website, http://cmns-tlc.umd.edu/tlcmeasurementtools.

Use of a Specific Innovation

An example of faculty interviews that ask about specific teaching practices is the Innovation Configuration that is part of the Concerns-Based Adoption Model (Hall and Hord, 2001; Heck, Stiegelbauer, Hall, & Loucks, 1981). Although not STEM specific, the Concerns-Based Adoption Model is a well-developed perspective on how teachers adopt new teaching strategies that were previously developed by others. Documenting how faculty members implement a new teaching strategy (i.e., their innovation configuration) is an important component of the model. Individual interviews or completed checklists can be used to measure the innovation configuration against a set of prescribed innovation components. Interviews are recommended especially for more complicated innovations. The interview protocol begins with open-ended questions, such as one requesting a description of how the instructor is using the innovation, and then becomes more specific—for example, asking about how the instructor assesses student learning. The interviewer also probes for specific details of the innovation if details are not given.

Investigating Teaching Practices

A number of research studies have used interviews to investigate faculty decision making regarding their teaching practices. These studies not only describe aspects of faculty teaching, but also attempt to understand what factors influence the practices.

Hora (2012) and Hora and Anderson (2012) conducted interview-based studies with faculty to identify organizational influences on, and perceived norms for, interactive teaching. Semistructured interviews with both specific and broad questions were carried out, and data were collected on factors such as the organizational context and its influence on teaching. Henderson and Dancy (2007, 2008) conducted an interview study with physics faculty to identify teaching practices and decision making about teaching. Similar to Hora's interviews, Henderson and Dancy's were semistructured, with questions that started broadly and included possible probing questions designed to gather more detail about issues that were not covered in the initial response. For example, in the interview protocol, the broad questions, such as "describe your introductory quantitative physics class?" "how is your course structured?" "what happens during class time?" and "what do you require students to do outside of class?" were followed by possible probing questions asking what students are required to do in class (e.g., listen, write, read, speak, share ideas with others), how open ended class time is structured (e.g., stick to a rigid schedule or adjust the schedule on the basis of students' responses), how students are assessed and what typical exam questions are like (e.g., similar to homework, conceptual, mathematical, open ended), and how assessments are graded (e.g., on an all-or-nothing basis, with partial credit, with feedback given). Another example of a semistructured interview protocol with both specific and broad questions is part of the Marbach-Ad et al. (2012) studies; the full protocol is available online at the University of Maryland Teaching and Learning Center website, http://cmns-tlc.umd.edu/tlcmeasurementtools.

Yerushalmi and colleagues used interviews based on concrete instructional artifacts to create simulated teaching environments (Henderson, Yerushalmi, Kuo, P. Heller, & K. Heller, 2004; Henderson, Yerushalmi, Kuo, K. Heller, & P. Heller, 2007; Yerushalmi, Henderson, K. Heller, P. Heller, & Kuo, 2007; Yerushalmi, Cohen, K. Heller, P. Heller, & Henderson, 2010). In these studies, faculty looked at several artifacts representing a range of teaching practices and described their practices and the reasons they adopted them. For example, one set of artifacts contained solutions to three different sample problems. Interview questions related to this artifact consisted of (1) general, open-ended questions designed to gather information about an instructor's ideas and (2) specific questions, often related to an artifact, designed to gather information about an instructional decision. An example of the latter type of questions is "here are several instructor solutions. … describe how they are similar [to] or different [from] your solutions." The complete protocol and artifacts are available as a supplemental appendix to Henderson et al. (2007) and are online at http://prst-per.aps.org/abstract/PRSTPER/v3/i2/e020110.

Beliefs about Teaching and Learning

Another set of studies focuses on identifying faculty beliefs about teaching and learning on the assumption that these beliefs are related to teaching practices. Like the studies of faculty decision making, these also provide information about faculty teaching practices.

Martin, Prosser, Trigwell, Ramsden, and Benjamin (2000) and Samuelowicz and Bain (1992) conducted interviews with college faculty to identify their conceptions regarding teaching and learning. Samuelowicz and Bain (1992) interviewed 13 instructors for 60 to 90 minutes each. The interviews were semistructured and were based on 14 questions. One question—"What is teaching?"—was aimed at eliciting conceptions of teaching. Two groups of questions, one focusing on teaching practice and another on student learning, sought to gain as full a description of the conception of teaching as possible in more indirect ways. The questions in the first group dealt with such aspects as the aim of teaching and the teacher's role; course design and revision; what is regarded as good teaching; pleasant and unpleasant teaching experiences; perceived obstacles to good teaching; and awareness of teaching conceptions believed to be held by others.

Kember and Kwan (2002) conducted interviews with 17 lecturers in three departments. The semistructured interviews focused on five broad areas. The first aimed at eliciting the faculty members' conceptions of good teaching. The second focused on the motivational strategies they employed in their teaching practice. The third inquired into the types of learning activities they expected their students to undertake inside and outside the classroom. Finally, instructors were asked about what they felt were the most effective teaching strategies and about their perceptions of how they taught classes with different types of student populations. During the analysis, interviewees were rated in terms of their conceptions of good teaching (e.g., transmission of knowledge vs. facilitation of learning) and their self-described teaching approaches (e.g., content centered vs. learning centered). The authors reported a very high correlation between conceptions of good teaching and teaching approaches.

Brown, Abell, Demir, and Schmidt (2006) conducted interviews designed to get an understanding of how STEM faculty viewed inquiry-based instruction. The researchers also set out to describe faculty perceptions of the challenges, constraints, and opportunities associated with designing and teaching inquiry-based labs.

STUDENT INTERVIEWS

Student interviews are used primarily as a way to understand students' perceptions of, or reactions to, aspects of teaching and learning. Student interviews focusing on a particular course can be extremely useful in improving the course. They can also be used in research studies to better understand aspects of the student experience. Although not commonly requested, important information could come from student interviews sometime after (e.g., six months or one year) they complete a course or from students who dropped a course.

Student Interviews for Course Improvement

One common use of student interviews is for course improvement. These interviews are often done in groups because the resources required for individual interviews may be prohibitive. For example, in their book on classroom assessment techniques, Angelo and Cross (1993) described the Group Instructional Feedback Technique (GIFT), a method for gauging students' reactions to teaching and teachers. A peer (e.g., another faculty member) interviews the students of his or her colleague to understand what is helping or hindering the students' learning and to solicit suggestions for improvement. Campus-based teaching and learning centers offer similar services, often to instructors who want to receive feedback at midsemester. A web search will identify many variations on the basic GIFT procedures. An example from Chemeketa Community College is found at http://oppcenter.chemeketa.edu/documents/GIFTFacilitatorsRole.pdf. A similar approach is known as the Small Group Instructional Diagnosis (SGID). Details and specific examples are found at http://wikipodia.podnetwork.org/Home/topics-for-discussion/small-group-individual-diagnosis.

Sheppard, Johnson, and Leifer (1998) described a more in-depth procedure for student involvement in the measurement of teaching practices. They used a list of important aspects of instruction, including instructor–group interactions, instructor–individual interactions, instructor dynamism and enthusiasm, an analytic–synthetic approach, organization and clarity of instruction, effort, and continuous development of the curriculum. At mid-term the faculty member whose course was being reviewed prepared a reflective memo, and at the end of the course two faculty peers convened student focus groups, using the list of aspects of instruction as the basis for discussion. On the basis of a videotape of the focus group sessions and the reflective memo, peers developed a "summary memo" with information that the instructor could use to reflect on and revise his or her teaching.

Student Interviews for Research

Although individual student interviews are often too resource intensive to use in seeking course improvement, they can be a valuable data source in research studies. For example, in their well-cited study *Talking about Leaving*, Seymour and Hewitt (1997) conducted interviews with hundreds of students with strong Math SAT scores who intended to major in a STEM field. The semistructured interviews lasted

between 45 and 75 minutes, were conducted in a conversational manner, and focused on students' experiences in STEM courses at their current institution and in other situations (e.g., high school). The study found that nearly all of the students who switched from a STEM major, as well as many of the students who persisted in a STEM major, cited poor teaching as a problem with introductory STEM courses.

In a more recent study, Thiry, Weston, Laursen, and Hunter (2012) conducted 40- to 80-minute interviews with 73 students. The semistructured exploratory interviews were designed to get information about students' perceived benefits from participating in undergraduate research, about the actual work that they were involved in, and about their interactions with their advisor and research group.

BOTTOM LINE

Although resource intensive, interviews provide an opportunity for researchers to explore complex and ill-defined problems; develop more in-depth understanding; explore faculty and student perceptions; pursue questions of causality; develop or validate other measurement methods, such as quantitative instruments and observational protocols; and document teaching practices for professional development. Researchers using interviews may need to educate STEM faculty about the value of qualitative research and the methodologies used to analyze data.

OBSERVING STEM TEACHING

Observation involves watching faculty and students in their natural teaching and learning environments, including classrooms, labs, the field, office meetings, and consultations. This method directly documents teaching practice as it unfolds, with the observers taking notes or categorizing instructional behaviors, either in real time or as they watch videotapes. The observers use a well-developed protocol to guide their observations. Although the technique has been practiced most widely in K–12 schools, classroom observations are being used increasingly in colleges and universities to support the study and improvement of STEM teaching.

Researchers, administrators, and others interested in supporting professional development or evaluating STEM faculty teaching for promotion and tenure often turn to observations in the classroom to document teaching practice. By documenting specific behaviors and student engagement, observers can document whether or not a particular approach to teaching is occurring, whether or not the observed interactions are consistent with the theory or goals of that approach, and how students are responding to the approach. Observation can also help investigators compile a list of consistent faculty behaviors and instructional designs that are routinely experienced in undergraduate teaching environments and that can contribute to and guide longitudinal research and reform efforts.

Although classroom observations can be used for a variety of purposes in undergraduate settings, they are most commonly used to support professional development activities or to evaluate teaching quality. For example, some undergraduate teaching and learning centers offer coaching and mentoring services in which a trained faculty developer observes a class, often with the use of a structured protocol, and then meets one-on-one with the instructor. In these cases, faculty developers frequently integrate pre- and post-class interviews or coaching sessions and provide targeted feedback to the instructor. Others use classroom observations for the purpose of evaluating teaching for employee performance reviews, usually in concert with other evaluation measures. Most often conducted by peers within a given academic department, the protocols vary from structured to unstructured (taking notes). Unstructured protocols

that allow for the emergence of new observation categories can be quite valuable in detecting unsuspected, but important, elements of teaching. To take full advantage of unstructured observations requires a familiarity with qualitative research methodology and theories of teaching and learning. Unfortunately, unstructured observations by untrained observers are too often used to assess faculty teaching effectiveness, resulting in inconsistency, lack of specificity about what practices or incidents are considered important, and charges of subjectivity.

Many observation protocols require observers to document certain aspects of classroom practice as well as evaluate the quality of teaching. This approach is attractive to those wanting a single measure of whether or not teaching reflects particular standards or expectations. The use of faculty observations to document STEM teaching practice has both inherent strengths and inherent challenges that should be evaluated in considering this measurement technique:

Observation Strengths

- Observations can capture contextualized accounts of what STEM instructors do and when they do it, as well as information about instructor–student interactions.
- Data obtained can be used for multiple purposes, including professional development, assessment, program evaluation, and research.
- Observations can result in specific and actionable knowledge that could be used to coach and mentor faculty.
- Observations can document nuances and details of practice dynamics that are not documentable through other techniques.
- Data obtained from observations are often perceived as more objective than self-reported data supplied by faculty members.

Observation Challenges

- Training is required in order to obtain reliable data across multiple observers.
- Because observations occur in real time, they are resource intensive and it is typically possible to observe only a small portion of any course; thus, observation of a single class session may not be representative of other class sessions.
- Observations can describe only that which is observable (i.e., some aspects of teaching cannot be captured through observation alone).
- It is a misconception that experts in observation who lack direct content knowledge cannot contribute to an observation-based evaluation.
- Some faculty are concerned that observers will upset classroom dynamics.

IMPLEMENTING SUCCESSFUL FACULTY OBSERVATIONS

When choosing or developing an observation protocol, investigators should clearly define what aspects of teaching will be described and then realistically anticipate how many categories one observer can attend to in real time while also addressing more summative aspects of teaching. Researchers should consider any biases observers (and even observed faculty members) might bring to the classroom environment during the observations and should consider whether they can be minimized in the development of the protocol. For example, observers with expertise in a particular field may be impressed by a coherent lecture and miss student disengagement or lack of understanding. It is also important to think about which types of teaching, and how many instances of each type, should be observed. In observing class sessions, as the stakes for the observations increase, so, too, should the number of class sessions observed in order to ensure that the observed sessions are representative.

The development of trust between faculty and observers is critical and can help create a neutral, nonthreatening atmosphere for those observing teaching practices. Such an atmosphere is particularly important, given evidence that some faculty are uncomfortable about being observed. A brief faculty interview in advance of the observation can help smooth the way as well as identify instructor goals, while a follow-up meeting can help confirm that the observations are representative. Videos can also help adjust for any individual biases, allowing for more extensive analysis, perhaps by more than one observer.

Caution should be exercised in using observation protocols for evaluative purposes, for two reasons. First, evaluative measures may “turn off” or alienate faculty and therefore be of limited utility for professional development purposes. Second, because observers make different judgments about what constitutes quality teaching, reliability is difficult to attain when analysts are required to not only describe teaching but also judge its quality in real time. In addition, asking observers to consider multiple factors, such as teaching methods, student engagement, and content assessment, can add significantly to the demands placed on the observers.

Developing Observation Procedures

As with any other measurement instrument or research, observation procedures should be designed, tested, and implemented, with careful attention paid to the methodological quality of the instrument. In particular, the key questions of validity and reliability of the procedures must be addressed. With respect to observations, some of the difficult aspects of validity have to do with the extent to which observed behaviors are open to interpretation by the observers, as well as how well the procedures adequately measure the domain of interest, in this case STEM teaching. As regards reliability, observers must consistently measure the same thing each time data are collected and different observers must use the procedures in a similar fashion while scoring the same instructor. These concerns should be addressed in the overall design and implementation of the observation procedures.

Clearly, much more goes into designing effective observation procedures than simply reproducing a check-the-box rating or asking a single evaluator to describe what he or she observed. Just as with other forms of research, time, care, and expertise go into the development of effective observation procedures and observers require training on how to use them effectively.

OBSERVATIONAL INSTRUMENTS

Observational instruments are divided into two varieties. With a holistic instrument, the observer watches an entire class session and then rates each item with regard to the lesson as a whole. With a segmented instrument, the class session is divided into short periods (usually a few minutes each) and the observer rates each item as it occurred in each period.

Holistic Observational Instruments

Reformed Teaching Observation Protocol (RTOP): The RTOP is a widely used classroom observation protocol, particularly among researchers and evaluators interested in "reformed" teaching practices. Based on the constructivist literature about teaching and learning, the RTOP focuses on the extent to which instructors adhere to those practices identified with the inquiry- and standards-based literature. Using a five-point scale ranging from "never occurred" to "very descriptive," all of the instrument's items measure the extent to which various practices are observed in the classroom. Sample items include "The teacher's questions triggered divergent modes of thinking," "participation of students was encouraged and valued," and "The focus and direction of the lesson was often determined by ideas originating with students." The content being taught in the class that is observed is not a central focus of the protocol. One of the critiques of the RTOP's forced-choice response options is the absence of a "not applicable" option, a shortcoming that may result in implausible ratings in some cases. The RTOP reference manual is found at http://www.public.asu.edu/~anton1/AssessArticles/Assessments/Chemistry%20Assessments/RTOP%20Reference%20Manual.pdf.

The UTeach Observation Protocol (UTOP): The UTOP was developed at the University of Texas at Austin to assess overall quality of instruction. The instrument is based on protocols created by Horizon Research and designed to evaluate the UTeach program. The protocol outlines UTeach expectations for quality instruction, as well as national reform standards. The developers argue that the UTOP is appropriate for describing a range of teaching styles without favoring one over another. Indicators (each rated on a scale of 1 = "not observed at all" to 5 = "observed to a great extent") used to rate class sessions include "The majority of students were on task throughout the class," "The structure of the class included opportunities for the instructor to gauge student understanding," "The resources selected for this class contributed to the purposes of the instruction," and "The significance of the math and science content, including how it fits into the 'big picture' of the discipline, was made explicit to the students." The protocol is found at https://wikis.utexas.edu/display/physed/UTeach+Observation+Protocol.

The Oregon Collaborative for Excellence in the Preparation of Teachers (OCEPT) Classroom Observation Protocol (OTOP): The OTOP was developed to study the effects of an instructional intervention in Oregon. To create the instrument, researchers focused on teacher and student behaviors. They included only 10 items in the protocol, but also included a post-observation interview to complement each observation, validate the data obtained, and elicit instructors' views on their own teaching. Each of the 10 items is rated globally on a scale of 1 to 4 based on a set of possible indicators. Items include "The lesson encouraged students to seek and value various modes of investigation or problem solving" and "The lesson promoted strongly coherent conceptual understanding in the context of clear learning goals." The OTOP Outcomes Research Study is found at http://opas.ous.edu/Work2009-2011/InClass/OTOP%20Instrument%20Numeric%202007.pdf.

Teaching Behaviors Inventory (TBI): The TBI protocol attempts to capture the key aspects of teaching behavior hypothesized to be linked to effective instruction and student learning (Erdle & Murray, 1986; Murray, 1983). The instrument does not require the analyst to judge the quality of instruction per se, but simply to report whether or not a particular teaching behavior occurred. The 95 specific classroom behaviors are rated on a five-point scale (1 = never; 2 = rarely; 3 = sometimes; 4 = often; 5 = always) and are grouped into 14 teaching behavior factors. Specific behaviors include "talks with students before or after class," "tells jokes or humorous anecdotes," "states objectives of each lecture," and "speaks in a monotone."

The TBI is evaluative in that the categories that it comprises, such as clarity and enthusiasm, are believed to be associated with effective teaching. The TBI has a significant history of use in empirical research and in publications about undergraduate teaching and is commonly adopted by teaching and learning centers as an easy-to-use protocol for peer review or professional development. The TBI is found at http://www.calvin.edu/admin/provost/documents/behaviors.pdf.

Segmented Observational Instruments

Flanders Interaction Analysis (FIA): The FIA is based on the assumption that interactions between students and teachers represent a key aspect of effective classrooms. The FIA distinguishes between two types of teacher talk in the classroom: direct teacher talk (i.e., lecturing, giving directions) and indirect teacher talk (i.e., praising, asking questions). It defines student talk as either a response (i.e., an answer to a question that has been posed) or an initiation (i.e., an interaction initiated by a student). Analysts code each type of talk every three to five seconds, with the intersection between the two representing the interaction in the classroom. While the FIA has been used mostly in K–12 classroom research, some (e.g., Gilbert and Haley, 2010) have argued for more widespread use in undergraduate settings because data obtained with the protocol are easy to log into the matrix and interpret. More information on FIA is available at http://files.eric.ed.gov/fulltext/ED088855.pdf.

Teaching Dimensions Observation Protocol (TDOP): The TDOP was designed as part of an NSF grant to study the cognitive, cultural, and organizational factors influencing instructional decision making and classroom practice in STEM departments. Based on a middle-school protocol, the instrument was substantively revised and adapted to specifically fit undergraduate classroom practices. The TDOP documents six categories of teacher and student behaviors: teaching methods, pedagogical strategies, cognitive demand, student–teacher interactions, student engagement, and instructional technology. A core feature of TDOP data collection is that the rater records observed behaviors in each of the six categories for two-minute intervals throughout the class period. Codes for the teaching methods category include “lecture,” “interactive lecture,” “small-group work/discussion,” and “student presentation.” Codes for pedagogical strategies include “moves into audience,” “humor,” “reads,” “assessment,” and “administrative task.”

In addition to collecting data on the observed behaviors, analysts take notes about the class content and other features of interest to the observer. The TDOP is available on a web-based platform, so all data collection, interrater reliability testing, and data management are automated. The protocol is found at http://tdop.wceruw.org/.

VaNTH Observation System (VOS): The VOS was designed by an NSF-supported multi-institutional research center, named VaNTH for the collaborating institutions: Vanderbilt University, Northwestern University, the University of Texas at Austin, and the Health Sciences and Technology Program at Harvard and Massachusetts Institute of Technology. The VOS involves collecting four types of data: student–teacher interactions, student academic engagement, narrative notes of classroom events, and ratings of specific indicators of effective teaching. As with the TDOP, the first three types of data are captured in repeating five- to six-minute cycles. Student–teacher interactions are coded in terms of *who* said *what* to *whom, how* the interaction was framed, and with what *media* the interaction was recorded. Student academic engagement is measured by counting the number of students engaged in both desirable and undesirable classroom activities. Narrative notes are taken to identify the content of the lesson, the type of instructional strategy used, and any extenuating circumstances. Finally, after the class session, the observer rates the lesson on 17 items (using a scale of 1 = slightly/somewhat, 2 = moderately, 3 = thoroughly/ well developed). Items include "making connections to prior learning," "ongoing assessment of students' understanding during a lesson," and "moving among students." (See Harris and Cox, 2003, and Cox and Cordray, 2008, for more details.)

Classroom Observation Rubric: The Classroom Observation Rubric was developed to focus on student–teacher dialogues in the context of the use of a clicker in undergraduate physics courses. The protocol focuses on the types of questions posed with clickers (e.g., content oriented or logistic); response options; the distribution of student responses; the professor's wait time for responses; actions taken during the wait time; and interactions, or what the researchers call "dialogic interactions." In their use of the Classroom Observation Rubric, Turpen and Finkelstein (2009) cite case studies of six physics professors to show how variations in these aspects of student–teacher dialogues led to different classroom norms that likely influenced the depth of student learning. (See Turpen and Finkelstein, 2009; the instrument is found at http://prst-per.aps.org/multimedia/PRSTPER/v5/i2/e020101/e020101_app.pdf.)

BOTTOM LINE

Observations are a strong method for documenting STEM teaching when they are conducted under well-defined protocols that capture what happens in a class session without the observer's subjective judgment regarding quality or the impact on student learning clouding the picture. Still, because many important aspects of teaching are not easily observable, participants at the national AAAS-hosted meeting strongly recommended adopting a mixed-methods approach with more than one set of observations, particularly when the results are used in high-stakes situations such as promotion and tenure decisions.

TEACHING PORTFOLIOS

Teaching portfolios afford an opportunity for STEM faculty to showcase their teaching through annotated collections of text and supporting artifacts. Much as the curator of a museum selects and interprets a collection of specimens, effective teaching portfolios guide the reviewer through the meaning of selected artifacts. Done well, portfolios not only provide the data needed to help document teaching, but also result in measurements that can be used as a baseline for subsequent change-oriented actions. Sometimes just the act of constructing, discussing, and evaluating teaching portfolios can lead to improvements in teaching. Although much of the scholarly work on portfolios has focused on K–12 classroom use, portfolios are becoming more common in documenting undergraduate teaching.

A portfolio offers a good overview of STEM teaching through the collection of representative activities combined with evidence of their effectiveness. Using teaching and learning artifacts as forms of documentation, an informative portfolio reflects an instructor's teaching practice. Portfolios usually include a statement about, or discussion of, what the portfolio documents, the instructor's philosophy of teaching and learning, and the guidelines applied for the creation of the portfolio. Portfolios can also include everything from a course syllabus, to demonstrations (e.g., video clips, examples of graded work) of how the instructor teaches and interacts with students, to examples of tests and written student work.

Portfolios are most often used to document teaching practice as part of a formal initiative in which faculty are given guidelines for assembling them. The guidelines usually outline how the portfolios should be created (e.g., individually or as part of a group process), what they should look like (e.g., electronic or hard copy), how they should be annotated, and how they will be analyzed and used. Specific choices for these variables will influence what can be measured and how the resulting measurements can be interpreted. Often, administrators and department heads request teaching portfolios for the purposes of promotion and tenure decisions or for rewarding teaching excellence. But teaching portfolios also can be used to encourage reflective practice and may serve as anchors for scholarly communities around teaching.

Like the other instruments examined in this report, portfolios present strengths and challenges:

Portfolio Strengths

- A portfolio results in rich multimedia depictions that encourage creativity, making it easy to represent different aspects of teaching and to create the potential for triangulation of claims.
- A portfolio captures both espoused practice (i.e., what is claimed in the teaching philosophy) and enacted practice (i.e., what is demonstrated via the artifacts) and therefore can help faculty members identify gaps by themselves.
- A portfolio provides a foundation for advising STEM faculty, either individually about teaching or in concert with others to describe the teaching of a department or a specific program.
- A portfolio can function as a valuable research tool when its contents are analyzed with an eye toward achieving a better understanding of how educators think about and practice teaching.

Portfolio Challenges

- Assembling a portfolio and evaluating it are both time intensive.
- Having faculty make their personal practice and teaching philosophy public is contrary to norms at most institutions.
- Because of their person-specific nature, portfolios can be difficult to interpret, measure, and compare.
- A portfolio is developed by an instructor and therefore represents his or her perceptions and perspectives, which may not align with other measurements.

IMPLEMENTING SUCCESSFUL TEACHING PORTFOLIOS

Teaching portfolios are highly personal and can include a number of different materials, depending on their purpose as well as the instructor's choices about what is important. Diverse depictions of teaching can be challenging to compare across instructors; guidelines for the construction of a portfolio can help streamline its preparation and analysis. For example, an undergraduate STEM teaching portfolio might include, but not be limited to, the following items:

- Teaching statement or other philosophy
- Examples of practice and why they were selected
- Statement of learning objectives, such as a course syllabus
- Demonstrations of how the instructor communicates with students (e.g., through e-mails or short videos)

- Statement about office hours or other insights into informal interactions with students
- Representative examples of student work
- Demonstrations of how the instructor gives feedback to students (e.g., graded work)
- Documentation of interactions (e.g., with colleagues, at professional meetings, at professional development sessions on campus) involving teaching and classroom practice
- Publications that highlight teaching practice (not just content expertise)
- Student course evaluations or classroom observations by colleagues or administrators
- Teaching honors or other recognitions

Depending on their final use, guidelines can be well defined or left open for interpretation. Clearly, guidelines that are more prescriptive will result in portfolios that can be more readily analyzed and compared.

However, as with any other form of self-reported data, the teaching philosophy that drives the portfolio represents what an instructor wants to say about his or her beliefs and practices related to teaching. Moreover, the artifacts selected to illustrate these beliefs and practices represent events taken out of context and framed in a particular way. Thus, an instructor's choice of what goes into a portfolio constitutes a kind of selection bias that needs to be taken into account by anyone interpreting the contents of the portfolio.

Significantly, the construction of a teaching portfolio has the potential to be a learning experience for faculty. This potential can be viewed as a strength, because helping faculty become more reflective about their teaching is central to improving the quality of teaching and learning in undergraduate STEM education. When associated with efforts to describe teaching, however, the potential for change can be problematic in that the measurement activity itself could alter the phenomenon being measured. Like all of the other approaches discussed in this report, portfolios are most revealing when viewed in combination with other measurements.

PORTFOLIO EXAMPLES

As mentioned earlier in the sidebar on page 6, Nancy Chism's *Peer Review of Teaching: A Sourcebook* provides a basic overview of how portfolios can be used to describe teaching in higher education across disciplines. This section introduces four applications of teaching portfolios involving STEM higher education. None of the applications focus on analyzing the content of the portfolios as a means to describe teaching, although that could have been done. Two of the applications focus specifically on graduate students.

Knowledge Exchange Exhibition and Presentation (KEEP) Toolkit

In 2002, the Carnegie Foundation for the Advancement of Teaching developed a collection of web-based, open-source tools to help its scholars at the Carnegie Academy for the Scholarship of Teaching and Learning (CASTL) "document, share, and reflect on some of the critical aspects of their efforts in transforming teaching and student learning." The Knowledge Exchange Exhibition and Presentation (KEEP) Toolkit website, opened to the public in 2004, helped postsecondary faculty create compact and compelling multimedia representations of their efforts to improve undergraduate teaching and learning. Using KEEP toolkit templates, faculty and instructors can upload and organize instructional materials into web-based "snapshots," describe the purpose of those materials, and include evidence to support claims of their effectiveness. The snapshots can then be used in various ways: as an instructional tool in the classroom; to prompt discussion with colleagues; and to make scholarship of teaching and learning efforts publicly accessible. Examples of KEEP snapshots made by CASTL scholars are found at http://gallery.carnegiefoundation.org/gallery_of_tl/castl_he.html.

In October 2009, the KEEP Toolkit was transferred to MERLOT, a nonprofit consortium that facilitates the use of peer-reviewed online instructional tools. MERLOT not only hosts current and future KEEP snapshots, but also provides user support services and an opportunity to have KEEP snapshots catalogued in the MERLOT repository.

Disciplinary Commons

In the Disciplinary Commons model, practicing computer science educators came together to discuss introductory computer science courses, with the preparation of a portfolio framed as an opportunity to gain useful ideas through interaction with others (and with the educators themselves). The portfolios produced were course specific and included course objectives, institutional and curricular context, course content and structure, teaching methods, teaching philosophy, evidence of student learning, grading, and self- and peer evaluation. The project supported its participants by organizing sessions at which they shared portfolio elements and got feedback and instructions. (See Tenenberg and Fincher, 2007; information on the Disciplinary Commons is found at http://depts.washington.edu/comgrnd/.)

Engineering Teaching Portfolio Program (ETPP)

The ETPP is a process devoted to helping engineering graduate students prepare for future teaching responsibilities. Portfolios produced were comprehensive and included a statement about teaching, five annotated artifacts, and a statement about racial and ethnic diversity. Graduate students were supported as a group as they prepared the portfolios and received feedback and instructions. (See Linse, Turns, Yellin, & VanDeGrift, 2004; information on the ETTP is found at http://faculty.up.edu/vandegri/Tenure/Papers/ASEE04/PreparingFutureFaculty.pdf)

Portfolios to Professoriate (P2P)

P2P is an initiative that uses the construction of a portfolio to help engineering graduate students prepare for the teaching and research responsibilities of their upcoming academic careers. The students produce comprehensive professional portfolios of teaching-related materials, including a teaching reflection and teaching artifacts. (See McNair and Garrison, 2012; information on the P2P program is found at http://www.asee.org/public/conferences/8/papers/5477/view.)

BOTTOM LINE

Constructing a teaching portfolio is a time-consuming and potentially emotional exercise. Without careful planning, it can result in noncompliance or in portfolios that are difficult to interpret. Therefore, regardless of the primary goal of constructing the portfolio—from documenting teaching practice to self-study—portfolio initiatives need to address ways to support the construction of teaching portfolios, particularly by STEM faculty. Although additional research is needed into the most effective use of portfolios in describing undergraduate STEM teaching, portfolios clearly have the potential to play a significant role in changing teaching for the better.

SELECTING AND COMBINING MEASUREMENT APPROACHES

Many STEM faculty, college and university administrators, state and national policymakers, education researchers, and other STEM education stakeholders have a commitment to improving undergraduate teaching. Whether the ultimate goal is to improve STEM teaching and learning or to respond to external requests (such as accreditation or grant requirements), describing the complex activity of teaching requires multiple measurements and approaches. All aspects of describing STEM instructional practice—from developing and implementing specific instruments or protocols to analyzing results—require a significant investment of time, energy, resources, and expertise to be successful. Table 1 provides an overview of some of the important types of uses, challenges, and choices involved with each technique.

Collecting data to describe STEM teaching presents the additional challenge of ethically studying faculty and students. Most campuses have guidelines and protocols that restrict how human subjects may be studied and how the resulting data may be used. Throughout the national AAAS meeting, participants recommended that those interested in collecting data on STEM teaching first consult with their human subjects institutional review board before initiating any studies, even if they are intended for internal evaluation only.

The Importance of Triangulation

While each of the methods discussed in this report has its own particular strengths, meeting participants urged caution in selecting one method over another, particularly in high-stakes situations, such as promotion and tenure decisions. Each measurement approach has its own benefits and applications to which it is well suited, but any plan that relies on a single method will be biased by the deficiencies associated with that method. As a result, measurement approaches that use multiple methods of data collection (i.e., triangulation) will yield more useful results than approaches that do not. Triangulation can result from a combination of different measurement approaches, the integration of qualitative and quantitative data, or the collection of data from both faculty and students. These mixed-methods approaches will help contribute to a better understanding of STEM teaching practice.

DESIGNING AND CONDUCTING MIXED-METHOD APPROACHES

A useful resource for those interested in exploring the benefits of taking a mixed-methods approach to research is Creswell, N., and Clark, V. (2007). *Designing and Conducting Mixed Methods Research*. Thousand Oaks, CA: Sage.

Table 1. **OVERVIEW OF THE FOUR BASIC MEASUREMENT TECHNIQUES**

CATEGORY	SURVEYS	INTERVIEWS	OBSERVATIONS	PORTFOLIOS/ ARTIFACTS
Purposes/Uses	Learning about practices, motivations, beliefs, and attitudes; professional development	Gaining a deeper understanding of respondent's interpretations, intentions, and perceptions; flexible and responsive	Offer a rich description of visible aspects of teaching practice	Provide documentation of intention and outcome
Typical Uses in STEM Undergraduate Education	Course evaluations completed by students; survey of teaching practices and attitudes completed by faculty members	Research studies; voluntary consultations with teaching and learning center staff	Faculty peer evaluations for personnel decisions (promotion and tenure), research and evaluation studies	Voluntary faculty professional development
Limitations/ Challenges	Relatively low response rates and possibility of nonrepresentative sampling (e.g., enthusiastic faculty may be more likely to respond to a survey about teaching); items may be interpreted differently by different people	Unrecognized interviewer bias that influences the conduct or interpretation of interviews; power imbalance between interviewer and interviewee	Unrecognized observer bias that influences the awareness or interpretation of observed behavior; power imbalance between person observed and observer; some important aspects of teaching are not observable	Absence of annotation can lead to shallow interpretation; non-uniform as research data
Requisite Resources	Knowledge of quantitative data analysis (and perhaps qualitative analysis for open-ended items); knowledge of guidelines for constructing good questions	Time for transcription and analysis; multiple analysts with experience analyzing qualitative data	Procedures to guide and document observations; time for interpretation; time to conduct observations	Guidelines for instructors for creating a portfolio and for selecting artifacts; time for interpretation
Key Choices	Forced-choice vs. open-ended questions; whom to sample	Level of structure in protocol; whom to sample	Methodology (descriptive, evaluative, ethnographic); whom to sample; timing of observations	Medium of portfolio (digital vs. physical); level of structure in guidelines; how to analyze data; whom to sample
Guidelines for Ensuring Validity and Reliability	Pilot instruments; consider established and validated instruments; test and validate one's own instrument	Pilot protocols; consider relationship between interviewer and interviewee; use follow-up questions for clarification— multiple analysts are needed for best results	Document the evidence; debrief instructor	Include annotations; include samples of student work

THE WORK HAS JUST BEGUN

Although a number of instruments and protocols are available for use or adaptation in STEM learning environments, many questions remain within the academic community about what techniques and protocols are appropriate for describing STEM teaching. Questions also persist about the validity of the resulting data. As a result, researchers and others interested in describing teaching practice need to consider when and how to design new techniques specifically for STEM teaching applications, how those techniques should be used, how questions are posed to elicit meaningful results, and how to analyze the results from a variety of sources and across different institutional environments.

More research may even be needed into how to advance the acceptance of measuring STEM teaching. For example, many academics do not see how describing STEM teaching could be viewed as research, because their experiences with surveys and classroom observations most often relate to program assessment or teaching evaluations. At many institutions, the measurement of teaching practices is seen as synonymous with student evaluations of teaching, of which many faculty are highly suspicious.

Whereas educational researchers may take for granted that being able to describe teaching is a necessary first step for advancing the study of teaching and learning, STEM faculty may not see or value the connections. Therefore, it is important to explain how the data collected can be used to support initiatives to improve student learning and retention, as well as what the educational research knowledge base has to say about how teaching choices affect these outcomes.

Clearly, the measurement of STEM instructional practice is not a trivial task. Moreover, without some level of confidence in the quality of the data collected, the descriptions obtained are of limited value in advancing STEM teaching and learning more broadly. The remainder of this section provides an overview of some of the outstanding research questions uncovered during the course of the three-day national meeting.

MEASURING STEM TEACHING: WHERE ARE WE NOW?

Many assume that colleges and universities have programs in place to measure STEM teaching and that much is known about teaching practice nationwide. However, although teaching practices have been documented at the classroom level or even across departments or institutions, few large-scale studies have been conducted to describe STEM teaching practices across different types of institutions or nationwide. Using well-defined and documented measurement techniques, large-scale studies could provide policymakers and researchers with a detailed accounting of the state of the nation's undergraduate STEM classrooms, the impact of more diverse student populations on teaching dynamics, and faculty beliefs about teaching. Such studies could also provide the baseline data needed to advance both STEM teaching and STEM learning.

In addition, much can be learned from the existing large-scale surveys of faculty teaching practices discussed earlier. For example, organizations such as HERI often break out results by discipline and also offer researchers the possibility of conducting secondary analyses of data from the organization's extensive database of faculty survey responses.

Measuring STEM Teaching in Online Education

The discussion related to the measurement of teaching practices presented in this report mirrors the current state of measurement activities in that they are almost entirely situated in traditional in-person educational settings. With the rapid increase in the popularity of online education, it is important to develop measurement techniques and tools to describe this type of teaching. In some cases, measurement techniques from in-person classes can be adapted fairly easily. In other cases (e.g., observations), it will be necessary to develop new techniques. The structure of online teaching also offers fresh opportunities for using measurement techniques that are not possible to use in a more traditional teaching environment.

Faculty Self-Reported Data

Researchers need to know more about the connections between faculty self-reports of their practice and their actual practice. Research into such a topic could provide valuable insights into STEM teaching by documenting when and why faculty perspectives stated in their self-reports on their teaching practices differ from the perspectives identified by other methods, such as observation. In addition, can researchers ask specific questions, in surveys or interviews, that will minimize the differences between perspectives stated in self-reports and perspectives identified by other methods? Knowing more about this issue can help researchers interpret past studies as well as help guide future investigations.

Student Self-Reported Data

Data generated by students, including course evaluations, assignments, and more focused commentary, could contribute critical insights into undergraduate STEM teaching. But more research is needed into what kinds of questions, time frames, and response formats yield the most useful and informative information. Researchers should also explore whether and how student perceptions of teaching behaviors vary over time and how disciplinary socialization influences the ways in which students perceive their instructors and instruction. Researchers can, of course, use data analysis to mine college and university databases to explore potential differences by field of study, but in-depth studies are needed to understand why some students perceive certain kinds of teaching practices in particular ways and whether or not these perceptions influence their perceptions of their instructors. Other areas for investigation include student experiences and perceptions of different classroom teaching practices and whether there are gaps between what STEM faculty intend and what students experience.

Response Rates and Nonresponse Bias

Students and faculty alike can become overwhelmed by electronic surveys and other education-related questionnaires—so much so, that they refuse to participate in them. In addition, some suspect that enthusiastic, confident teachers are more likely than others to participate in teaching-related studies. Similarly, many faculty believe that students are more likely to complete course evaluations or other data collection if they have particularly strong opinions about the course. More research is needed into how to improve response rates, particularly from students who can reach "survey fatigue" early in their academic careers as a result of attending to frequent requests for participation by their institutions or individual researchers. Can response rates be improved, for example, if the researcher explains why the surveys are being conducted, how the data will be used, and whether or not participants will have access to the results? How can sampling structures be used to increase response rates without sacrificing generalizability? Researchers need more insight into who does not respond to surveys and how their nonresponse might affect the resulting data.

Use and Development of Observation Protocols

Although observation is becoming a more widely used method for documenting STEM classroom practice, little is known about how observations are used across STEM disciplines, what documentation (if any) results from a given observation, how observation affects the observed and observer's practice, and what (if any) training is provided to observers. Also, it is still an open question as to whether and how content knowledge affects what observers document across STEM disciplines. Finally, more work needs to be done on the development of non-evaluative observational protocols that still capture the dynamics of STEM teaching and on minimizing the potential for disrupting those dynamics by having an observer in the classroom.

Validity and Reliability

In some areas, such as observations, a number of validated protocols already exist. In others, including interviews and surveys, research is advancing so rapidly that standardized instruments specific to STEM instruction have not yet been developed. Researchers should scrutinize existing instruments carefully with regard to their origins, psychometric properties, and issues related to their appropriate use in documenting STEM teaching. Also, more work needs to be done on judging the validity and reliability of observation protocols, especially in relation to multiple observers. With growing evidence that asking classroom observers to both describe and evaluate the quality of teaching can result in data that are less reliable, the use of these kinds of protocols and their interpretation of findings need further research. Studies combining and comparing the instruments and methods described in this report (along with others) may lead to additional insights regarding validity and reliability.

Advancing the Use and Analysis of Portfolios

Although portfolios have been used widely in K–12 environments, less is known about how to use and analyze portfolios to document undergraduate STEM teaching. Given the promise of this instrument as a measurement technique, research is needed to identify ways to increase, and develop guidelines for, the use of portfolios by undergraduate STEM instructors. At the same time, researchers need to better understand what aspects of STEM teaching can be measured by portfolios and, once that understanding is achieved, develop protocols to collect and analyze the data obtained. Finally, because the construction of a portfolio can be time consuming, more researchers should consider the difference that new technology might make in that regard.

Lessons from K–12 Instruction

Much of what researchers currently know about using observations, portfolios, and other instruments to measure teaching practices was developed and tested in K–12 classrooms. Investigators working on developing measurement protocols and instruments for use in undergraduate STEM classrooms should therefore explore the lessons learned—both positive and negative—in K–12 schools. (see, e.g., the Bill and Melinda Gates Foundation website, http://www.gatesfoundation.org/united-states/Pages/measures-of-effective-teaching-fact-sheet.aspx; and Alexandra Beatty, Rapporteur; Committee on Highly Successful Schools or Programs for K–12 STEM Education; Board on Science Education (BOSE); Board on Testing and Assessment (BOTA); Division of Behavioral and Social Sciences and Education (DBASSE); and National Research Council. (2011). Successful STEM Education: A Workshop Summary. Washington, DC: National Academies Press. The publication is available online at http://www.nap.edu/catalog.php?record_id=12820.) Note that, in adapting instruments from K–12 settings, it is important to consider how the instructional context differs in higher education.

Institutional Change

In much the same way that describing teaching practices can provide the insights and documentation needed to improve learning environments, so, too, does understanding institutional culture help stakeholders implement systemic change. Institutional and departmental policies affect everyone, yet most investigators researching undergraduate STEM teaching practice lack the tools and expertise to document institutional change. Researchers need to know more about how measurement can support the evaluation and planning of change efforts and even serve as a driver of change. Also, investigators need to know how to affect departmental, divisional, and institutional policy in order to influence teaching practices and how measurement can work systemically throughout STEM undergraduate education. Additional questions concern the most productive uses of measurement in professional development or institutional change. Finally, it is important to be aware of potential unintended consequences of measurement and whether or not measurement might result in standardized testing or changes to policies affecting academic freedom.

FROM MEASUREMENT TO IMPROVED STEM EDUCATION

A number of different stakeholders—from faculty, administrators, and policymakers to faculty developers, researchers, and evaluators—are being asked to describe STEM teaching. The resulting investigations vary from documenting classroom practice to researching effective teaching methods, but ultimately, they can all be used to improve undergraduate STEM teaching. As set forth in this report, several tools and techniques are available to assist with the effort. The overviews presented of the four basic techniques that can be used to describe STEM teaching—surveys, interviews, observations, and portfolios—provide a good starting point for those who wish to engage in such measurements.

Each of the four basic measurement techniques has its strengths and weaknesses, and the best descriptions of STEM teaching involve the use of multiple techniques. Still, there is much room for growth and development in our ability to describe that teaching. Indeed, the organizers of, and participants in, the national AAAS meeting hope that this report can also serve as a foundation for the development of improved description and measurement techniques. This important, necessary work will continue to strengthen our efforts to improve undergraduate STEM education.

WORKS CITED AND OTHER SOURCES

Introduction

Chism, N. (2007). *Peer Review of Teaching: A Sourcebook* (2nd ed.). Bolton, MA: Anker.

National Research Council. (2012). *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering.* Washington, DC: The National Academies Press.

President's Council of Advisors on Science and Technology. (2012). *Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics.* Washington DC: Executive Office of the President, President's Council of Advisors on Science and Technology.

Surveying STEM Faculty and Students

Dey, E. L. (1997). Working with low survey response rates: The efficacy of weighting adjustments. *Research in Higher Education,* 38(2), 215–227.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Fowler, F. J. (2009). *Survey Research Methods* (4th ed.). Thousand Oaks, CA: Sage.

Lattuca, L. R., Terenzini, P. T., & Volkwein, J. F. (2006). *Engineering Change: Findings from a Study of the Impact of EC2000, Final Report.* Baltimore: ABET.

McMartin, F., Iverson, E., Wolf, A., Morrill, J., Morgan, G., & Manduca, C. (2008). The use of online digital resources and educational digital libraries in higher education. *International Journal on Digital Libraries,* Special Issue on Digital Libraries and Education.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology,* 16, 263-280.

Ritter, L. A., & Sue, V. M. (2007). Using online surveys in evaluation. *New Directions for Program Evaluation,* 115, 1–64.

Salant, P., & Dillman, D. (1994). *How to Conduct Your Own Survey.* New York: John Wiley and Sons.

Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education,* 44(4), 409–432.

Sheehan, K. B. (2001). E-mail survey response rates: A review. *Journal of Computer-Mediated Communication,* 6 (2).

Faculty Surveys

Borrego, M., Cutler, S., Prince, M., Henderson, C., & Froyd, J. (2013). Fidelity of implementation of research-based instructional strategies (RBIS) in engineering science courses. *Journal of Engineering Education,* 102(3).

Borrego, M., Froyd, J., & Hall, T. S. (2010). Diffusion of engineering education innovations: A survey of awareness and adoption rates in U.S. engineering departments. *Journal of Engineering Education,* 99(3), 186–207.

Brawner, C. E., Felder, R. M., Allen, R., & Brent, R. (2002). A survey of faculty teaching practices and involvement in faculty development activities. *Journal of Engineering Education–Washington,* 91(4), 393–396.

Center for Postsecondary Research at Indiana University Bloomington. F*aculty Survey of Student Engagement 2012*. http://fsse.iub.edu/pdf/2012/FSSE12_TS.pdf.

Dancy, M., & Henderson, C. (2010). Pedagogical practices and instructional change of physics faculty. *American Journal of Physics,* 78(10), 1056–1062.

Froyd, J. E., Borrego, M., Cutler, S., Henderson, C., & Prince, M. (2013). Estimates of use of research-based instructional strategies in core electrical or computer engineering courses. *IEEE Transactions on Education,* in press.

Henderson, C., & Dancy, M. (2009). The impact of physics education research on the teaching of introductory quantitative physics in the United States. *Physical Review Special Topics: Physics Education Research,* 5(2), 020107.

Henderson, C., Dancy, M., & Niewiadomska-Bugaj, M. (2012). The use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process? *Physical Review Special Topics: Physics Education Research,* 8(2), 020104.

Hurtado, S., Eagan, K., Pryor, J. H., Whang, H., & Tran, S. (October 2012). Undergraduate Teaching Faculty: The 2010–2011 HERI Faculty Survey.

Iverson, E., Fox, S., & Manduca, C. (2008). *SERC Pop Up Survey Results.* Northfield, MN: Science Education Resource Center. http://serc.carleton.edu/files/serc/report_email_poll_2007.v2.doc.

Kuh, G. D., Chen, D., & Laird, T. N. (2007). Why teacher–scholars matter: Some insights from FSSE and NSSE. *Liberal Education,* 93(4), 40.

Macdonald, R. H., Manduca, C. A., Mogk, D. W., & Tewksbury, B. J. (2005). Teaching methods in undergraduate geoscience courses: Results of the 2004 On the Cutting Edge Survey of U.S. Faculty. *Journal of Geoscience Education,* 53(3), 237.

Marbach-Ad, G., Schaefer-Zimmer, K. L., Orgler, M., Benson, S., and Thompson, K. V. (2012). Surveying research university faculty, graduate students and undergraduates: Skills and practices important for science majors. Paper presented at the annual meeting of the American Educational Research Association (AERA), Vancouver, BC.

National Center for Education Statistics. *National Study of Postsecondary Faculty.* (Washington, DC: National Center for Education Statistics, 2005). http://nces.ed.gov/surveys/nsopf/.

Nottis, K., Feuerstein, A., Murray, J., & Adams, D. (2000). The teacher belief inventory: Measuring the theoretical and practical orientations of preservice teachers. *Education*, 121(1), 90–101.

Trigwell, K., and Prosser, M. (2004). Development and use of the approaches to teaching inventory. *Educational Psychology Review,* 16(4), 409–424.

Zieffler, A., Park, J., Garfield, J., delMas, R., & Bjornsdottir, A. (2012). The Statistics Teaching Inventory: A survey of statistics teachers' classroom practices and beliefs. *Journal of Statistics Education,* 20(1).

Student Surveys

Bell, S., Galilea, P., and Tolouei, R. (2010). Student experience of a scenario-centered curriculum. *European Journal of Engineering Education,* 35(3), 235–245.

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education,* 17(1), 48–62.

Fencl, H., & Scheel, K. (2005). Engaging students: An examination of the effects of teaching strategies on self-efficacy and course climate in a nonmajors physics course. *Journal of College Science Teaching,* 35(1), 20–24.

Giesey, J. J., Chen, Y., & Hoshower, L. B. (2004). Motivation of engineering students to participate in teaching evaluations. *Journal of Engineering Education,* 93(4), 303–312.

Greenwald, A. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist,* 52(11), 1182–1186.

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. In M. Theall, P. C. Abram, & L. A. Mets (eds.), *The Student Ratings Debate: Are They Valid? How Can We Best Use Them?* (New Directions for Institutional Research, No. 109) (pp. 9–25). San Francisco: Jossey-Bass.

Lizzio, A., Wilson, K., and Simons, R. (2002). University students' perceptions of the learning environment and academic outcomes: implications for theory and practice. *Studies in Higher Education,* 27(1), 27–52.

Lumkes, J. H. (2009–2010). Survey of three different methods of delivering engineering content in lecture. *Journal of Educational Technology Systems,* 38(3), 349–366.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. A. Smart (eds.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective* (pp. 319–383). New York: Springer.

Murray, H. G. (1987). Acquiring student feedback that improves instruction. *New Directions for Teaching and Learning* 1987(32), 85–96.

Porter, S. (2011). Do college student surveys have any validity? *Review of Higher Education,* 35(1), 45–76.

Porter, S. R., & Whitcomb, M. E. (2005). Non-response in student surveys: The role of demographics, engagement, and personality. *Research in Higher Education,* 46(2), 127–152.

Rodrigues, R. A. B., & Bond-Robinson, J. (2006). Comparing faculty and student perspectives of graduate teaching assistants' teaching. *Journal of Chemical Education,* 83(2), 305312.

Terenzini, P. T., Cabrera, A. F., Colbeck, C. L., Parente, J. M., & Bjorklund, S. A. (2001). Collaborative learning vs. lecture/discussion: Students' reported learning gains. *Journal of Engineering Education,* 90, 123–130.

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? In M. Theall, P. C. Abrami, & L. A. Mets (eds.), *The Student Ratings Debate: Are They Valid? How Can We Best Use Them?* (New Directions for Institutional Research, No. 109) (pp. 45–56). San Francisco: Jossey-Bass.

Wachtel, H. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education,* 23(2), 191–212.

Wilson, K., Lizzio, A., and Ramsden, P. (1997). The development, validation and application of the Course Experience Questionnaire. *Studies in Higher Education,* 22(1), 33–53.

Surveys Comparing Student and Faculty Responses

Balam, E. M., & Shannon, D. M. (2010). Student ratings of college teaching: A comparison of faculty and their students. *Assessment and Evaluation in Higher Education,* 35(2): 209–221.

Kagesien, O., & Engelbrecht, J. (2007). Student group presentations: A learning instrument in undergraduate mathematics for engineering students. *European Journal of Engineering Education,* 32(3): 303–314.

Lattuca, L. R., Terenzini, P. T., & Volkwein, J. F. (2006). *Engineering Change: A Study of the Impact of EC2000.* Baltimore: ABET.

Lopatto, D. (2004). Survey of Undergraduate Research Experiences (SURE): First findings. *Cell Biology Education,* 3(4), 270–277.

Interviewing STEM Faculty and Students

Angelo, T. A., & Cross, K. P. (1993). *Classroom Assessment Techniques: A Handbook for Faculty* (2nd ed.). San Francisco: Jossey-Bass.

Corbin, J., & Strauss, A. (2013). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (4th ed.). Thousand Oaks, CA: Sage.

Creswell, J. W. (2012). *Qualitative Inquiry and Research Design: Choosing Among Five Approaches.* Thousand Oaks, CA: Sage.

Kvale, S. (2007). *Doing Interviews.* Thousand Oaks, CA: Sage.

Laursen, S., Hunter, A.-B., Seymour, E., Thiry, H., & Melton, G. (2010). *Undergraduate Research in the Sciences: Engaging Students in Real Science.* San Francisco: Jossey-Bass.

Marshall, C., & Rossman, G. B. (2010). *Designing Qualitative Research* (5th ed.). Thousand Oaks, CA: Sage.

Maxwell, J. A. (2013). *Qualitative Research Design: An Interactive Approach* (3rd ed.). Thousand Oaks, CA: Sage.

Merriam, S. B. (2009). *Qualitative Research: A Guide to Design and Implementation* (2nd ed.). San Francisco: Jossey-Bass.

Miles, M. B. & Huberman, A. M. (1994). *Qualitative Data Analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Patton, M. (2001). *Qualitative Research & Evaluation Methods* (3rd ed.). Thousand Oaks, CA: Sage.

Saldana, J. (2009). *The Coding Manual for Qualitative Researchers.* Thousand Oaks, CA: Sage.

Seidman, I. (2006). *Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences* (3rd ed.). New York: Teachers College Press.

Weiss, R. S. (1995). *Learning from Strangers: The Art and Method of Qualitative Interview Studies.* New York: Free Press.

Weston, C., Gandell, t., Beauchamp, J., McAlpine, L., Wiseman, C., & Beauchamp, C. (2001). Analyzing interview data: The development and evolution of a coding system. *Qualitative Sociology,* 24(3), 381–400.

Faculty Interviews

Brown, P. L., Abell, S. K., Demir, A., & Schmidt, F. J. (2006). College science teachers' views of classroom inquiry. *Science Education,* 90(5), 784–802.

Hall, G. E., & Hord, S. M. (2001). *Implementing Change: Patterns, Principles, and Potholes.* Boston: Allyn & Bacon.

Heck, S., Stiegelbauer, S., Hall, G. E., & Loucks, S. (1981). *Measuring Innovation Configurations: Procedures and Applications.* Austin, TX: The University of Texas at Austin, Research and Development Center for Teacher Education.

Henderson, C., & Dancy, M. (2008). Physics faculty and educational researchers: Divergent expectations as barriers to the diffusion of innovations. *American Journal of Physics* (Physics Education Research Section), 76(1), 79–91.

Henderson, C., & Dancy, M. (2007). Barriers to the use of research-based instructional strategies: The dual role of individual and situational characteristics. *Physical Review Special Topics: Physics Education Research,* 3, 020102.

Henderson, C., Yerushalmi, E., Kuo, V., Heller, P., & Heller, K. (2004). Grading student problem solutions: The challenge of sending a consistent message. *American Journal of Physics,* 72 (2), 164–169.

Henderson, C., Yerushalmi, E., Kuo, V. H., Heller, K., & Heller, P. (2007). Physics faculty beliefs and values about the teaching and learning of problem solving. II. Procedures for measurement and analysis. *Physical Review Special Topics: Physics Education Research,* 3(2), 020110.

Hora, M. T. (2012). Organizational factors and instructional decision-making : A cognitive perspective. *Higher Education,* 35(2), 207–235.

Hora, M. T., & Anderson, C. (2012). Perceived norms for interactive teaching and their relationship to instructional decision-making: a mixed methods study. *Higher Education,* 64(4), 573–592.

Kember, D., & Kwan, K.-P. (2002). Lecturers' approaches to teaching and their relationship to conceptions of good teaching. In N. Hativa & P. Goodyear (eds.), *Teacher Thinking, Beliefs and Knowledge in Higher Education* (pp. 219–239). Dordrecht, The Netherlands: Kluwer.

Marbach-Ad, G., Schaefer, K. L., & Thompson, K. V. (2012). Faculty teaching philosophies, reported practices, and concerns inform the design of professional development activities of a disciplinary teaching and learning center. *Journal on Centers for Teaching and Learning,* 4, 119–137.

Martin, E., Prosser, M., Trigwell, K., Ramsden, P., & Benjamin, J. (2000). What university teachers teach and how they teach it. *Instructional Science,* 28(5), 387–412.

Samuelowicz, K., & Bain, J. D. (1992). Conceptions of teaching held by academic teachers. *Higher Education,* 24(1), 93–111.

Seymour, E., & Hewitt, N. (1997). *Talking about Leaving: Why Undergraduates Leave the Sciences.* Boulder, CO: Westview Press.

Sheppard, S., Johnson, M., & Leifer, L. (1998). A model for peer and student involvement in formative course assessment. *Journal of Engineering Education,* 87(4), 349–354.

Southerland, S. A., Gess-Newsome, J., & Johnston, A. (2003). Portraying science in the classroom: The manifestation of scientists' beliefs in classroom practice. *Journal of Research in Science Teaching,* 40(7), 669–691.

Yerushalmi, E., Cohen, E., Heller, K., Heller, P., & Henderson, C. (2010). Instructors' reasons for choosing problem features in a calculus-based introductory physics course. *Physical Review Special Topics: Physics Education Research,* 6(2), 020108.

Yerushalmi, E., Henderson, C., Heller, K., & Heller, P., & Kuo, V. (2007). Physics faculty beliefs and values about the teaching and learning of problem solving. Part I: Mapping the common core, *Physical Review Special Topics: Physics Education Research,* 3(2), 020109.

Student Interviews

Ding, L., Reay, N. W., Lee, A., & Bao, L. (2009). Are we asking the right questions? Validating clicker question sequences by student interviews. *American Journal of Physics,* 77(7), 643–650.

Seymour, E., & Hewitt, N. (1997). *Talking About Leaving: Why Undergraduates Leave the Sciences.* Boulder, CO: Westview Press.

Sheppard, S., Johnson, M., & Leifer, L. (1998). A model for peer and student involvement in formative course assessment. *Journal of Engineering Education,* 87(4), 349–354.

Thiry, H., Weston, T. J., Laursen, S. L., & Hunter, A.-B. (2012). The benefits of multi-year research experiences: Differences in novice and experienced students' reported gains from undergraduate research, *Life Sciences Education,* 11(3): 260–272.

Observing STEM Teaching

Cash, A. H., Hamre, B. K., Pianta, R. C., & Meyers, S. S. (2012). Rater calibration when observational assessment occurs at large scales: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly,* 27(3), 529–542.

Chism, N. (2007). *Peer Review of Teaching: A Sourcebook* (2nd ed.). San Francisco: Jossey-Bass.

Eison, J. (1988). Designing effective peer observation programs. *Journal of Staff, Program, and Organization Development,* 6(2), 51–59.

Guarino, C., & Stacy, B. (2012). *Review of Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.* Boulder, CO: National Educational Policy Center.

Henry, M. A., Murray, K. S., & Phillips, K. A. (2007). *Meeting the Challenge of STEM Classroom Observation in Evaluating Teacher Development Projects: A Comparison of Two Widely Used Instruments.* St. Louis: M.A. Henry Consulting.

Henry, M. A., Murray, K. S., Hogrebe, M., & Daab, M. (2009). *Quantitative Analysis of Indicators on the RTOP and ITC Observation Instruments.* St. Louis: M.A. Henry Consulting.

Millis, B. J. (1992). Conducting effective peer classroom observations. In D. H. Wulff & J. D. Nyquist (eds.), *To Improve the Academy. Vol. 11: Resources for Faculty, Instructional, and Organizational Development* (pp. 189–201). Still water, OK: New Forums Press.

Observational Instruments

Adamson, S. L., Bank, D., Burtch, M., Cox, F., III, Judson, E., Turley, J. B., Benford, R., & Lawson, A. E. (2003). Reformed undergraduate instruction and its subsequent impact on secondary school teaching practice and student achievement. *Journal of Research in Science Teaching,* 40(10), 939–957.

Amidon, E. J., & Flanders, N. A. (1967). *The Role of the Teacher in the Classroom: A Manual for Understanding and Improving Teachers' Classroom Behavior* (rev. ed.). Minneapolis: Association for Productive Teaching.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing Faculty Work.* San Francisco: Jossey-Bass.

Brent, R., & Felder, R. M. (2004). A protocol for peer review of teaching. *Annual ASEE Conference Proceedings.* Washington, DC: ASEE.

Cox, M. F., & Cordray, D. S. (2008). Assessing pedagogy in bioengineering classrooms: Quantifying elements of the "How People Learn" model using the VaNTH Observation System (VOS). *Journal of Engineering Education,* 97(4), 413–431.

Erdle, S., & Murray, H. G. (1986). Interfaculty differences in classroom teaching behaviors and their relationship to student instructional ratings, *Research in Higher Education,* 24(2), 115–127.

Ferrare, J. J., & Hora, M. T. (2012). Cultural models of teaching and learning: Challenges and opportunities for undergraduate math and science education (WCER Working Paper No. 2012-8). Madison, WI: Wisconsin Center for Education Research. http://www.wcer.wisc.edu/publications/workingPapers/papers.php.

Gilbert, M. B., & Haley, A. (2010). Faculty evaluations: An alternative approach based on classroom observations. Faculty Focus. http://www.facultyfocus.com/articles/faculty-evaluation/faculty-evaluations-an-alternative-approach-based-on-classroom-observations/.

Harris, A. H., & Cox, M. F. (2003). Developing an observation system to capture instructional differences in engineering classrooms. *Journal of Engineering Education,* 92 (4), 329–336.

Hora, M. T., & Ferrare, J. J. (2012). Instructional systems of practice: A multi-dimensional analysis of math and science undergraduate course planning and classroom teaching. *Journal of the Learning Sciences,* first published on September 24, 2012.

Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology,* 75, 138–149.

Sawada, D., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The Reformed Teaching Observation Protocol. *School Science and Mathematics,* 102(6), 245–252.

Teaching Development Evaluation Tools: Teaching Behaviors Inventory. Grand Rapids, MI: Calvin College. http://www.calvin.edu/admin/provost/teaching/instructional/tools/behaviors.htm.

Turpen, C., & Finkelstein, N. D. (2010). The construction of different classroom norms during peer instruction: Students perceive differences. *Physical Review Special Topics: Physics Education Research,* 6(2), 020123.

Turpen, C., & Finkelstein, N. D. (2009). Not all interactive engagement is the same: Variation in physics professors' implementation of peer instruction. *Physical Review Special Topics: Physics Education Research,* 5(2), 020101.

Wainwright, C., Morrell, P. D., Flick, L., & Shepige, A. (2004). Observation of reform teaching in undergraduate level mathematics and science courses. *School Science and Mathematics,* 104(7), 322–335.

Walkington, C., Arora, P., Ihorn, S., Gordon, J., Walker, M., Abraham, L., & Marder, M. (2011). *Development of the UTeach Observation Protocol: A Classroom Observation Instrument to Evaluate Mathematics and Science Teachers from the UTeach Preparation Program* (UTeach Technical Report 2011-01). Austin, TX: University of Texas, UTeach Natural Sciences.

Teaching Portfolios

Admiraal, W., Hoeksma, M., van de Kamp, M., & van Duin, G. (2011). Assessment of teacher competence using video portfolios: Reliability, construct validity, and consequential validity. *Teaching and Teacher Education,* 27(6), 1019–1028.

Brookfield, S. (1995). *Becoming a Critically Reflective Teacher.* San Francisco: Jossey-Bass.

Fitzpatrick, M. A., & Spiller, D. (2010). The teaching portfolio: Institutional imperative or teacher's personal journey? *Higher Education Research & Development,* 29(2), 167–178.

Hutchings, P., & Shulman, L. S. (1998). *The Course Portfolio: How Faculty Can Examine Their Teaching to Advance Practice and Improve Student Learning.* Herndon, VA: Stylus Publishing.

Martinez, J. F., Borko, H., Stecher, B., Luskin, R., & Kloser, M. (2012). Measuring classroom assessment practice using instructional artifacts: A validation study of the QAS Notebook. *Educational Assessment,* 17(2–3), 107–131.

Seldin, P., Miller, J. E., & Seldin, C. (2010). *The Teaching Portfolio: A Practical Guide to Improved Performance and Promotion/Tenure Decisions.* San Francisco: Jossey-Bass.

Watson, R., & Doolittle, P. (2012). *International Journal of e-Portfolio: About.* http://www.theijep.com/about.html.

Portfolio Examples

Chism, N. (2007). *Peer Review of Teaching: A Sourcebook* (2nd ed.). San Francisco: Jossey-Bass.

Linse, A., Turns, J., Yellin, J., & VanDeGrift, T. (2004). Preparing future engineering faculty: Initial outcomes of an innovative teaching portfolio program. *Proceedings of the American Society for Engineering Education Annual Conference.* Salt Lake City, UT.

McNair, L. D., & Garrison, W. (2012). Portfolios to professoriate: Helping students integrate professional identities through E-Portfolios. *Proceedings of the American Society for Engineering Education Annual Conference.* San Antonio, TX.

Ni, L., Guzdial, M., Tew, A. E., Morrison, B., & Galanos, R. (2011). Building a community to support HS CS teachers: The Disciplinary Commons for Computing Educators. *Proceedings of the 42nd ACM technical symposium on Computer science education* (pp. 553–558), Dallas, TX.

Tenenberg, J., & Fincher, S. (2007). Opening the door of the computer science classroom: The Disciplinary Commons. *Proceedings of the 2007 SIGCSE Conference* (pp. 514–518). Covington, KY. .

Combining Measurement Approaches

Borrego, M., Douglas, E. P., and Amelink, C. T. (2009). Quantitative, qualitative, and mixed research methods in engineering education. *Journal of Engineering Education* 98(1), 53–66.

Caracelli, V. J., & Greene, J. C. (1997). Crafting mixed-method evaluation designs. *New Directions for Program Evaluation,* 74, 19–30.

Creswell, J. W., Klassen, A. C., Clark, V. L. P., & Smith, K. C. (2010). *Best Practices for Mixed Methods Research in the Health Sciences* Bethesda, MD: The Office of Behavioral and Social Sciences Research (OBSSR) of the National Institutes of Health.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis,* 11(3), 255–274.

Johnson R. B., & Onwuegbuzie. A. J. (2004). Mixed method research: A research paradigm whose time has come. *American Educational Researcher,* 33(7), 14–26.

Leech, N.L. & Onwuegbuzie, A.J. (2009). A typology of mixed methods research designs. *Quality and Quantity,* 43 (2), 265–275.

Tashakkori, A., & Teddlie, C. (2002). *Handbook of Mixed Methods in Social and Behavioral Research.* Thousand Oaks, CA: Sage.

Mixed-Methods STEM Examples

Creswell, N., and Clark, V. (2007). *Designing and Conducting Mixed Methods Research*. Thousand Oaks, CA: Sage.

Ebert-May, D., Derting, T. L., Hodder, J., Momsen, J. L., Long, T. M., & Jardeleza, S. E. 2011. What we say is not what we do: Effective evaluation of faculty development programs. *BioScience* 6(17), 550–558.

Hora, M. T. & Anderson, C. D. (2012). Perceived norms for interactive teaching and their relationship to instructional decision-making: A mixed methods study. *Higher Education,* 64 (4), 573–592.

Lattuca, L. R., Terenzini, P. T., & Volkwein, J. F. (2006). *Engineering Change: Findings from a Study of the Impact of EC2000, Final Report.* Baltimore, MD: ABET.

Thiry, H., Weston, T. J., Laursen, S. L., and Hunter, A-B. (2012). The Benefits of Multi-Year Research Experiences: Differences in Novice and Experienced Students’ Reported Gains from Undergraduate Research. *Life Sciences Education,* 11(3), 260–272.

Participant List

LAST NAME	FIRST NAME	AFFILIATION
Borko	Hilda	Stanford University
Borrego	Maura	National Science Foundation
Boylan	Myles	National Science Foundation
Braddy	Linda	Mathematical Association of America
Breslow	Lori	Teaching and Learning Laboratory & MIT Sloan School of Management
Bretz	Stacey Lowery	Miami University, Oxford
Cole	Renee	University of Iowa
Connolly	Mark	University of Wisconsin, Madison
Dancy	Melissa	University of Colorado
Davis	Kirsten	Boise State University
Domina	Thurston	UC Irvine
Ebert-May	Diane	Michigan State University
Fairweather	James	Michigan State University
Ferrare	Joseph	University of Wisconsin-Madison
Finelli	Cynthia	University of Michigan
Franklin	Scott	Rochester Institute of Technology
Froyd	Jeffrey	Texas A&M University
George	Yolanda S.	AAAS and NSF PI
Gomez	Luanna	SUNY College of Buffalo
Grissom	Scott	Grand Valley State University
Grunert	Megan	Western Michigan University
Henderson	Charles	Western Michigan University
Hora	Matthew	Wisconsin Center for Educational Research
Hsieh	Sheng-Jen	Texas A&M University
Iverson	Ellen	Carleton College
Karlin	Jennifer	South Dakota School of Mines and Technology
Krause	Stephen	Arizona State University
Lattuca	Lisa	University of Michigan
Laursen	Sandra	University of Colorado Boulder

LAST NAME	FIRST NAME	AFFILIATION
Lea	Peter	National Science Foundation
Lee	Kevin	University of Nebraska
Light	Gregory	Northwestern University
Litzinger	Thomas	Penn State University
MacIsaac	Daniel	Buffalo
Malcom	Shirley	AAAS
Marbach-Ad	Gili	University of Maryland
Marrongelle	Karen	Oregon University System & Portland State University
McGuire	Sharon	Boise State University
McKagan	Sarah “Sam”	PER User's Guide
Meltzer	David	Arizona State University
Reese	Mike	Johns Hopkins Center for Educational Resources
Russell	Arlene	University of California, Los Angeles
Sankar	Chetan	Auburn University
Sayre	Eleanor	Kansas State University
Simon	Beth	University of California, San Diego
Slakey	Linda	University of Massachusetts Amherst
Smith	Diane	Smith Communications
Smith	Toby	Association of American Universities
Smith	Karl	University of Minnesota
Speer	Natasha	University of Maine
Turns	Jennifer	University of Washington
Van Dyke	Ray	Virginia Tech
Van Staaden	Moira	Bowling Green State University
Varma-Nelson	Pratibha	Indiana University, Purdue University, Indianapolis
Vigeant	Margot	Bucknell University
White	Susan	American Institute of Physics
White	Gary	National Science Foundation

This report can be found at ccliconference.org/measuring-teaching-practices/.

This report is based upon work supported by the National Science Foundation under grant No. DUE 1252972.

ADVANCING SCIENCE, SERVING SOCIETY