A PROTOCOL FOR PEER REVIEW OF TEACHING

Rebecca Brent/Richard M. Felder
Education Designs, Inc./North Carolina State University

Abstract

A peer review protocol that serves both formative and summative functions has been implemented at North Carolina State University. For summative evaluation, two or more reviewers use standardized checklists to independently rate instructional materials (syllabus, learning objectives, assignments, tests, and other items) and at least two class observations, and then reconcile their ratings. For formative evaluation, only one rater completes the forms and the results are shared only with the faculty member being rated rather than being used as part of his/her overall teaching performance evaluation. Pilot test results of the summative protocol show a high level of inter-rater reliability. This paper presents a brief overview of the reasons for including peer review in teaching performance evaluation and the problems with the way it has traditionally been done, describes and discusses the protocol, summarizes the pilot test results, and demonstrates how the use of the protocol can minimize or eliminate many common concerns about peer review of teaching.

Introduction

Mounting pressures on engineering schools to improve the quality of their instructional programs have been coming from industry, legislatures, governing boards, and ABET. An added impetus for improving engineering instruction is a growing competition for a shrinking pool of qualified students. If enrollment falls below a critical mass, the loss in revenues from tuition and other funds tied to enrollment could place many engineering schools in serious economic jeopardy.

A prerequisite to improving teaching is having an effective way to evaluate it. Standard references on the subject all agree that the best way to get a valid summative evaluation of teaching is to base it on a portfolio containing assessment data from multiple sources—ratings from students, peers, and administrators, self-ratings, and learning outcomes—that reflect on every aspect of teaching including course design, classroom instruction, assessment of learning, advising, and mentoring.[1–4] A schematic diagram of a comprehensive evaluation system that incorporates these elements is shown in Figure 1.[5] This paper deals with the peer review component of the system. Other references may be consulted for information regarding student ratings of teaching[6–9] and teaching portfolios.[4,10–12]

Why, How, and How Not to Do Peer Review

For the last half century, the standard way to evaluate teaching has been to collect course-end student rating forms and compile the results. While student ratings have considerable validity,[6] they also have limitations. Among other things, students are not qualified to evaluate

an instructor's understanding of the course subject, the currency and accuracy of the course content, the appropriateness of the level of difficulty of the course and of the teaching and assessment methods used in its delivery, and whether the course content and learning objectives are consistent with the course's intended role in the program curriculum (for example, as prerequisite to other courses). Only faculty colleagues are in a position to make these judgments. Moreover, students have limited ability to provide individual formative feedback to their instructors; only colleagues can freely provide such feedback. Recognizing these limitations of student ratings, growing numbers of institutions and departments have begun to include peer review in their faculty performance evaluations.

Peer review is not without its own problems, however. In the customary approach to it, a faculty member observes a class session and jots down notes about whatever happens to catch his or her attention. This approach has several flaws.

- One class may not provide a representative picture of someone's teaching, and the presence of an observer in the class could increase the likelihood of an atypical performance by the instructor (possibly better and possibly worse).
- Different observers are likely to focus on different things and interpret what they see in different ways, so that same class session could get a good report from one observer and a poor one from another.
- Simply watching someone teach a single class provides little information about the currency or accuracy of the course content, the appropriateness of the assignments and tests, and whether or not the students are being equipped with the knowledge and skills needed to move on in the curriculum and to satisfy program accreditation requirements.

Other common concerns about peer review include wide variations in faculty opinions about what constitutes good teaching, controversies over who is qualified to be a peer reviewer, the possibility of personal biases affecting ratings, and excessive time demands on the reviewers.

Peer review procedures that address these concerns have been developed by professional educators.[1,2] One such procedure recently implemented in the N.C. State University Chemical Engineering Department involves evaluation of instructional materials and at least two class observations by two or more independent reviewers, who subsequently reconcile their ratings.

Design and Pilot Test of the N.C. State Peer Review Procedure

The department faculty committee assigned to formulate a peer review procedure began by developing checklist rating forms for classroom observations and course materials, with the checklist items being selected from lists of well-established characteristics of effective teaching.[2] The forms are shown in Tables 1 and 2. The following strategy was then devised:

1. A committee of peer reviewers was formed. Two reviewers ("raters") were assigned to each faculty member ("instructor") to be reviewed.

2. The raters met with the instructor to discuss the instructor's goals for the course, arrange two class observation dates, specify the course materials to be collected (syllabi, course

learning objectives, policies and procedures, handouts, representative lecture notes, assignments and tests, and grade distributions), and go over the two rating forms.

3. The raters observed the first class and independently filled out class observation rating forms (Table 1). Immediately afterward, they met to reconcile their ratings of each item on the form and entered the reconciled ratings on a consensus form. If they could not agree on how to rate an item, their ratings were averaged and rounded up to the next highest integer. The same procedure was subsequently carried out for the second class observation.

4. At the end of the semester, the raters collected the specified course materials, independently filled out course material rating forms (Table 2), and reconciled them to arrive at a consensus rating. They then drafted a report summarizing their findings and gave it to the review committee chair.

5. The chair drafted a letter that summarized and discussed the instructor's strengths and areas that needed improvement. The letter was first given to the raters to be reviewed for accuracy and revised if necessary, and copies of the revised letter were sent to the department head and the instructor. The instructor was welcome to submit a dissenting report if he/she disagreed with any of the findings, but none of the instructors reviewed in the pilot test saw a need to do so.

6. All instructors who were reviewed were invited to meet with their raters and the review committee chair to discuss the evaluation and formulate measures they might take to improve their teaching.

Each rater spent about seven hours on this process: two meeting with the instructor, two observing classes, and three reviewing course materials, reconciling forms, and preparing reports.

In a test of the class observation rating form, one of the task force members observed a class taught by a senior faculty member known to be an outstanding lecturer and gave it the top rating of 5 in eight of the ten categories and 4 in the other two, for an average of 4.8. The full procedure was then implemented for three assistant professors. The average consensus ratings in the six class observations varied from a high of 4.0 to a low of 2.9. (Average ratings were calculated only for reliability analysis; they are not normally included in the peer review summary reports.)

There was a gratifying level of inter-rater consistency in ratings of both class observations and course materials. The average ratings for the same instructor differed from one rater to another by no more than half a unit. Out of 60 item ratings submitted by individual raters for the first class observations (10 items for each of three professors, with each item being rated by two evaluators), the two raters agreed 25 times, differed by one unit 28 times, and differed by two units seven times. The between-rater differences for the second set of class observations were even lower than those for the first set. The agreement for the first set would undoubtedly have been even greater if the raters had observed one or two practice sessions and discussed how to rate each item before progressing to the actual observations. In 30 ratings of individual items

related to course materials (Table 2), the two raters agreed 23 times and differed by only one unit 7 times. No item ratings differed by more than one unit.

The between-session differences in ratings for each instructor were quite small. The overall consensus ratings differed from one session to another by 0.4 units, 0.2 units, and 0.4 units for the three faculty members reviewed, probably reflecting normal variations in teaching effectiveness from day to day. The consensus ratings for specific items in the two observed classes were identical 16 times, differed by one unit 13 times, and differed by two units once. Besides corresponding closely to each other, the class observation ratings for each instructor were consistent with the student evaluations collected at the end of the semester. The committee concluded that the class sessions they observed were truly representative of the instruction delivered throughout the semester.

After reviewing these results, the department faculty voted to adopt the procedure and it has been used successfully for three years. The high inter-rater reliability observed in the pilot test has been consistently maintained, and no instructors have filed dissenting reports.

Recommended Peer Review Protocol

Peer review has two possible functions: summative (to provide data to be used in personnel decisions or award nominations) and formative (to improve teaching). Based on our review of the peer review literature and our experience with the procedure described above, we recommend the following protocol for both summative and formative peer review.

1. *Design class observation and course material rating forms using the formats shown in Tables 1 and 2*. Select items that have been shown to correlate with effective teaching from lists given in References 1 and 2. Obtain consensus approval of the department faculty for the items included in the final forms.

2. *At the beginning of the fall semester or quarter, form a departmental peer review committee that will function for the next academic year*. The committee should consist of a chair within the department who oversees the peer review process and a cadre of faculty raters who may come from within the department or from other departments in related disciplines. Guidelines for selecting raters are suggested in the next section.

3. *Early in the fall, provide a 1–2 hour training session to the raters*. The trainer (an experienced rater from previous years or a faculty development consultant) should present an illustrative set of course materials and one or two mini-lectures or videotaped excerpts of real lectures, and the participants should complete the rating forms and discuss their reasons for assigning the ratings they did. Presenting two mini-lectures that vary in quality makes the experience more instructive.

4. *Summative review*. For faculty members being considered for reappointment, promotion, or tenure or undergoing post-tenure review, the summative procedure described previously should be used (preliminary meeting to go over the procedures, at least two raters and two class observations for each faculty member reviewed, reconciliation of independently completed checklists, final meeting to discuss the results and identify steps for improvement if necessary). The results should be included in a portfolio along with a

summary of student ratings for the preceding three years and other items specified in Figure 1.

Formative review. A modification of the summative procedure should be implemented for formative peer review. The preliminary interview, two classroom observations, and course material review may be performed by only one rater, who completes the rating sheets as above but shares and discusses the results only with the instructor. Such constructive feedback provided to faculty members in their first few years should increase the chances of their meeting or exceeding departmental standards for teaching in subsequent summative reviews.

Resolving Concerns about Peer Review

In the introductory section, we raised several common concerns about peer review. In what follows, we suggest how these concerns are addressed by the protocol just described.

- *Concern: There is no universal agreement among faculty members about what constitutes good teaching, and the chances of getting agreement in most departments are slim.*

Extensive research has demonstrated that certain characteristics of instruction correlate significantly with students' motivation to learn, learning outcomes, and satisfaction with their education. The suggested checklist rating items in References 1 and 2 are based on those research findings. The references list far more items than would be practical to include in rating forms, and even the most disputatious department faculty should be able to reach consensus on a subset of them.

- *Concern: Many faculty members are not qualified to review someone else's teaching, and those who* are *qualified may be in short supply and overworked*.

We are not aware of research-based eligibility criteria for being a peer reviewer, but certain criteria are suggested by experience and common sense. We propose that reviewers (both summative and formative) should be:

(1) *tenured faculty or faculty or non-tenure-track faculty with primarily teaching and advising responsibilities.* Untenured assistant professors should not have to rate colleagues who may later be in a position of evaluating their candidacy for tenure. (Another way to avoid this situation is to use raters from different departments, subject to the knowledgeability condition of Criterion 3.)

(2) *experienced.* Faculty with less than three years of teaching experience should generally not be called upon to rate someone else's teaching.

(3) *knowledgeable*. Raters should understand the criteria to be used in the peer review process, and to a reasonable extent, the broad discipline of the course being reviewed if not the specific course content. Asking a mechanical engineer to review instruction in certain civil or chemical engineering courses, for example, would be generally acceptable, but asking a medieval historian to review instruction in an engineering course would not. As for understanding the rating criteria, the suggested preliminary rater training should be adequate to provide it.

(4) *competent*. While it is not necessary to use only winners of outstanding teacher awards as peer reviewers (there may not be enough of them to meet departmental needs), using poor teachers to evaluate their colleagues' teaching would clearly be a bad idea.

(5) *flexible*. There is no single correct way to teach. Instructors whose styles vary from traditional lecture-based instruction to full-bore active, cooperative, problem-based learning may all be excellent teachers. Faculty with a rigidly narrow view of what constitutes acceptable teaching should not be peer reviewers.

(6) *unbiased.* Individuals who have strong personal or philosophical differences with a faculty colleague should not be asked to serve as peer reviewers for that colleague. If they are asked to do so, they have an ethical responsibility to decline.

Many engineering faculty members meet these criteria, so at most institutions it should not be too difficult to find enough qualified raters to cover all scheduled summative peer reviews in a given year.

- *Concern: Peer review that goes beyond a single class observation imposes too much of a time burden on faculty members*.

The total time required for a summative review using the suggested protocol is about seven hours per rater. This obligation is equivalent to serving on a committee that meets for two hours every other week in a semester, a level of commitment routinely required of faculty members. Moreover, in the proposed system faculty members would generally undergo summative reviews no more than once in three years, so that most faculty members would only be required to serve as reviewers every two or three years. The time burden of peer review is thus considerably less than that imposed by typical committee service.

- *Concern: Two observed classes may not be representative of the entire course.*
- *Concern: The presence of an observer in a class necessarily affects the instructor and possibly also the students, so that any observed class cannot be representative of the course (the "observer effect")*.
- *Concern: Raters may be biased against the instructor and unable to maintain objectivity in their reviews.*

These are legitimate concerns. Since the protocol uses multiple raters and observations and the observations are only one component of the review process, it is unlikely but possible for a good teacher to get a poor evaluation or vice versa because of atypical class sessions. Similarly, even though the suggested reviewer selection process should screen out bias, it is possible—albeit highly improbable—for two raters to share the same unacknowledged bias toward the instructor they are evaluating.

These concerns simply reinforce the idea that peer review should be only one component of the system used to evaluate faculty teaching performance. If multiple sources are used in the review—say, student ratings and peer ratings—and they converge to the same conclusion about an instructor's teaching performance, the chances are great that the common conclusion is correct. On the other hand, if the two sets of ratings yield considerably different conclusions,

then either something is wrong with at least one set or the instructor's teaching in the reviewed course was not truly representative of his/her usual teaching. At that point, further investigation could and should be undertaken.

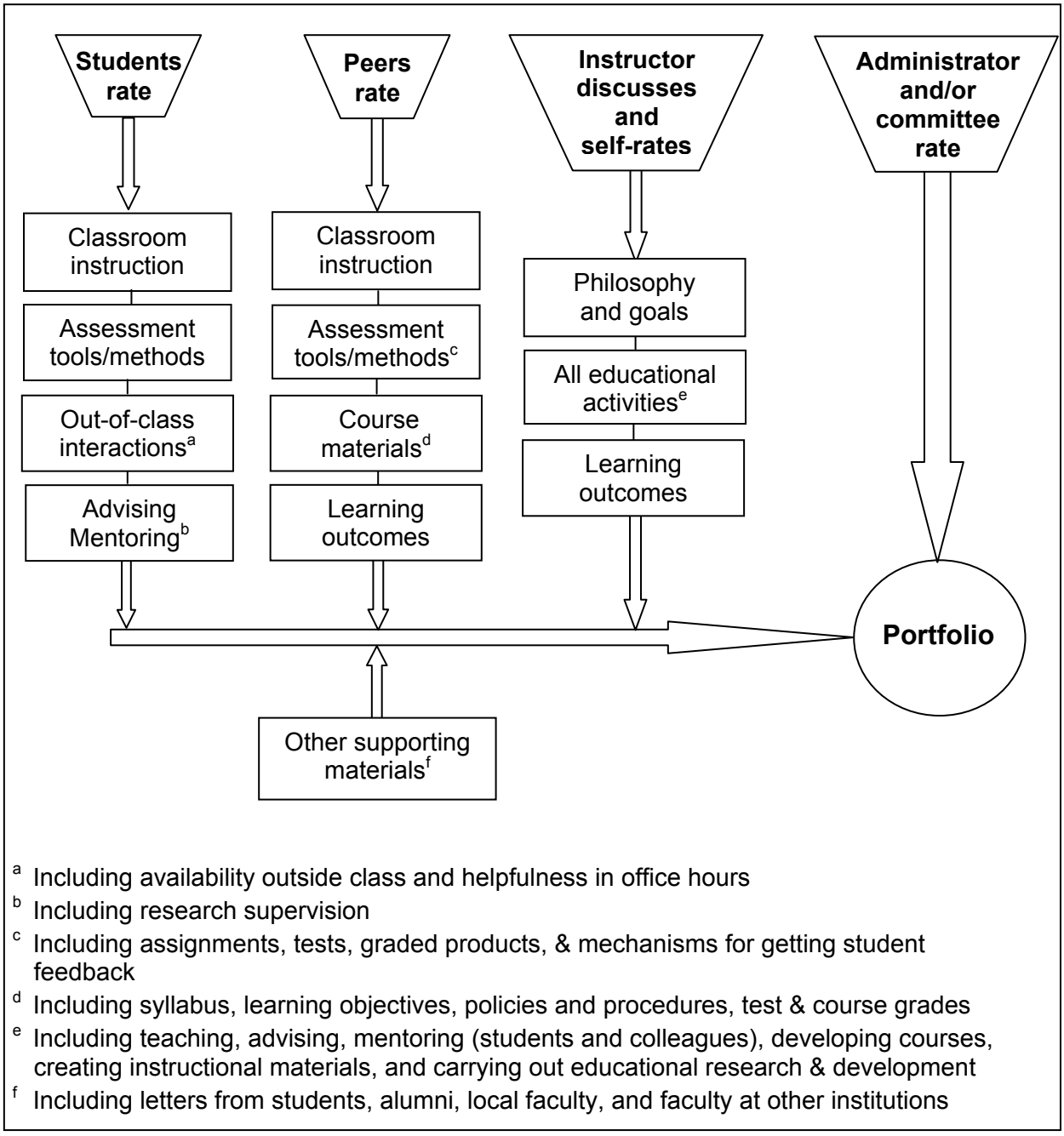
One way to increase the reliability of multiple-source evaluations is to make sure that there is some overlap in the information the sources provide. For example, if the class observation rating sheet includes items related to preparedness for lectures, clarity of explanations, and respect for students, then the evaluation forms completed by the students should ask for ratings of the same attributes.

Summary

A protocol for summative peer review of teaching has been outlined and tested. It is based on research on teaching effectiveness, consistent with accepted best practices in evaluation, and reliable, and does not impose undue time demands on the faculty. If it is part of a multiple-source assessment system of the type illustrated in Figure 1, it should provide an evaluation of teaching performance with a validity acceptable by any reasonable standard, but more extensive testing will be required to confirm that hypothesis. The protocol also provides a good basis for formative evaluation, which if implemented in the first few years of a faculty member's career should significantly increase the likelihood that a subsequent summative review will be favorable.

References

1. N. Van Note Chism, *Peer Review of Teaching*, Bolton, MA, Anker Publishing, 1999.
2. M. Weimer, J.L. Parrett, and M. Kerns, *How am I Teaching?* Madison, WI, Magna Publications, 1988.
3. D.P. Hoyt and W.H. Pallett, "Appraising Teaching Effectiveness: Beyond Student Ratings," IDEA Paper No. 36, Kansas State University Center for Faculty Evaluation and Development, <*www.idea.ksu.edu*>, November 1999.
4. National Research Council, *Evaluating and Improving Undergraduate Teaching in Science, Technology, Engineering, and Mathematics*, Washington, DC, National Academies Press, 2003.
5. R.M. Felder and R. Brent, "How to Evaluate Teaching," *Chem. Engr. Education,* in press (2004).
6. W.E. Cashin, "Student Ratings of Teaching: The Research Revisited," IDEA Paper No. 32, Kansas State University Center for Faculty Evaluation and Development, <*www.idea.ksu.edu*>, September 1995.
7. R.M. Felder, "What Do They Know, Anyway?" *Chem. Engr. Education, 26*(3), 134-135 (1992), <*http://www.ncsu.edu/effective_teaching/Columns/Eval.html*>.
8. R.M. Felder, "What Do They Know, Anyway? 2. Making Evaluations Effective," *Chem. Engr. Education, 27*(1), 28-29 (1993), <*http://www.ncsu.edu/effective_teaching/Columns/Eval2.html*>.
9. W.J. McKeachie, "Student Ratings: The Validity of Use," *American Psychologist, 52*(11), 1218–1225 (1997).
10. R.M. Felder, "If You've Got It, Flaunt It: Uses and Abuses of Teaching Portfolios," *Chem. Engr. Education, 30*(3), 188-189 (1996), <*http://www.ncsu.edu/effective_teaching/Columns/Portfolios.html*>.
11. P. Seldin, *The Teaching Portfolio: A Practical Guide to Improved Performance and Promotion/Tenure Decisions,* 2nd Edition, Bolton, MA, Anker Publishing Co., 1997.
12. R. Edgerton, P. Hutchings, and K. Quinlan, *The Teaching Portfolio: Capturing the Scholarship in Teaching,*Washington, DC, American Association for Higher Education, 1991.



[a] Including availability outside class and helpfulness in office hours
[b] Including research supervision
[c] Including assignments, tests, graded products, & mechanisms for getting student feedback
[d] Including syllabus, learning objectives, policies and procedures, test & course grades
[e] Including teaching, advising, mentoring (students and colleagues), developing courses, creating instructional materials, and carrying out educational research & development
[f] Including letters from students, alumni, local faculty, and faculty at other institutions

Figure 1. Comprehensive Evaluation of Teaching Performance

Table 1
Class Observation Checklist

Course: _____ Instructor: _____ Date: _____

Circle your responses to each of the questions and then add comments below the table.

	Exceeds expectations in all respects	Meets expectations in all respects	Meets expectations in most respects	Meets expectations in some respects	Meets expectations in few or no respects
1 – was well prepared for class	5	4	3	2	1
2 – was knowledgeable about the subject matter	5	4	3	2	1
3 – was enthusiastic about the subject matter	5	4	3	2	1
4 – spoke clearly, audibly, and confidently	5	4	3	2	1
5 – used a variety of relevant illustrations/examples	5	4	3	2	1
6 – made effective use of the board and/or visual aids	5	4	3	2	1
7 – asked stimulating and challenging questions	5	4	3	2	1
8 – effectively held class's attention	5	4	3	2	1
9 – achieved active student involvement	5	4	3	2	1
10 – treated students with respect	5	4	3	2	1

What worked well in the class? (Continue on back if necessary)

What could have been improved? (Continue on back if necessary)

Rater(s)_____

Table 2
Course Material Checklist

Course: _____ Instructor: _____ Date: _____

Circle your responses to each of the questions and then add comments below the table.

	Exceeds expectations in all respects	Meets expectations in all respects	Meets expectations in most respects	Meets expectations in some respects	Meets expectations in few or no respects
1. Course content includes the appropriate topics	5	4	3	2	1
2. Course content reflects the current state of the field	5	4	3	2	1
3. Course learning objectives are clear and appropriate	5	4	3	2	1
4. Course policies and rules are clear and appropriate	5	4	3	2	1
5. Lecture notes are well organized and clearly written	5	4	3	2	1
6. Supplementary handouts and web pages are well organized and clearly written	5	4	3	2	1
7. Assignments are consistent with objectives and appropriately challenging	5	4	3	2	1
8. Tests are consistent with learning objectives and appropriately challenging	5	4	3	2	1
9. Tests are clearly written and reasonable in length	5	4	3	2	1
10. Student products demonstrate satisfaction of learning objectives	5	4	3	2	1

What are the strengths of the course materials? (Continue on back if necessary)

What could have been improved? (Continue on back if necessary)

Rater(s) _____

REBECCA BRENT, Ed.D., is President of Education Designs, Inc., with interests that include faculty development in the sciences and engineering, support programs for new faculty members, preparation of alternative licensure teachers, and applications of technology in the K-12 classroom. She was formerly a professor of education at East Carolina University. She is co-director of the ASEE National Effective Teaching Institute.

RICHARD M. FELDER, Ph.D. (<*www.ncsu.edu/effective_teaching*>) is Hoechst Celanese Professor Emeritus of Chemical Engineering at North Carolina State University. He is co-author of *Elementary Principles of Chemical Processes* (Wiley, 2000), author or co-author of over 200 papers on engineering education and chemical process engineering, a Fellow Member of the ASEE, and co-director of the ASEE National Effective Teaching Institute.