Assessment of teaching effectiveness: Lack of alignment between instructors, institutions, and research recommendations

Charles Henderson,[1] Chandra Turpen,[2*] Melissa Dancy,[3] and Tricia Chapman[4*]

[1]*Department of Physics and Mallinson Institute for Science Education, Western Michigan University, Kalamazoo, Michigan, 19050, USA*

[2]*Department of Physics, University of Maryland, College Park, Maryland, 20742, USA*

[3]*Department of Physics, University of Colorado, Boulder, Colorado 80309, USA*

[4]*Department of Education, Nazareth College, Rochester, New York, 14618, USA*

(Received 23 September 2013; published 19 February 2014)

Ideally, instructors and their institutions would have a shared set of metrics by which they determine teaching effectiveness. And, ideally, these metrics would overlap with research findings on measuring teaching effectiveness. Unfortunately, the current situation at most institutions is far from this ideal. As part of a larger interview study, 72 physics instructors were asked to describe how they and their institutions assess teaching effectiveness. Results suggest that institutions typically base most or all of their assessment of teaching effectiveness on student evaluations of teaching. Instructors, on the other hand, base most or all of their assessment of teaching effectiveness on student exam performance and nonsystematic formative assessments. Few institutions and instructors use assessment practices suggested by the research literature. In general, instructors are much more positive about the methods they use to evaluate their teaching than the methods their institutions use to evaluate their teaching. Both instructors and institutions could benefit from broadening the assessment sources they use to evaluate teaching effectiveness through increased use of standardized measures based on student learning and greater reliance on systematic formative assessment.

DOI: 10.1103/PhysRevSTPER.10.010106 PACS numbers: 01.40.Fk

I. INTRODUCTION

Strong assessment methods are broadly recognized as integral to effective instruction [1]. Three current trends in higher education are encouraging institutions to pay more attention to assessing teaching effectiveness: increasing accountability pressures, increased competition for students, and increased use of evidence-based management techniques [2]. Thus, since there is significant movement towards both increasing and improving assessment practices in higher education, now is an ideal time to influence these changes in a positive direction.

With respect to educational change, it has been argued that changing assessment methods will be instrumental in promoting more innovative, research-based teaching practices [3]. In addition to serving as potential levers for instigating change, the assessment methods used by institutions and instructors to determine teaching effectiveness will influence instructors' determinations about whether their attempts to use research-based instructional strategies are working. In this way, assessment practices influence judgments about the relative advantage and compatibility [4] of educational innovations and thus the continuation or discontinuation of new instructional strategies.

At the same time as assessment is seen as being increasingly important in higher education, there is very little research available about what types of assessment are actually used by instructors and their institutions and the alignment of assessment practices between these groups. Without knowing about the current status of assessment in higher education it will be difficult or impossible to know how to productively move forward.

II. RESEARCH QUESTIONS

This paper will address four core research questions related to physics instructors' perceptions about the techniques that they and their institutions use to evaluate teaching effectiveness.

(1) What assessment techniques do physics instructors report using to determine whether they are teaching effectively?
   (a) How do instructors use each technique?
   (b) What do instructors perceive as the strengths and weaknesses of each technique?

(2) What assessment techniques do physics instructors report their institutions using to determine whether instructors are teaching effectively?
   (a) How do institutions use each technique?

[*]Formerly at Western Michigan University, Kalamazoo, Michigan 49008, USA.

(b) What do instructors perceive as the strengths and weaknesses of each technique?

(3) To what extent do the assessment techniques that physics instructors report using align with the assessment techniques that they report their institutions using?

(4) To what extent do the assessment techniques that physics instructors and their institutions use align with the research literature on assessment of teaching effectiveness?

III. ORGANIZATION OF THE PAPER

The paper will begin with a literature review of the role of assessment in higher education. We first discuss the role of assessment in general in shaping teaching decisions. Then, we discuss specific assessment techniques identified in a literature search. This is followed by a description of the research methods that we used to conduct and analyze the 72 interviews with physics instructors. The results will be presented around eight techniques that can be used to assess teaching effectiveness. Finally, we end the paper with conclusions and implications for practice.

IV. ASSESSMENT OF TEACHING EFFECTIVENESS—LITERATURE REVIEW

A. Why worry about assessment of teaching effectiveness?

Assessment methods are very important and changing assessment methods are seen as a key for promoting more innovative, research-based teaching practices [3]. For educational quality to improve at higher education institutions, "instructors should experiment with new teaching and learning methods, and evaluate those methods' results" [5] (p. 41). In addition, the culture of universities must change to value such approaches to teaching: "We must identify the full range of teaching skills and strategies that might be used, describe best practices in the evaluation of teaching effectiveness ..... and define how these might be used and prioritized during the promotion process" [6] (p. 153).

Without valid methods of assessing teaching effectiveness, teaching methods are more likely to be based on intuition and/or tradition, which may not result in the desired outcomes. After all, if neither instructors nor their institutions objectively knows if, when, or to what extent students are meeting learning goals then intuitive but ineffective practice is more likely to continue and may even be inadvertently rewarded. Likewise, when efforts are undertaken to improve teaching without valid assessments, the success of the reform may not be noted or rewarded; assessments communicate what is valued in the system [7]. This could be especially detrimental to early reform efforts when the time commitment to the reform is large or other implementation difficulties are significant. Without feedback about what is working well and what is not working well, instructors cannot make informed decisions about whether to continue with a new instructional strategy and how to improve their implementation of the strategy.

B. Assessment methods from the literature

In this section we will provide an overview of assessment methods as discussed in the research literature. The assessment methods are organized into eight categories that will be discussed in the results section as well as an additional "other" category for assessment methods that were not found in the analysis of instructor interviews. The categories (see Table I) were developed iteratively by the entire research team based on both the interview data and our understanding of the research literature. Thus, assessment methods in the other category will be discussed in the literature review section and in the conclusions or implications, but not in the results section. It is also important to note that, for the purposes of this paper, we will confine our discussions of teaching effectiveness to course-level assessment. This is one piece of the larger conversation on assessment. A common unit of analysis for assessment that will not be discussed in this paper is the program level.

It was necessary for us to develop a set of categories of assessment methods because it is uncommon in the literature for authors to use categorization schemes that combine assessment techniques used by instructors and institutions. One exception is Berk [8], who uses the literature on the measurement of teaching effectiveness to identify and describe 12 strategies for measuring teaching effectiveness: student ratings, peer ratings, self-evaluation, videos, student interviews, alumni ratings, employer ratings, administrator ratings, teaching scholarship, teaching awards, learning outcomes, and teaching portfolios. Note, however, that Berk does not include informal or formative measures of teaching effectiveness that have been identified by others as important (see e.g., Refs. [5,9,10]). See Table I for a comparison of our categories to Berk's categories.

In the following sections we provide an overview of each of the assessment categories.

1. Student evaluations of teaching

Student evaluations of teaching (SETs) are the most common method institutions use to assess teaching effectiveness [5,8,11–13]. As many have noted, there is a large body of research- and opinion-based literature related to SETs and there is considerable disagreement about the value of these evaluations (see e.g., Ref [14]). The purpose of this paper is not to enter this debate. Rather, the purpose of this paper is to identify possible disconnects that exist between the methods that institutions use to assess teaching and the methods used by instructors.

Thus, related to the use of SETs, we will simply point out that different researchers have come to different

TABLE I. Categories of assessment information used in the analysis compared to categories developed by Berk [8].

Assessment categories for analysis	Definition	Match with Berk (2005) categories
Student evaluations of teaching	All structured collection of student evaluative feedback about a course that occurs within the term that the course is conducted.	Student ratings, student interviews, alumni ratings
Peer observations of teaching	Having peer instructors or university administrators observe an instructor's course and provide feedback (often through written reports and sometimes verbal feedback in face-to-face meetings).	Peer ratings, videos
Teaching portfolios	Having instructors self-report how they teach, sometimes providing references to the research literature about the evidence for the success of the instructional methods they are using.	Peer ratings, self-evaluation, administrator ratings, teaching portfolios
Research-based assessments	Typically involving pre-post or post-only testing with the use of multiple-choice conceptual inventories developed through research-based techniques (i.e. developed though an iterative process of reliability and validity checks, peerreviewed, etc.).	Learning outcomes
Exams, quizzes, or homework	Using students' performance on regular course exams, quizzes, or homework and taking this performance as an indicator of the success of instruction.	Learning outcomes
Systematic formative assessment	Gathering a sampling of students' performance on low stakes and often in-class tasks as an indicator of teaching effectiveness (e.g., having students submit votes or walking around the room and systematically observing the solutions of multiple groups).	N/A
Informal formative assessment	All other forms of formative assessment, such as students' verbal comments in class or in an instructor's office, the look of confusion in students' eyes, whether or not students are awake, or whether or not students are asking questions.	N/A
Postcourse feedback from students	Spontaneous or solicited comments from student(s) semesters or years after the course has ended.	N/A
Other		Teaching awards (awards that the instructor has won) Teaching scholarship (the extent to which an instructor presents or publishes in the scholarship of teaching and learning) Employer ratings (for program level assessment)

conclusions. Some argue that SETs are valid since they are correlated with student achievement [13–15]. Others argue that the SETs are merely popularity contests and that the use of SETs is a barrier to more effective teaching, since instructors are hesitant to be more rigorous due to fears of lower SETs [11,16–18]. However, there is almost universal agreement that assessment of something as complicated, nuanced, and important as teaching effectiveness should be made using multiple methods (see e.g. Refs [3,6–8,11,14,19–21]).

The literature suggests that, although widely used by institutions to judge teaching effectiveness, instructors tend not to value SETs as providing useful information and typically do not use SETs to improve instruction [13,15,22]. Similarly, many department chairs do not believe that SETs are a good measure of teaching effectiveness [22]. It is sometimes thought that a fear of decreased SETs is one barrier that keeps some instructors from using innovative instructional styles [16,18,23].

One of the criticisms of typical SETs is that many questions ask students about their preference for particular aspects of the class rather than about what they learned from these aspects [24]. Based on this criticism, Seymour and colleagues developed an alternative instrument, student assessment of their learning gains (SALG), that asks students to rate how much they have gained from the class

[24]. Instructors can use SALG for their own assessment even if this information is not valued by their institutions.

2. Peer observations of teaching

Peer observations are used for both instructor self-development (formative) and institutional evaluation (summative) of teaching [8,19,22]. However, Berk [8] cautions that it may be more difficult to get instructor buy-in for summative use. In addition to being useful in evaluating teaching effectiveness, one of the benefits of peer assessment is that it helps instructors within a department create a shared understanding of good teaching [19,22]. This development of a shared understanding among instructors in a department regarding teaching through observations and discussion is thought to promote more effective teaching throughout the department [22].

Chism [19] criticizes typical summative use of peer observations by institutions where "uninformed peers make brief visits and report from the perspective of their own biases" (p. 75). She offers seven guidelines (p. 77) for useful peer observations: (1) observers should receive training on classroom observations, (2) a single classroom observation is not sufficient, (3) preobservation information about the course and goals for the class session are necessary to provide context to the observer, (4) the approach used by an observer should help to focus the observations (e.g., via a checklist or rating form), (5) the observer should not disrupt the class operations, (6) observations should last for the entire time of a typical 50-minute class session or at least one hour of a longer class, and (7) observation reports should be completed as soon after the observation as possible.

3. Teaching portfolios

Similar to peer observations, teaching portfolios are used for both instructor self-development and institutional evaluation of teaching [19,25]. A teaching portfolio consists of documents and other artifacts compiled by the instructor [19]. Typically, a portfolio also includes a narrative from the instructor explaining the significance of the artifacts and often a self-reflection on teaching [19]. The compiled portfolio is then reviewed by colleagues or administrators [19]. A wide variety of materials can be included in a teaching portfolio. Examples include syllabus, course guides, sample course handouts and tests, self-evaluation statement, description of course development efforts, presentations, summaries of student evaluations of teaching, reports of peers who have observed the class, and samples of graded student work [19] (p. 113).

An important strength of portfolios is that they allow for a multidimensional portrayal of teaching [19,25]. In particular, teaching portfolios are the only assessment method we discuss that can capture a teacher's thinking about the teaching process and efforts to improve. This is notable because current literature stresses the importance of continuous cycles of assessment and improvement in high quality teaching practices [5,19,26–28].

4. Research-based assessments

Researchers in physics and other STEM disciplines have developed a number of research-based assessments that can be used to measure teaching effectiveness. These instruments can be used for both instructor and institutional evaluation of teaching. By far the most common of these are concept inventories [29,30]. Concept inventories are generally multiple-choice instruments that are developed based on research into common student difficulties with a particular concept or set of concepts. Development also includes a cyclic process of checking the validity and reliability of the instrument with refinements made to the instrument as the development process proceeds [31,32]. Most commonly, concept inventories are given pre- and post-instruction and learning gains are calculated.

Concept inventories have been used in a variety of situations to assess teaching effectiveness [29,30,33–36]. For example, strong conceptual inventory gains were helpful in maintaining the Technology Enabled Active Learning program at MIT in spite of strong student criticism [33]. Conceptual inventories have also been identified as important learning measures that can be convincing for individual instructors [37].

In addition to concept inventories, other research-based assessments exist that focus on other types of learning outcomes for science courses, such as students' beliefs and attitudes about learning [40].

5. Exams, quizzes, or homework

Student performance on tests are a common way for instructors to gauge their instructional effectiveness [8,9]. Instructors, especially in the sciences, "tended to believe that they could quantify student achievement based on demonstrated ability to supply information or solve problems on tests" [9] (p. 238). Some institutions, such as the University of Phoenix, also evaluate instructors using student performance on university-developed subject matter tests [41].

6. Systematic formative assessment

Systematic formative assessment (sometimes called classroom assessment techniques) are designed for the dual purpose of helping students learn and helping teachers assess how well the students are learning [9,42]. Formative assessment is typically integrated into teaching activities and, thus, formative assessment data would need to be processed in some way for use by the institution for assessing teaching effectiveness.

Angelo and Cross [42] describe a wide variety of systematic formative assessment techniques. These include minute papers (in the last few minutes of a class the

instructor asks students to write short responses to identify the most important thing they learned in class and what questions they still have), muddiest point (students are asked to identify the muddiest point in a lecture, reading, homework assignment, etc.), and what's the principle (students are given a few problems and asked to identify the principle or principles that best apply to each). Muddiest point has been adapted by Etkina and colleagues as part of their weekly report homework assignment [43]. In physics, however, the most common form of formative assessment is the use of multiple-choice conceptual questions for use during class—often using a classroom response system. This technique was popularized by Mazur as part of Peer Instruction [44]. More recently, additional work into the structure of questions and the development of banks of well-developed questions has been undertaken by several research groups [45–47].

7. Informal formative assessment

Although less is written about this type of formative assessment, informal exchanges with students during class and out of class are known to be meaningful sources of feedback for college instructors [9,10] and may be one of the primary sources of information about teaching effectiveness used by K-12 teachers [48].

8. Postcourse feedback from students

Similar to informal formative assessment, postcourse feedback from students is not well documented in the literature. For high school teachers, there is some evidence that students returning after graduation to thank teachers are an important source of assessment of success [47]. Exit interviews with students at the end of a degree program are sometimes used by institutions to get feedback about specific instructors and courses. Some institutions also request feedback from former students (who may have already graduated) for instructors being considered for tenure or promotion.

9. Other

The eight ways that instructors or institutions can evaluate teaching effectiveness listed above were all represented in the literature as well as in the interviews we conducted with 72 physics instructors for this study. In addition to these methods, there were some other methods that were mentioned in the research literature. These include the following:

(i) Teaching awards. Berk [8] identifies teaching awards as one of the 12 strategies to measure teaching effectiveness. In his review of the literature on teaching awards, he largely dismisses the value of teaching awards as a measure of teaching effectiveness due to high variability between institutions and often lack of transparency about how teaching award winners are selected [8]. Gibbs [49] agrees with these criticisms of teaching awards and provides guidance for those who wish to develop more meaningful procedures for clarifying criteria and selecting recipients for teaching awards.

(ii) Instructor engagement in scholarship of teaching and learning. Several have argued that an important indicator of teaching effectiveness is the extent to which instructors engage in systematic study of their own teaching, a process often known as the scholarship of teaching and learning [8,50,51].

(iii) Institutional data on student performance. Institutionally available data can be used to evaluate teaching effectiveness. This includes student performance in later courses [52] and student attrition—percentage of students not passing a class [33].

(iv) Student attendance. The percentage of students who attend lectures has also been suggested as a measure of teaching effectiveness [52].

V. METHODS

In the fall of 2008, a randomly selected sample of U.S. physics instructors were asked to complete an online survey about their instructional goals and practices as well as their knowledge and use of a set of 24 research-based instructional strategies that could be used for teaching introductory quantitative physics (see Ref. [53] for more details). Respondents included instructors from both four- and two-year institutions. The overall response rate for the survey study was 50.3%. A subset of survey respondents was purposefully chosen to participate in an associated interview study. Interviewees were selected to represent users, former users, and knowledgeable nonusers of two research-based instructional strategies: Peer Instruction [44,54–56] and Workshop Physics [57,58]. Interviewees were also selected to represent both two-year and four-year colleges and universities as well as both male and female instructors. Of the 100 instructors we attempted to contact for interviews, 72 (72%) agreed to participate. Instructors not interviewed were approximately equally split between declining to participate and not responding to our requests. Interviews were conducted via telephone and lasted approximately one hour. Interviewees were given a $75 stipend for participating in the interview.

The semistructured interviews were primarily focused on the participants' knowledge about and use of one of the two target instructional strategies (Peer Instruction or Workshop Physics). However, participants were asked to explicitly discuss assessment issues through the following interview questions: (A) How do you know if your instruction is working? (B) What evaluation criteria does your institution use to evaluate teaching? (C) Is there an observational component to your teaching evaluation? If so, are there any specific criteria or specific behaviors that are being looked for in these observations? (D) Do you receive feedback on

your teaching from your students? What kinds of things do students use to evaluate good teaching? In addition, issues related to the assessment of teaching often came up in other parts of the interview. Interviews were audio recorded and transcribed. The complete interview transcript was used for analysis.

For the purposes of understanding interviewees' views about assessment of teaching, each interview was first coded using a broad category of "commentary on assessment." This code was applied to all discussions of course level and institutional level assessment practices. These assessment issues ranged, for example, from reading the expressions on students' faces to judging their level of understanding to the institutional use of SETs to determining the teaching competence of an instructor for promotion and tenure decisions.

Once all of the portions of the interview that were related to assessment were identified through the broad coding, a more detailed coding scheme was developed iteratively by two of the authors (C. H., C. T.). The detailed coding categories were developed through the iterative analysis of approximately 12 of the interviews as well as our understanding of the research literature. The more detailed coding scheme consisted of eight common sources of assessment information (see Table I). With respect to each source, during the analysis we sought to answer the following questions:

(1) Was this source mentioned as potentially relevant for an instructor's (or institution's) assessment of teaching? An assessment source was rated as mentioned if it could be inferred from the interview that the interviewee felt that the type of information could be used by an instructor (or institution) to assess teaching effectiveness. It is important to note that lack of mention of an assessment source does not necessarily mean that the interviewee is unaware of it or that it is not used by the interviewee or his or her institution. Given the multiple questions about assessment throughout the interview, though, it is reasonable to infer that lack of mention of an assessment source likely means that this source is not a particularly important influence (either positively or negatively) on the instructor's or institution's decision-making process.

(2) Was this source of information actually used by the instructor (or institution) to assess teaching effectiveness? An assessment source was rated as used if it could be inferred from the interview that the interviewee (or institution) actually used this type of information to assess teaching effectiveness. If information was collected for other purposes, then it was not counted as being used for assessment of teaching effectiveness. A good example is the use of instructor-developed tests. It is likely that all of the instructors we interviewed give tests as an important way to assign grades to students. Yet, not all interviewees indicated that they use student test performance as a way to assess their teaching effectiveness.

(3) Did the interviewee see value in their own use (or the institution's use) of this source of information? An assessment source was rated as valued if it could be inferred from the interview that the interviewee felt that the type of assessment information could be useful for their own (or their institution's) assessment of teaching effectiveness. Note that the source of information does not actually need to be used to be valued, and vice versa.

Each question was answered in one of three ways (yes, no, unclear) along with a short justification. The criterion for answering "yes" or "no" was the researchers' inference from the interview transcript. This criterion was used because many interviewees did not make explicit statements about these issues. We were quite careful in making inferences to articulate to ourselves the basis for each inference. For example, referring to the use of student evaluations, one instructor said, "I've gotten decent teaching evaluations so I can't say that I've been disappointed with what I've been doing" (G14).[1] In this statement, the instructor does not explicitly answer any of the three questions. However, it is quite clear from the statement that the instructor believes that student evaluations can be used by an instructor to assess teaching effectiveness, actually uses student evaluation results to assess his or her teaching, and finds these student evaluation results to be a valuable assessment source. These inferences were made because (1) the interviewee statement "I've gotten decent teaching evaluations" allows us to infer that the instructor looks at his or her teaching evaluations and is aware of how students rate his or her course on these evaluations, and (2) the interviewee statement "I can't say that I've been disappointed by what I've been doing" suggests that the interviewee connects the student evaluations with an attitude about the success of the course and allows us to infer that the insturctor uses student evaluations as a source of assessment and that the instructor finds this source of assessment valuable. Finally, we inferred that, since student evaluations were used and valued, the instructor was aware that student evaluations could be used. If it was not possible to make a reasonable inference from the text, then the item was rated as "unclear."

In addition to answering these yes-or-no questions, a short (1–3 sentence) description was created for each type of assessment used by each interviewee. The researchers also holistically assessed whether each interviewee felt

[1]Each interviewee is identified by a unique code. The letter indicates the type of institution (T: two-year college; B: four-year college or university with a bachelor's degree in physics; G: four year university with a graduate degree in physics). The number uniquely identifies the individual within that type of institution.

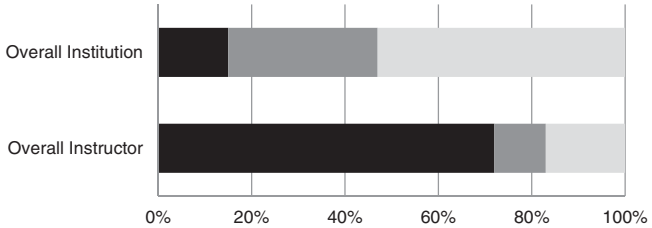
positively or negatively about how well they were overall able to evaluate their own teaching effectiveness. Similarly, the researchers holistically assessed whether each interviewee felt positively or negatively about how well they thought their institutions were able to evaluate teaching effectiveness. These holistic assessments were made based on the entire set of evidence collected. For example, if an interviewee discussed each of the institutional assessment methods negatively, then we rated the instructor as feeling negative overall about the way the institution assesses teaching effectiveness. Because these driving analytical questions arose from the analysis of the interviews and were not asked directly in the interview, in some cases there were insufficient data upon which to classify a particular interviewee.

Each of the 72 transcripts was coded independently by at least two of the four authors. The coding was then compared and any differences were resolved through discussion. Once all of the interviews were coded, summaries of the quantitative and qualitative data were developed. These are presented in the following section.

VI. RESULTS

A. Overall attitudes towards assessment

We begin by examining interviewees' holistic impressions about the success of instructor and institutional assessment practices (as shown in Fig. 1). Of the 72 instructors who discussed their assessment of teaching, most (72%) were judged as feeling positively about their efforts to assess whether their instruction was working. However, only a small minority of instructors (15%) were judged as feeling positively about their institution's assessment of teaching effectiveness. A considerable fraction of instructors (32%) were judged as feeling negatively about how their institution assesses teaching effectiveness. There were also many instructors (53%) for whom we were unable to make an overall judgment about their feelings towards institutional assessment practices.



	Overall Instructor	Overall Institution
% Positive	72	15
% Negative	11	32
% Unclear	17	53

FIG. 1. Interviewees' overall attitudes of how well they and their institutions are able to assess teaching effectiveness.

Figure 1 suggests that instructors have more confidence in their own assessment practices than they do in the assessment practices of their institutions. Because the judgments about the interviewees' overall attitudes were made holistically based on data from throughout each interview, in most cases there was not a simple and coherent supporting quotation available. However, some instructors made their attitudes clear within a fairly confined quotation, such as those used in the examples below. The two most common overall attitudes found were positive attitudes of how well faculty judge their own teaching effectiveness and negative attitudes of how well institutions judge their teaching effectiveness.

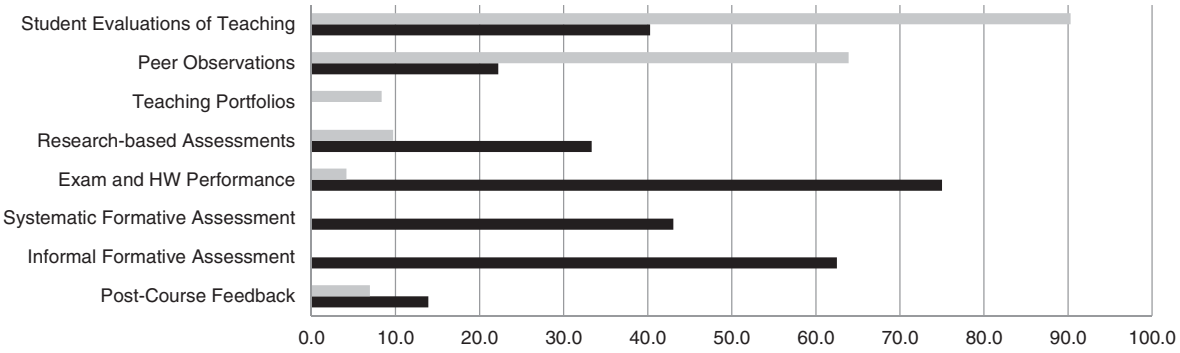
For example, most interviewees felt positively about their ability to assess their own teaching effectiveness. One interviewee said, "*So I think that is what I am looking for [being able to talk like a physicist], and if I can feel like the majority of the class is headed in that direction by the end of the semester, really being able to explain why they think what they think, and listen to an argument, and really be able to pick at it and question it, that is when I feel good about what I did*" (B14).

Another said "*one of the things that I do, when I get the test, I make a spreadsheet of what the score that each student got on each question, so that I can kind of judge is my technique for this question working or not, for this particular subject matter. So that is something that I do to kind of try to judge. Another thing is just feedback from my students. I am an approachable person; my students actually come and see me. So I ask them is that helping, is that working, and use their feedback*" (G21).

Negative attitudes about institutional assessment of teaching effectiveness were explicitly held by about 1/3 (32%) of interviewees. For example, one interviewee said, "*it's the professor's job to figure out how they can best learn. But until you're tenured the reality is, you know, your evaluations had best sing*" (B16). Another said, "*Isn't that terrible? No one holds me accountable to my teaching style ……There really isn't much evaluation of teaching happening*" (T4).

B. Use of assessment strategies

Figure 2 summarizes the quantitative counting of the types of assessment strategies that instructors say they use and the types of assessment strategies that instructors say their institutions use. For this analysis "use" was defined as using to assess teaching. It is possible that an instructor mentioned something being a component of their course but did not mention using it as a source of assessment. So, for example, if an instructor reported being required to collect end of semester student evaluations as part of the official evaluation system of his or her institution, but did not indicated using the student evaluations to inform his or her own teaching, we would code student evaluations as being used by the institution but not by the instructor.



	Post-Course Feedback	Informal Formative Assessment	Systematic Formative Assessment	Exam and HW Performance	Research-based Assessments	Teaching Portfolios	Peer Observations	Student Evaluations of Teaching
Used by Institution	6.9	0.0	0.0	4.2	9.7	8.3	63.9	90.3
Used by Instructor	13.9	62.5	43.1	75.0	33.3	0.0	22.2	40.3

Percentage of Instructors Reporting Use

FIG. 2. Reported use of various sources of assessment information by instructor and by institutions in judging teaching effectiveness.

The figure shows that there is very little overlap in the sources of assessment information used by instructors and those reportedly used by institutions. Instructors report that institutions primarily use student evaluations of teaching (90%) and peer observations of teaching (64%) to judge teaching effectiveness. Instructors primarily use students' performance on exams, quizzes, and homework (75%) and informal formative assessment (63%).

Instructor and institutional characteristics were examined to identify possible differences in the types of assessment methods. Characteristics identified were gender of instructor and type of institution (two-year college, four-year college with a physics B.A. as highest physics degree, or four-year college with a physics graduate program). With these two characteristics we tested four null hypotheses: (1) there is no difference between the types of information that male and female instructors report using to assess teaching effectiveness, (2) there is no difference between the types of information that male and female instructors report their institutions using to assess teaching effectiveness, (3) there is no difference between the types of information that instructors at the three different types of institutions report using to assess teaching effectiveness, and (4) there is no difference between the types of information that instructors at the three different types of institutions report institutions using to assess teaching effectiveness.

Testing each null hypothesis involved conducting eight statistical tests; one for each of the eight types of assessment information. Fisher's exact test was used to test hypotheses 1 and 2 (due to small numbers of people in some categories) and the Freeman-Halton test (an extension of Fisher's exact test that allows for a 2×3 contingency table) was used to test hypotheses 3 and 4. For example, for null hypothesis 1 we conducted a two-tailed Fisher's exact test to test null hypothesis 1a) there is no difference between the use of student evaluations of teaching that male and female instructors report using to assess their teaching effectiveness. The p value of this test was 0.41. The p values for tests of the other seven types of assessment information (null hypotheses 1b-1i) returned p values between 0.108 and 1.00. Because we were conducting eight statistical tests, we needed to adjust the critical p value to avoid type I errors (false positives). Thus, according to the Bonferroni correction, an appropriate critical p value would be the typical 0.05 divided by the number of tests, or $p_{\text{critical}} = 0.05/8 = 0.00625$. Thus none of our eight tests for null hypothesis 1 (1a–1i) were significant and we failed to reject the null hypothesis. Similar procedures were followed for null hypotheses 2–4. For null hypothesis 2, p values ranged from 0.159 to 0.429. For null hypothesis 3, p values ranged from 0.038 to 1.00. For null hypothesis 4, p values ranged from 0.011 to 0.429. None of the tests resulted in p values lower than our adjusted critical value. So, we failed to reject all four of the null hypotheses.

Based on these statistical tests we were unable to identify any differences in the types of assessment methods used based on instructor or institutional characteristics. Of course, given our relatively small sample size we would not expect to be able to detect subtle differences. Nonetheless, this result suggests that there is a large degree of homogeneity among instructors and higher education institutions with respect to assessment of teaching.

The following sections provide more detailed results for each of the eight assessment methods. In order to be clear about the portion of the interview sample represented by each result, instead of identifying the percentage of interviewees we instead identify the number of interviewees.

C. Use and perception of assessment techniques

1. Student evaluations of teaching

Nearly all instructors (65/72) identified student evaluations as a way to assess teaching effectiveness. Every instructor who mentioned student evaluations indicated that their institutions use student evaluations to evaluate teaching effectiveness (65/72). Many instructors (29/72) also said that they use student evaluations to evaluate their own teaching effectiveness.

Most instructors (61/65) who mentioned student evaluations were referring to the standard, anonymous, end of semester student evaluations conducted by the institution. Instructors indicated that there are usually questions about the professor (how prepared they were, how approachable or helpful they were, etc.) and about the class (how much did you learn?, would you recommend the class to others?, etc.). These standard questions are usually evaluated by students on a 4- or 5-point scale. There is also typically an open-ended section. Depending on the institution, this might contain several open-ended questions (e.g., separate questions asking about the class, professor, learning, etc.) or just a single comments section for all open-ended comments.

In the words of one instructor, student evaluations ask questions about whether or not the student agrees with a set of statements:

> *"There is a strongly agree, agree, disagree, strongly disagree category, so there's only four categories. And —let's see, they have you—'Please rate your reaction to each of the following statements regarding the quality of instruction for this course.' And it says, 'The instructor outlined the expectations. The instructor answered questions effectively. The instructor knew the subject matter. The instructor taught with enthusiasm. The instructor explained the material well. The instructor motivated me to do my best work, applied the class policy fairly, was available outside for help, concerned about the students' progress, had appropriate expectations. Presentations were interesting. Returned graded work promptly. And the graded work had careful and helpful comments on them—was graded carefully. The course is well organized. Assignments were a good use of my time. Examination questions reflected the course contents and the emphasis.' "(B24).*

A few instructors (7/65) reported using nonstandard evaluations to assess their own teaching, either exclusively or in addition to the standard evaluations. This usually involved instructor-developed rating forms given out during the semester (3/7) or at the end of the semester (2/7). For example, one instructor talked about the student evaluation he gives during the semester:

> *"I try to at least ….. at least once a semester, give out just a little small 10-minute questionnaire to students to say, you know, 'What have you liked about the course so far? What would you like to see being done differently? What has been posing any difficulties for you?' So that's just my own little tool to get a little bit of feedback to see if I'm missing ….. you know, completely missing the boat on something" (T6).*

A sizeable minority (29/72) of instructors interviewed indicated that they use student evaluations to assess their own teaching. These instructors generally feel positively about their own use of student evaluations (22/29). Some mentioned using student evaluations to see if the students were learning (7/29) or to make changes in their course (7/29). For example, one instructor said, "*So, for next semester, based on their responses, based on my thinking about what was done, I will be able to make changes for the new semester*" (T3). Overall, the most commonly mentioned positive part of student evaluations were the open-ended comments (12/29). Those who felt negatively (5/29) or had mixed feelings (2/29) about student evaluations in general often said the only useful part about them was the student comments (6/7):

> *"A lot of times students do not understand that the bubbles are actually less useful, especially if they are all just sort of vaguely—just kind of—the ranking is not excellent, but it is sort of good. We cannot tell anything unless they write some comments, and a lot of times they do not put in comments. Sometimes the comments are mostly that either they felt the instructor was great, or they have a very, very long complaint about the instructor" (G17).*

Some instructors also mentioned using student evaluations to see if there were serious problems, but thought student evaluations had little other value (4/29): "*To tell you the truth I used to use the evaluations only to detect problems, not to tell me how the class was but only to detect serious problems*" (G11). Only one instructor reported finding value in the numerical part of student evaluations. "*One part of what the institution is looking for is for me not to be down in the low part of the scale because that would be really weird …… and in fact long before my administrators get it, I've had a chance to look at it, and I've surely noticed the same things, and I'm busily trying to figure out what to do about it*" (T9). The most common complaints among those who use student evaluations were that students do not take the time to write valuable comments (5/29), and/or do not know how to evaluate teaching (4/29). "*The consensus is that the students' evaluations tend to be more of a popularity contest than really evaluation of real learning that happens in the classroom*" (B14).

Most instructors interviewed (65/72) said that their institutions use student evaluations to assess teaching effectiveness. Instructors generally feel negatively about how institutions use student evaluations (30/65) although some instructors feel positively (18/65). Many (25/65) explicitly said student evaluations factor into tenure and promotion decisions. Of those who mentioned that their institution uses student evaluation data for tenure and promotions decisions, about a third (9/25) felt positively about this use of student evaluations and two-thirds (16/25) felt negatively. Many thought that students did not know how to evaluate teaching (13/65). Many also said that student evaluations were like a popularity contest, because it just depends on whether the students like you or not (24/65). "*This turns out to be pretty much of a popularity contest and not really in anything based on student learning*" (B13).

Of the instructors who were rated as positive about their institutions' use of student evaluations, only a few actually explicitly said something positive about them (7/18). "*If there are real problems they will come out in those evaluations. If someone is just not teaching well, those instruments are pretty blunt, I'll admit. But frankly the measurement just has to be really coarse. Are you doing a really bad job, or are you okay*" (G7). More commonly, a mildly positive attitude was inferred by the researchers when an instructor mentioned institutional use of student evaluations and did not say anything negative about them (11/18):

> *"Well I think it is how they feel about the person primarily, that is teaching the class. And I think the other thing that they are keyed into is how well they were able to succeed in the class. I think those are the big things that students put in there. But they also put in things where they seem to indicate that they like the activities that we do in class" (T10).*

Such a description of SETs implies that the instructor feels that student responses to SETs are grounded, at least in part, in meaningful sentiments from students.

2. Peer observations of teaching

Most instructors interviewed (49/72) mentioned the use of teaching observations as a way to assess teaching effectiveness. Teaching observations are when a peer or supervisor attends one or more class session and creates a written or oral report. This report could be in the form of a letter, rating form, or informal conversation. Many instructors thought that student engagement was an important factor for which observers were looking (13/49). Of the instructors who described the observation in detail (29/49), over half said there was no explicit criteria for which the observers were looking (17/29). "*And it's also something that's highly individual—that everybody that does the evaluation really makes up their criteria for what they want to see. So when I go in and I evaluate somebody, it's actually sort of my decision what I get to evaluate*" (T2). Less than half said that there were some broad categories of things that typically guided the observations (12/29).

> *"So for instance, you have to have mastered the subject matter. You have to have interaction with students as far as rapport and partial respect and humor. You have to have classroom presence, which are things like awareness of physical conditions in the classroom, avoiding distracting behavior, awareness of students as opposed to group and individuals. So there are lots of things, but also making sure that the students are actually actively involved in the class" (B4).*

All 49 instructors who mentioned teaching observations did so in the institutional context. Most (36/49) explicitly said that teaching observations are used in an evaluative way as a factor in tenure and promotion decisions. Most of the others (10/49) indicated that they were required to have teaching observations so it is likely that these observations factored into tenure and promotion decisions, but this was not explicitly stated. The remaining instructors (3/49) expressed interest in having peer observations, but their institutions were not using them. In most of the institutions, teaching observations can be done by any other instructor (40/49), but in some institutions it is required that at least one observation be done by the chair or other administrator (12/49).

Most instructors (34/49) feel positively about how their institutions use peer observations. This is generally because these observations help the institution get a more complete understanding of their teaching skills. A few (4/34) also mentioned teaching observations being a better measure of teaching effectiveness than other methods such as student evaluations. Some (14/34) explicitly talked about good feedback they received from teaching observations.

> *"I had a speech person who was on my peer team and she said, 'Do you realize what you do when you ask a question?' I said, 'No, what do I do?' And she said, 'Well what you do is you're all animated, and engagement, and you're chatting with them, and smiling, and all the rest of the stuff, and then you drop out this question, then your face gets really flat. And in a nonverbal kind of way what you've actually done is gone from being very accessible to be[ing] very threatening. And it sort of takes the people in your class aback when you do that.' And then sometimes because I'm almost two meters tall, I go stand next to somebody when I ask them a question. She said, 'Do you have any appreciation for how damn tall you are when you stand next to somebody in a chair?' She said, 'Stand back a little further, you won't be so threatening. They won't*

think you're going to eat them.' I said, 'Oh that's all very helpful stuff.' So those kinds of things come up a lot" (T9).

A few instructors (4/49) felt negatively about institutions using peer observations. One did not trust the people doing the observing to be fair, one thought their institution was too traditional in their requirements, one indicated that observers did not observe enough of the class to get a fair reading of it, and one just did not like how open-ended the evaluation could be. In all four cases, these peer observations were used in the tenure and promotion process.

Some instructors (16/49) also say they also use peer observations for their own assessment of teaching. There were no instructors who used peer observations for themselves where peer observations were not also used by their institution. All 16 felt positively about their own use of peer observations. Most often they find the direct feedback from the observers useful in adjusting their own teaching (15/16).

"Well the interactions that I have after they come in and sit in on my class. There is a lot of discussion—'Okay we saw that you did it in this way. I did that once and I found that if I did it this way it worked a little bit better. Have you tried that—you did, did it work.' And so there isn't this, 'You have to do it this way because that is how we think it works.' There tends to be a lot of open discussion" (B2).

3. Teaching portfolios

Teaching portfolios were not commonly mentioned as a method for assessing teaching effectiveness. Only a few instructors (6/72) said that their institutions used teaching portfolios to assess teaching effectiveness, and no instructors indicated using teaching portfolios to assess their own teaching. Although none of the instructors mentioned specific institutional requirements for portfolios, they described their portfolios as including assignments, handouts, exams, and other student documents (4/6) and/or a self-evaluation written by the instructor (3/6).

While only half (3/6) of instructors explicitly mentioned the portfolios being used for tenure or promotion decisions, all indicated they were included in the general evaluation of their teaching. Most (5/6) did not seem to have strong feelings one way or the other about these teaching portfolios and just mentioned them in passing. "*You have to do like a self-evaluation every couple of years*" (T22). One was clearly positive because it allowed him to report if he had tried something new in his instruction that might be throwing off the student evaluations. "*I think if you're trying a new technique, since you can mention that you're doing that, the kind of the understanding is that maybe if it doesn't work so well the first time, people won't really hold that against you as much*" (B23).

4. Research-based assessments

Approximately half (38/72) of instructors interviewed mentioned research-based assessments such as Force Concept Inventory (FCI) [59] or Force and Motion Concept Evaluation [60] as a way to assess teaching effectiveness. These are commonly used research-based conceptual tests for introductory-level classes that can help to document students' conceptual learning during a course and how the conceptual learning in a particular class of students compares to published learning gains. Of these, most (26/38) reported using a research-based assessment. Research-based assessments are typically used pre- and post-instruction to see how much the students have gained (21/26).

"I wanted to see whether the students actually learned something in my course. So what I did was, I made them do the entire test during one of their first recitations to get a beginning of semester calibration point. And then I actually made them take the whole test again as part of the final exam at the end of the semester" (G19).

Occasionally, however, only post-instruction tests are given (2/26) assuming the pretest scores did not vary over time. "*I found that my pretest scores were pretty much the same all the time. So I stopped doing that because I didn't feel it was a productive use of our time. It wasn't telling me anything I didn't already know*" (T10). In three cases, it was unclear whether the assessments were used pre-post, post-only, or some other way.

Most (24/26) instructors use the research-based assessments as a way to assess their own teaching. "*I try to do a pre-post, kind of analysis. I think I'm pretty well known for taking the feedback and trying to fold it back in*" (B3). All 24 instructors felt positively about using research-based assessments. In addition, five instructors indicated not using research-based assessments but liked the idea of the research-based assessments being used to assess teaching effectiveness. "*I remember he [another instructor] made a—he had a test. He would ask certain questions, basically in a test scenario, before he got started. And then when he was done with this subject matter, then he would do it again. But that way he will—could demonstrate how much people learned*" (G23).

Only seven instructors indicated that their institutions use research-based assessments as a way to assess teaching effectiveness, typically at the department level. In five of these cases, both the instructor and the institution made use of this information. Only two instructors mentioned their institution using research-based assessments when they did not also use the assessments to help assess their own teaching effectiveness. All seven instructors who mentioned their institutions using research-based assessments felt positively about it. "*We've talked about whether we thought there were other books we should use or whether*

this was serving the students well. The improvement on the FCI shows that we're continuing to be effective" (B18). Some instructors expressed interest in their particular institution using research-based assessments (5/38), but that their institution does not currently do so. Most (4/5) of these instructors already use research-based assessments for themselves. One does not but wishes it was used at the institution level. "*At the department level, I'm sorry to say, I think we do a pretty poor job [of assessment]. So like I said, we've never had any push to do something like the FCI*" (B7). Only one instructor was negative about the idea of research-based assessment being used even though the institution does not use it. "*Yeah okay you've done a great job of teaching them the concepts but can they solve problem? Can they do the math? Can they actually calculate anything at the end of the day?* (G22). Five instructors mentioned research-based assessments as a method that could be used to measure teaching effectiveness in some way, but did not elaborate and did not use them.

5. Student performance on exams, quizzes, or homework

Most (54/72) instructors interviewed mentioned the possibility of using student performance on regular coursework as a way to assess teaching effectiveness. This student performance could be any kind of graded work such as exams, quizzes, or homework. All 54 of these instructors used student performance measures to gauge their own teaching effectiveness. Most (49/54) felt positively about their use of student performance to judge teaching effectiveness. Two instructors felt negatively about using student performance to judge teaching effectiveness. One indicated that they did not think exam performance provided enough information about how the students were doing (T2) and another indicated that it is difficult to develop a sufficiently detailed picture of student understanding from either homework or exam scores (B21). There were also three instructors who did not clearly express how they felt about using student performance as a way to judge teaching effectiveness.

Only a few instructors (3/72) indicated that their institutions use student performance on regular coursework to measure teaching effectiveness. One of these three was positive about it and indicated that the institution's guidelines ask instructors to evaluate university objectives such as critical thinking by analyzing student test responses on selected tests using a university-developed rubric. It was unclear how the other two instructors felt about their institutions using student performance on regular assignments to determine teaching effectiveness. In both cases these instructors mentioned that their departments use common final exams to compare student performance in parallel sections of the same course.

The most common type of coursework used in this category was student performance on tests or exams (47/54). "*So, I collect the tests. I grade them. I pass them back during the period and we talk about it. And then I collect them again and keep them. So I'm able to compare how successful students have been on a particular question or issue from semester to semester*" (T5). Some of these instructors (14/47) specifically mentioned using conceptual questions as part of the tests, however none mentioned using only conceptual questions. "*We just concentrated more on trying to focus on the conceptual bits and to get some baseline on the traditional problems similar to the ones that had been offered in the past*" (G1). The second most common type of coursework in this category was graded homework (17/47). "*Seeing what their homework problems look like not just if they're right or wrong but—we don't have TAs for grading homework so we grade all of our homework. We see what they do. We require that they put in steps. And so we can get a sense of how well did they get those intermediate steps*" (B6). Nearly as many instructors (11/47) mentioned quizzes as an important type of coursework used in this category.

There were also a few instructors (4/54) who mentioned using pre- and post-testing with their own, nonresearch-based tests to assess their teaching effectiveness.

6. Systematic formative assessment

Nearly half (33/72) of the instructors interviewed mentioned systematic formative assessment techniques as a way to gauge teaching effectiveness. None mentioned their department or institution using systematic formative assessment to gauge teaching effectiveness. Systematic formative assessment is any methodical sampling of student work. This could include, for example, using clickers, cards, or a show of hands during class; systematic sampling of group work during class; or using Just-in-Time Teaching techniques [61] outside of class. Most of the instructors (31/33) who mentioned the possibility of using systematic formative assessment to judge teaching effectiveness indicated that they use it to judge their own instruction. "*I mean definitely the way that I've used it is really for getting a sense of whether students understand very basic concepts. So getting that quick assessment of whether or not they get this idea, and how we can move on*" (B14). All 31 instructors who used systematic formative assessment were positive about their use of it to gauge how their instruction was going.

The most common way that systematic formative assessment was described (22/31) was as part of Peer Instruction. (It is important to keep in mind that approximately half of the interviewees were specifically selected because they knew about Peer Instruction.) This typically involves having the students answer a question individually, discuss it with their neighbors, and then answer the same question again in hopes that they will convince each other of the correct response.

"I always want them to do it individually first because I want to see where they are. If only one person gets it wrong then I don't see the point in having them discuss it very much. So yeah, I wait until I see a mistake and then —or I would say disagreement amongst the students— then I think it's valuable for them to interact and talk to somebody that has a different point of view, and then they can try to come to an agreement" (T12).

Approximately half the instructors (16/31) explicitly mentioned using conceptual questions as part of their systematic formative assessment. Personal response systems (i.e., "clickers") were the most common way (13/31) to gather student responses from students.

Some instructors (8/31) also mentioned using a systematic sampling of group responses to gauge teaching effectiveness. "*I actually go around to the groups so I can actually see how they're doing. I actually ask them. It's kind of like a little mini oral exam every class because I'll go around. I'll say, 'Okay. Well why did this happen?' And if they can explain it to me, then I can see that they know what's going on. Many times I can see that they don't know what's going on*" (B4).

A few interviewees (4/31) mentioned using Just-in-Time Teaching [61] as a way to gauge their teaching effectiveness. One instructor mentioned looking at the quality of student questions in a systematic way (1/31). "*When I can, I sometimes do minute papers where they will just write down a question that they might have on the material*" (G7). There were also a few instructors (2/33) who liked the idea of using systematic formative assessment but did not use it in their own classrooms. "*One of the strengths [of Peer Instruction] is that you obviously get some instantaneous feedback on how well students are understanding it and then what sorts of misconceptions there might be. That's the main reason that I'd like to start doing this is because you do get that kind of feedback*" (B23).

7. Informal formative assessment

Many of the instructors interviewed (46/72) mentioned informal formative assessment as a way to judge teaching effectiveness. Of these, almost all (45/46) use informal formative assessment to judge their own teaching effectiveness. Informal formative assessment is any informal feedback about the instructional success that the instructor gets from his or her students, either in or out of class. It can be anything from listening in on group work happening in class to see if the students are understanding, to the types of questions students ask during office hours, to a general sense of how engaged the students are. While this information is often very valuable to the teacher, none of the instructors interviewed indicated that their institutions or departments use this information as a measure of teaching effectiveness.

There was also one instructor who mentioned the possibility of using informal formative assessment but does not use it due to concerns about the validity of this type of evidence. "*I should stress that I sort of feel, at least from my point of view, that I'm a bad judge of what is working and what is not working from the front of the classroom*" (T7).

Of the 45 instructors who indicated using informal formative assessment to assess their own teaching effectiveness, nearly all (43/45) felt positively about it. "*This is pretty good feedback as to how well the class is doing*" (B12). The others (2/45) did not express a clear attitude. The type of informal formative assessment used most often was the instructor's general sense from the front of the room of how engaged the students were in the lesson, or how much the students seemed to be understanding (18/45). "*And then from my point of view, I also, if I observe what they are doing then I can tell whether or not they are understanding what I'm talking about*" (T12). Other types of informal formative assessment used were individual conversations with students, such as during office hours (16/45), responses to questions asked in class (14/45), as well as the types of question students ask during class (13/45).

"When I'm doing some talking from up front, I'll frequently ask the class a question. And, I won't always use the peer instruction method. I'll just say 'Well, what do you think would happen if I did this?' And, I just wait to see if folks respond. And if several students eagerly respond, that doesn't necessarily mean that anybody knows what's going on other than those several students. But if no students respond and they all look completely confused then I know that I'm not getting through" (B17).

A few instructors also judge their teaching effectiveness by walking around while students are doing group work and listening to students talk to one another (8/45). This was considered informal formative assessment unless they described a systematic process for doing this.

"I guess a lot of it comes from when I'm walking around and talking to the students during the class. You know, you kind of have the—you kind of can get the feeling whether they're getting it or not from that exercise, like from the questions they ask you and, you know, what you see them doing" (B16).

8. Postcoursefeedback from students

Some of the instructors interviewed (14/72) mentioned getting feedback from students after the course as a way to assess teaching effectiveness. This feedback can take many forms such as an informal e-mail from a former student to the professor, an exit interview, or the institution requesting

a letter from former students of a professor up for tenure. Most of the instructors use the postcourse student feedback for themselves (10/14) and all feel positively about it. "*I talk to them to see how they are doing, and if they seem to be performing well, and they say 'Oh, I am so glad I took physics before taking the strength of material class',' because they are using the information they learned in the physics class in the next class. So, I am happy to learn they are successful*" (T3). All of the instructors who use postcourse student feedback use student initiated feedback from spontaneous e-mails, conversations, or as part of ongoing communications with former students.

There were also a few interviewees (5/14) who mentioned that their institutions use student feedback after the course to gauge teaching effectiveness. In all of these cases the feedback was formally requested from the students by an institutional representative. This was either in the form of letters requested from students of an instructor up for tenure or promotion (1/5), exit interviews of physics majors (2/5), student surveys a while after the course (1/5), or systematically contacting transfer students (1/5). In two cases, the information was used as part of the tenure and promotion process. Most (4/5) instructors felt positively about their departments using this information to judge their teaching effectiveness. "*And I've actually been part of the people planning on this. Trying to figure out what students have learned after they graduate or when they graduated. Make sure they've learned certain topics*" (G14). One instructor felt negatively about how the feedback was used. This person's department requests letters from former students when a faculty member is up for tenure or promotion.

> *"When you got up to the higher levels that were just reading these student letters, that was the only criteria. So that's, you know, the dean who's a history professor and has no understanding of what it is to teach physics, you know, was like oh, well five students were unhappy with you. It must mean you're a terrible teacher. So I mean, I think it's an incredibly stupid way to evaluate teaching" (B16).*

VII. CONCLUSIONS AND DISCUSSION

In this section we use the results to answer the four main research questions.

A. What assessment techniques do physics instructors report using to determine whether they are teaching effectively?

Instructors report using primarily student performance on tests and other assignments along with informal formative assessment as ways to assess their teaching effectiveness. Most instructors report being happy about their use of these techniques to assess their teaching.

B. What assessment techniques do physics instructors report their institutions using to determine whether instructors are teaching effectively?

Instructors report that their institutions use student evaluations of teaching and peer observations as the primary methods of evaluating teaching effectiveness. For 16 instructors (approximately one-quarter of those that report their institution uses student evaluations), student evaluations are the only measure that their institutions use to assess teaching effectiveness. However, even when used with other measures, instructors report that student evaluations are typically given the most weight in institutional evaluation of teaching effectiveness.

When institutions use peer observations of teaching as a way to assess teaching effectiveness, faculty report that these observations rarely have any formal criteria or rating rubric, although some institutions do suggest categories upon which the ratings should be based. Most instructors find the peer observation component useful and indicate that it is good to have another measure in addition to student evaluations.

C. To what extent do the assessment techniques that physics instructors report using align with the assessment techniques that they report their institutions using?

There is very little overlap. Approximately one-third of the instructors who report their institutions use peer observations to evaluate teaching also report using this information themselves. Although Fig. 2 shows that approximately half of the instructors who report that their institutions use student evaluations also use this information themselves, the instructors primarily value the open-ended comments over the numerical portion of the evaluations. Instructors report that the assessment techniques they use most frequently (student performance on tests, informal formative assessment, and systematic formative assessment) are rarely, if at all, used by their institutions.

D. To what extent do the assessment techniques that physics instructors and their institutions use align with the research literature on assessment of teaching effectiveness?

There is very little overlap. We will focus here on seven important areas where practice is not aligned with the research literature.

1. Many institutions assess teaching primarily or exclusively based on student evaluations of teaching. Research suggests that multiple measurement techniques are necessary.

The most common technique that institutions use to assess teaching effectiveness is through student evaluations

of teaching. In this study, 90% of instructors reported that their institutions use this measure. For 16 instructors (approximately one-quarter of those who reported their institutions use student evaluations) this is the only measure that their institutions use to assess teaching effectiveness. There has been much concern about the value of student evaluations expressed in the research literature. Although there is disagreement about whether there is any value in student evaluations, there is strong agreement that student evaluations should not be the only way to assess teaching effectiveness [3,6–8,11,14,19–21].

2. Institutions that do use peer observations rarely have any formal criteria or rating rubric. Research suggests that useful observations have systematic criteria.

The second most common technique that institutions use to assess teaching effectiveness is peer observations. Approximately two-thirds of faculty reported that their institutions use this measure. Although instructors are generally happy with this measure, nearly all of the peer observation techniques described are subject to Chism's [19] criticisms of typical and low quality peer observation techniques. In particular, none of the interviewees reported that there was any training for observers, that more than one class session was observed, or that any rating form or rubric was used. In addition to Chism's suggestions for making teaching observations, there are a number of observational instruments that have been designed for research purposes to assess the alignment of teaching with research-based instructional strategies [62,63]. These instruments could likely be adapted for the purpose of peer observations.

3. Neither faculty nor institutions make much use of research-based assessment instruments. Research has found these instruments to be very useful.

An important contribution of physics education research to the assessment of teaching effectiveness is the development of research-based assessment instruments, such as Force Concept Inventory [29,30,33–36,38]. Yet, these instruments are not widely used by instructors or their institutions. In our interviews, approximately one-third of instructors reported using these instruments to assess their teaching and less than 10% said that their institutions used these instruments. It is important to note that our interview sample was specifically chosen based on some familiarity with the research-based instructional strategies Peer Instruction or Workshop Physics. Thus, it is likely that the use of research-based assessments in the general population of physics instructors is lower.

4. Neither faculty nor institutions make much use of portfolios. Research suggests that the continuous reflection-improvement process is an indicator of high quality teaching.

One of the least used assessment measures overall was teaching portfolios. Interviewees reported that only 8% of their institutions used portfolios to evaluate teaching effectiveness and none of the interviewees indicated that they used teaching portfolios for their own assessment. This is in contrast to the literature on reflection in teaching that suggests that continuous cycles of data collection, reflection, and improvement are key indicators of teaching effectiveness [5,19,26–28]. Teaching portfolios are the only assessment measure discussed in this paper that would allow an instructor to document their continuous improvement process.

5. Faculty most commonly use student test performance as a measure of teaching effectiveness. Research suggests that students' ability to solve numerical problems is not a good measure of learning.

The most common technique that instructors use to assess their teaching effectiveness is student performance on tests and other assignments. Approximately three-quarters of instructors reported using this method. Although some instructors (approximately one-third of those who said they used student test performance) indicated that their tests contain conceptual questions, given typical practice, it is very likely that most instructor-developed tests consist primarily of traditional numerical problems for students to solve. A strong result from the physics education research literature is that students can learn how to solve these problems by rote without having an understanding of the underlying physics reasoning [30,64,65]. Thus, the use of student test performance may lead some instructors to overestimate their teaching effectiveness.

6. Faculty use informal formative assessment measures more frequently than systematic formative assessment measures. Research suggests that both are important.

The second most common technique that instructors use to assess their teaching effectiveness is informal formative assessment. Use was reported by nearly two-thirds of instructors. Using informal and incomplete information to make decisions in real time about how to proceed is no doubt a very important part of teaching [66]. High levels of use of informal formative assessment should be expected and encouraged. However, there has been an increase in research on the importance of additional systematic formative assessment [30,67] and the availability of tools (i.e., clickers) to make systematic formative assessment easier. Thus, the 42% of instructors who

reported using systematic formative assessment is not consistent with the importance of systematic formative assessment noted in the research literature, which would suggest it should be integral to all high-quality teaching [5,26].

7. Neither faculty nor institutions make use of institutional data on student performance. Limited research suggests that these measures can contribute to the assessment of teaching effectiveness.

Finally, institutional data on student performance have been suggested as potentially useful in assessing teaching effectiveness [33,52]. Student performance in later courses and the percentage of students passing a class are both potentially valuable metrics that could be relatively easily tracked. Yet, those types of data were not mentioned by any of our interviewees as a way to assess teaching effectiveness. These types of data are particularly attractive because most institutions already have this data available in their institutional databases. Thus, it would only be necessary to develop the analysis and reporting methods.

VIII. IMPLICATIONS

In this study, we have found that there is little overlap in how instructors evaluate their own teaching, how institutions evaluate instructors' teaching, and research on the effective evaluation of teaching. To some extent this is to be expected as each entity has a slightly different purpose for evaluating teaching. For example, instructors are typically very concerned with day-to-day evaluation, i.e. how are students doing with this particular activity or at meeting a narrow learning goal, and with using feedback to inform what they are doing in the moment. Institutions are more concerned with evaluating teaching over the course of a semester or across multiple semesters. So an instructor's stronger reliance on formative assessment than their institution is to be expected. However, it is likely that this lack of overlap between the assessment measures used by instructors and institutions is one of the characteristics of the current higher education system that limits the use of research-based instructional strategies since it means that each group is paying attention to different things. Similarly, the inconsistency of many of these measures with research recommendations for assessment means that instructors and institutions are often evaluating teaching effectiveness using low quality or incomplete measures. Below we make recommendations for instructors and institutions and recommendations for physics education research. While this study was done in the context of teaching introductory physics, it seems likely that the results and implications are applicable more broadly throughout higher education.

A. Recommendations for instructors and institutions

1. Instructors and institutions should broaden the sources of assessment they use

This study has found that both instructors and institutions use a limited repertoire of the possible assessment methods. Both groups would benefit from including additional methods.

One readily available and easy to use assessment method that would be valuable for both instructors and institutions is research-based assessment instruments, such as the Force Concept Inventory. Approximately one-third of instructors reported using these instruments to assess their teaching and few institutions use these instruments. Although these instruments only measure one type of student learning outcome (conceptual understanding of a particular topic), student performance on these instruments has been found to be a useful measure of course success in a variety of circumstances [34]. These instruments are also desirable because they make it relatively easy to compare student learning outcomes across different institutions. The ability to compare with others, locally and nationally, will increase the ability of instructors and institutions to know when reforms have resulted in positive outcomes, thus making continuation of the reforms more likely. Comparing results across institutions may also help faculty and the research community learn more about what is critical for high-quality implementations of these innovations across institutional and classroom contexts.

Institutions' overreliance on student evaluation data to evaluate teaching is at best an insufficient measure and at worst may be undermining effective teaching. Thus, it is important for institutions to reduce the importance of student evaluations. Probably the most politically acceptable way to do this is to add other assessment measures. Since institutions and instructors already both place value on peer observations, this is perhaps the easiest place to start. Institutions should seek to increase the use of and systematize the procedures for peer observations. In addition, institutions can seek ways to recognize instructor use of formative assessment practices. For example, instructors could be encouraged to report on the types and frequency of formative assessment practices they use.

2. Instructors and institutions should coordinate multiple sources of assessment

One thing that we noticed as we were analyzing the assessment methods used is that it was very unusual for interviewees to relate one assessment measure to another. There were a few cases when an instructor said that high performance on one measure made up for low performance on another measure in an institution's assessment of teaching effectiveness. For example, good peer observation reports can make up for poor student evaluations.

Thus, we recommend that not only should multiple assessment measures be used, but that the use of these multiple measures be coordinated. An excellent way to do this is via the use of teaching portfolios, an assessment measure that is infrequently used. A teaching portfolio allows an instructor to develop a narrative about a particular course that can include multiple sources of evidence to tell a complete story. This is particularly important when an instructor is trying out something new, which may not go smoothly the first time.

3. Instructors and institutions should assess teaching effectiveness based on evidence of reflection as well as evidence of positive outcomes

One issue that we became aware of in the literature on effective teaching is that effective teaching is sometimes described as a process (e.g., in terms of reflective teaching) and other times in terms of outcomes (e.g., student learning, specific teaching behaviors). The assessment literature is highly focused on teaching outcomes. Outcomes are important, but there are potential benefits of also encouraging effective teaching processes. Portfolios are an ideal way to document this process.

The literature on effective teaching processes suggests that good teachers are constantly collecting evidence of student learning on a variety of time scales. On the very short time scales this is what we have called informal formative assessment. On longer time scales, this would include ideas of scholarly teaching [68], teaching as research [50], or reflective practice [27]. Although many faculty use informal formative assessment, the more systematic, longer term reflective processes did not come up as important aspects of assessment for faculty. As some have argued, portfolios can be used as private documents for self-reflection [69] and, thus, we encourage faculty to engage in more systematic self-reflection.

At the institutional level, it is argued that instructors should be able to document that they are involved in an ongoing process of teaching improvement that includes the collection of evidence, reflection, and making changes. Again, this process can be described in a portfolio and portfolios can be evaluated based not only on the outcome measures reported but also on the extent to which the instructor is involved in continuous improvement [5,19].

B. Recommendations for physics education research (PER)

We now briefly turn our attention to suggestions of what the educational research community (especially PER) can do to improve some of the shortcomings in evaluation of teaching effectiveness that we have identified in this study. Given the current national emphasis on accountability in higher education, we suggest that this is an ideal time for the PER community to join the conversation. PER could play a variety of roles in improving and aligning assessment of teaching; here we discuss two important roles: (1) continue to develop and promote the use of tools to document course outcomes, and (2) conduct research on processes that instructors and institutions can use to reach the recommendations for improved assessment of teaching identified in the previous section.

1. PER should continue to develop and promote the use of tools to document course outcomes

While currently available conceptual inventories (such as the FCI) have been shown to be useful in assessing student learning in particular courses, our findings suggest that these instruments are not widely used. In addition to conceptual inventories, educational researchers have also developed attitudinal instruments [70], such as the Colorado Learning About Science Survey [38] or the Maryland Physics Expectations Survey [39]. Based on our interviews, these instruments are rarely used by instructors or institutions. However, these instruments are easy to administer and have been shown to be useful in identifying important shifts in student attitudes and expectations related to learning physics [71,72]. Thus, there is clearly a need for PER to do a better job of promoting the use of these available instruments.

There are, of course, many course outcomes that instructors and institutions value that are not measured by these instruments. Examples include student problem solving ability, scientific thinking skills, or critical thinking skills. Research-based instruments in these areas would be a useful addition to the assessment tools available and PER should ramp up efforts to develop these instruments.

In addition to promoting existing instruments and developing a wider variety of instruments to assist faculty in matching an evaluation to their unique situation and learning goals, there is also a need to assist faculty in interpreting the results of their instruments when they use them (i.e. if an instructor is teaching general physics at a small regional university, using interactive engagement methods, and the students score 30% on the pre FCI and 46% on the post, is that acceptable or is it a concern?). To date only limited information is publically available to assist faculty in making sense of their results when they use one of the existing instruments. A potential solution to this problem is to create a database where faculty can both share and compare data across populations.

One other strong finding of our study is that peer observations, when they exist, are not done in a systematic way and often do not focus on important aspects of instruction. Thus, PER can work to develop peer observation protocols that focus on observable aspects of the class that are aligned with research-based teaching. As mentioned earlier, a potentially productive way to approach this task would be to start with one or more of the observational instruments that have been developed for research purposes [62,63] and make appropriate adjustments for the purpose

of peer observations. We believe the development and promotion of the use of research-based peer observation protocols would be a strong leverage for reform. It is difficult to imagine that formally evaluating teaching along the dimensions of level of interactivity and student-student engagement would not result in an increase in the use of research-based instructional strategies.

A final type of new assessment procedure that this study can point to as promising is the coordinated use of formative assessment and standardized research-based summative assessments. This would help to encourage instructors to be more systematic in their use of formative assessment and also to package this information in a way that institutions could use. PER has developed a number of research tools and techniques that are commonly used to assess student learning, but so far PER has not significantly promoted the use of these tools in the assessment of teaching effectiveness. For example, the University of Washington Physics Education Group frequently uses student performance on open-ended conceptual questions, often embedded in exams of other normal coursework, to assess student understanding of particular topics [73,74]. These questions are also sometimes used by researchers assessing secondary implementation of the University of Washington curricular materials [75]. We believe that the PER community is well situated to (a) provide scaffolds that build on instructors' current assessment practices based on student performance on exams and homework, and formative assessment (e.g., research-based rubrics to evaluate student work [76]), and (b) encourage the complementary use of standardized assessment measures (such as conceptual inventories or other research-based assessments).

2. Conduct research on processes that instructors and institutions can use to reach the recommendations for improved assessment of teaching identified in the previous section

We realize that changing the customary ways that instructors and institutions assess teaching will not be easy. There are many logistical and political barriers that will need to be identified and addressed. For example, if institutional assessment of teaching effectiveness becomes more nuanced in ways that we discuss above, who will be well situated to coordinate this important work within institutions? What training, assistance, and resources do faculty need to change their assessment practices? How should institutions go about revising their procedures for assessing teaching effectiveness? To the best of our knowledge these, and many related questions, are all currently unaddressed by the educational research literature. We suggest that education researchers and change agents build on the currently available literature on instructional change (see e.g., Refs. [77,78]) to develop and research ways to change assessment practices. It will also be useful to conduct research on and coordinate change initiatives with the growing number of extrainstitutional policies that are currently being implemented to hold higher education institutions accountable for their educational missions.

C. Summary

In brief, we found there is little overlap in the ways faculty evaluate their own teaching, how their teaching is evaluated by their institutions, and what research suggests are best practices in assessment and evaluation of teaching. This is a critical finding for those interested in the reform of teaching toward more research-based instructional strategies. It is much less likely for teaching practices to change if desired outcomes are not being measured or used in the evaluation of teaching. We suggest that changing the culture and practice around how faculty and institutions evaluate teaching will be a productive leverage point in efforts to improve STEM instruction in higher education.

ACKNOWLEDGMENTS

This paper is based upon work supported by the National Science Foundation under Grant No. 0715698. We thank Andrew Lewandowski for his contributions to this paper by summarizing the quantitative data reported. We also thank all of the professors who shared their time and experiences with us.

[1] J. M. Atkin and J. Coffey, *Everyday Assessment in the Science Classroom*, Science Educators' Essay Collection (NSTA Press, Arlington, VA, 2003).

[2] P. Ewell, *Making the Grade: How Boards Can Enhance Academic Quality* (Association of Governing Boards of Universities and Colleges, Washington, DC, 2006).

[3] E. Seymour, Tracking the process of change in us undergraduate education in science, mathematics, engineering, and technology, Sci. Educ. **86**, 79 (2002).

[4] E. M. Rogers, *Diffusion of Innovations* (Free Press, New York, 2003), 5th ed..

[5] W. F. Massy, S. W. Graham, and P. M. Short, *Academic Quality Work: A Handbook for Improvement* (Anker Publishing Co., Bolton, MA, 2007).

[6] W. A. Anderson*et al.*, Changing the culture of science education at research universities, Science **331**, 152 (2011).

[7] National Academy of Engineering., *Developing Metrics for Assessing Engineering Instruction: What Gets*

Measured Is What Gets Improved: Report from the Steering Committee for Evaluating Instructional Scholarship in Engineering (National Academies Press, Washington, DC, 2009).

[8] R. A. Berk, Survey of 12 strategies to measure teaching effectiveness, *Int. J. Teach. Learn. Higher Educ.* **17**, 48 (2005).

[9] L. R. Lattuca and J. S. Stark, *Shaping the College Curriculum: Academic Plans in Context* (Jossey-Bass, San Francisco, CA, 2009), 2nd ed..

[10] C. Pfund *et al.*, Summer institute to improve university science teaching, Science **324**, 470 (2009).

[11] S. E. Carrell and J. E. West, Does professor quality matter? Evidence from random assignment of students to professors, J. Polit. Econ. **118**, 409 (2010).

[12] K. A. O'Meara, Beliefs about post-tenure review: The influence of autonomy, collegiality, career stage, and institutional context, J. Higher Educ. **75**, 178 (2004).

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

[14] K. A. Feldman, in *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, edited by R. P. Perry and J. C. Smart (Springer, Dordrecht, The Netherlands, 2007), p. 93.

[15] T. L. P. Tang, Teaching evaluation at a public institution of higher education: Factors related to the overall teaching effectiveness, *Public personnel management* **26**, 379 (1997).

[16] R. Arum and J. Roksa, *Academically Adrift: Limited Learning on College Campuses* (University of Chicago Press, Chicago, 2011).

[17] P. Babcock, Real costs of nominal grade inflation? New evidence from student course evaluations, Economic inquiry **48**, 983 (2010).

[18] V. E. Johnson, *Grade Inflation: A Crisis in College Education* (Springer, New York, 2003).

[19] N. V. N. Chism and C. A. Stanley, *Peer Review of Teaching: A Sourcebook* (Anker Publishing Co., Boston, MA, 1999).

[20] G. Rhoades and B. Sporn, Quality assurance in Europe and the U.S.: Professional and political economic framing of higher education policy, Higher Educ. **43**, 355 (2002).

[21] K. K. Schultz and D. Latif, The planning and implementation of a faculty peer review teaching project, Am. J. Pharm. Educ. **70**, 32 (2006).

[22] M. Wright, Always at odds?Congruence in Faculty Beliefs about Teaching at a Research University, J. Higher Educ. **76**, 331 (2005).

[23] R. M. Felder and R. Brent, The national effective teaching institute: Assessment of impact and implications for faculty development, J. Eng. Educ. **99**, 121 (2010).

[24] E. Seymour S. M Daffinrud, D. J. Wiese, and A. B. Hunter, Creating a better mousetrap: On-line student assessment of their learning gains, paper to the *National Meeting of the American Chemical Society, San Francisco, CA, 2000* (unpublished).

[25] P. Schafer, E. Hammer, and J. Berntsen, in *Effective Evaluation of Teaching: A Guide for Instructors and Administrators*, edited by M. E. Kite (Society for the Teaching of Psychology, 2012).

[26] P. Maki, *Assessing for Learning: Building a Sustainable Commitment across the Institution* (Stylus Publishing, Sterling, VA, 2010), 2nd ed..

[27] C. Amundsen and M. Wilson, Are we asking the right questions? A conceptual review of the educational development literature in higher education, Rev. Educ. Res. **82**, 90 (2012).

[28] P. Hutchings and AAHE Teaching Initiative, *The Course Portfolio: How Faculty Can Examine Their Teaching to Advance Practice and Improve Student Learning* (American Association for Higher Education, Washington, DC, 1998).

[29] K. Garvin-Doxas, M. Klymkowsky, and S. Elrod, Building, using, and maximizing the impact of concept inventories in the biological sciences: Report on a national science foundation sponsored conference on the construction of concept inventories in the biological sciences, CBE Life Sci. Educ. **6**, 277 (2007).

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

[31] R. J. Beichner, Testing student interpretation of kinematics graphs, Am. J. Phys. **62**, 750 (1994).

[32] J. Richardson, in *Inventions and Impact* (AAAS, Washington, DC, 2004), pp. 19.

[33] L. Breslow, Wrestling with pedagogical change: The teal initiative at MIT, Change: The Magazine of Higher Learning **42**, 23 (2010).

[34] R. R. Hake, Interactive engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).

[35] G. Marbach-Ad, Volker Briken, Kenneth Frauwirth, Lian-Yong Gao, Steven W. Hutcheson, Sam W. Joseph, David Mosser, Beth Parent, Patricia Shields, Wenxia Song, Daniel C. Stein, Karen Swanson, Katerina V. Thompson, Robert Yuan, and Ann C. Smith, A faculty team works to create content linkages among various courses to increase meaningful learning of targeted concepts of microbiology, CBE Life Sci. Educ. **6**, 155 (2007).

[36] T. M. Andrews, M. J. Leonard, C. A. Colgrove, and S. T. Kalinowski, Active learning not associated with student learning in a random sample of college biology courses, CBE Life Sci. Educ. **10**, 394 (2011).

[37] J. P. Mestre, Facts and myths about pedagogies of engagement in science learning, *Peer Review* **7**, 24 (2005).

[38] W. K. Adams, K. Perkins, N. Podolefsky, M. Dubson, N. Finkelstein, and C. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado learning attitudes about science survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).

[39] E. F. Redish, J. M. Saul, and R. N. Steinberg, Student expectations in introductory physics, Am. J. Phys. **66**, 212 (1998).

[40] N. G. Lederman, F. Abd-El-Khalick, R. L. Bell, and R. S. Schwartz, Views of nature of science questionnaire (VNOS): Toward valid and meaningful assessment of

learners' conceptions of nature of science, J. Res. Sci. Teach. **39**, 497 (2002).
[41] R. G. Ehrenberg, in *Reinventing Higher Education: The Promise of Innovation*, edited by B. Wildavsky, A. Kelly, and K. Carey (Harvard Education Press, Cambridge, MA, 2011), pp. 101.
[42] T. A. Angelo and K. P. Cross, *Classroom Assessment Techniques: A Handbook for College Teachers* (Jossey-Bass, San Francisco, 1993), 2nd ed..
[43] K. A. Harper, E. Etkina, and Y. F. Lin, Encouraging and analyzing student questions in a large physics course: Meaningful patterns for instructors, J. Res. Sci. Teach. **40**, 776 (2003).
[44] E. Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Upper Saddle River, NJ, 1997).
[45] A. Lee, L. Ding, N. W. Reay, and L. Bao, Single-concept clicker question sequences, Phys. Teach. **49**, 385 (2011).
[46] R. J. Dufresne and W. J. Gerace, Assessing-to-learn: Formative assessment in physics instruction, Phys. Teach. **42**, 428 (2004).
[47] I. D. Beatty, W. J. Gerace, W. J. Leonard, and R. Dufresne, Designing effective questions for classroom response system teaching, Am. J. Phys. **74**, 31 (2006).
[48] M. Fullan, *The New Meaning of Educational Change* (Teachers College Press, New York, 2001), 3rd ed..
[49] G. Gibbs, *Designing teaching award schemes*, (The Higher Education Academy, York, England,2008).
[50] M. R. Connolly, J. L. Bouwma-Gearhart, and M. A. Clifford, The birth of a notion: The windfalls and pitfalls of tailoring an SOTL-like concept to scientists, mathematicians, and engineers, Innovative Higher Educ. **32**, 19 (2007).
[51] G. R. Lueddeke, Professionalising teaching practice in higher education: A study of disciplinary variation and 'teaching-scholarship', Stud. Higher Educ. **28**, 213 (2003).
[52] D. Glenn, One measure of a professor: Students' grades in later courses, in *Chronicle of Higher Education* (2011) [http://teachpsych.org/ebooks/evals2012/index.php].
[53] C. Henderson and M. H. Dancy, The impact of physics education research on the teaching of introductory quantitative physics in the United States, Phys. Rev. ST Phys. Educ. Res. **5**, 020107 (2009).
[54] C. H. Crouch *et al.*, in *Research-Based Reform of University Physics, Reviews in PER*, edited by E. F. Redish and P. J. Cooney (American Association of Physics Teachers, College Park, MD, 2007), Vol. 1.
[55] A. P. Fagen, C. H. Crouch, and E. Mazur, Peer instruction: Results from a range of classrooms, Phys. Teach. **40**, 206 (2002).
[56] E. Mazur and C. H. Crouch, Peer instruction: Ten years of experience and results, Am. J. Phys. **69**, 970 (2001).
[57] P. W. Laws, Calculus-based physics without lectures, Phys. Today **44**, 24 (1991).
[58] P. W. Laws, *Workshop Physics Activity Guide* (John Wiley & Sons, New York, 1997).
[59] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).
[60] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, Comparing the force and motion conceptual evaluation and the force concept inventory, Phys. Rev. ST Phys. Educ. Res. **5**, (2009).
[61] G. M. Novak, *et al.*, *Just-in-Time Teaching: Blending Active Learning with Web Technology* (Prentice Hall, Upper Saddle River, NJ, 1999).
[62] D. Sawada, M. D. Piburn, E. Judson, J. Turley, K. Falconer, R. Benford, and I. Bloom, Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol, School Sci. Math. **102**, 245 (2002).
[63] M. T. Hora and J. J. Ferrare, Instructional systems of practice: A multidimensional analysis of math and science undergraduate course planning and classroom teaching, J. Learn. Sci. **22**, 212 (2013).
[64] L. C. McDermott, Millikan lecture 1990: What we teach and what is learned—closing the gap, Am. J. Phys. **59**, 301 (1991).
[65] E. F. Redish, *Teaching Physics with the Physics Suite* (John Wiley & Sons, Hoboken, NJ, 2003).
[66] F. Erickson, Some thoughts on "proximal" formative assessment of student learning, Yearbook of the National Society for the Study of Education **106**, 186 (2007).
[67] National Research Council, *Promising Practices in Undergraduate Science, Technology, Engineering, and Mathematics Education: Summary of Two Workshops* (National Academies Press, Washington, DC, 2011).
[68] E. L. Boyer, *Scholarship Reconsidered: Priorities of the Professorate* (Jossey-Bass, San Francisco, 1990).
[69] D. R. Woods, *Motivating and Rewarding University Teachers to Improve Student Learning: A Guide for Faculty and Administrators* (City University of Hong Kong Press, Hong Kong, 2011).
[70] L. C. McDermott and E. F. Redish, Resource letter: PER-1: Physics education research, Am. J. Phys. **67**, 755 (1999).
[71] E. Brewe, L. Kramer, and G. O'Brien, Modeling instruction: Positive attitudinal shifts in introductory physics measured with CLASS, Phys. Rev. ST Phys. Educ. Res. **5**, 013102 (2009).
[72] T. Gok, The impact of peer instruction on college students' beliefs about physics and conceptual understanding of electricity and magnetism, Int. J. Sci. Math. Educ. **10**, 417 (2012).
[73] L. C. McDermott, P. S. Shaffer, and M. D. Somers, Research as a guide for teaching introductory mechanics: An illustration in the context of the Atwood's machine, Am. J. Phys. **62**, 46 (1994).
[74] L. C. McDermott, Oersted medal lecture 2001: Physics education research—the key to student learning, Am. J. Phys. **69**, 1127 (2001).
[75] S. J. Pollock and N. D. Finkelstein, Sustaining educational reforms in introductory physics, Phys. Rev. ST Phys. Educ. Res. **4**, 010110 (2008).
[76] E. Etkina, A. Karelina, S. Murthy, and M. Ruibal-Villasenor, Using action research to improve learning and formative assessment to conduct research, Phys. Rev. ST Phys. Educ. Res. **5**, 010109 (2009).
[77] M. Borrego and C. Henderson, Theoretical perspectives on change in STEM higher education and their implications for engineering education research and practice (unpublished).
[78] C. Henderson, A. Beach, and N. Finkelstein, Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature, J. Res. Sci. Teach. **48**, 952 (2011).