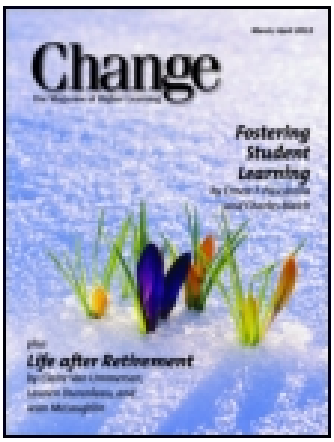


Change: The Magazine of Higher Learning
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/vchn20

A Better Way to Evaluate Undergraduate Teaching

Carl Wieman[a]
[a] Graduate School of Education at Stanford University
Published online: 06 Feb 2015.

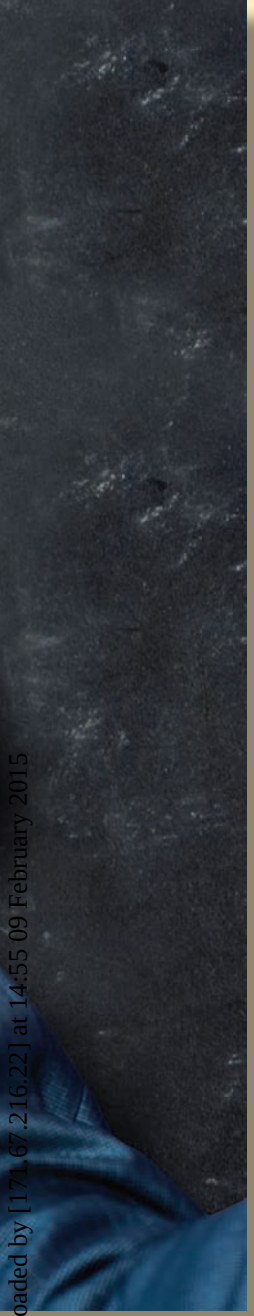
To cite this article: Carl Wieman (2015) A Better Way to Evaluate Undergraduate Teaching, Change: The Magazine of Higher Learning, 47:1, 6-15, DOI: 10.1080/00091383.2015.996077

To link to this article: http://dx.doi.org/10.1080/00091383.2015.996077

In Short

- Current evaluation methods do not allow an instructor or an institution to objectively determine the quality of teaching in a course or show how it can be improved.
- An optimum method for evaluating teaching, based on a detailed inventory of the teaching practices used in a course, allows a quantitative determination of the proportion of the teaching that is done using practices that research has shown result in improved student learning.
- The imbalance between the weighting of research and teaching in the university incentive system is partially due to the inferior quality of current teaching-evaluation strategies.
- For teaching to become a credible component in the incentive system, its evaluation must be valid, enable meaningful comparisons, be fair and practical, and lead to improvement.

Carl Wieman (cwieman@stanford.edu) is a professor of physics and a professor in the Graduate School of Education at Stanford University. He is the founder of the Carl Wieman Science Education Initiative (CWSEI) at the University of British Columbia and the Science Education Initiative at the University of Colorado. He is a Nobel laureate in physics and served as the associate director for science in the White House Office of Science and Technology Policy.



A Better Way to Evaluate Undergraduate Teaching

By Carl Wieman

A major problem in higher education is the lack of a good way to measure the quality of teaching. This makes it very difficult for faculty to objectively determine the quality of their teaching, work systematically to improve it, and document that quality. Institutions in turn are unable to incentivize, track, or demonstrate to external stakeholders improvement in the quality of the teaching that they provide. These deficiencies are becoming increasingly conspicuous in light of the calls for greater accountability in higher education and for the adoption of more effective, research-proven teaching practices, particularly in the areas of science, technology, engineering, and mathematics (the STEM disciplines).

In this paper I examine the context in which teaching evaluation is done at research universities. I then consider the requirements for methods of evaluating teaching quality in higher education and how well current evaluation methods meet those requirements. Finally, I propose a new method based on the notion that the teaching methods used by an instructor are a more accurate proxy for teaching effectiveness than anything else that is practical to measure—a concept that has emerged from the results of STEM education research.

Current Methods of Evaluating Teaching Quality

Definitions and Context

The definition of "teaching quality" is surprisingly vague, with most discussions of that quality focusing on the values and traits of the teacher. I propose a simple and operational definition: the effectiveness with which the teacher is producing the desired learning outcomes for the given student population.

There are a number of outcomes we want for students, but at the level of the individual course or instructor, the degree to which students master the material and complete the course are primary, with attitudes about the subject (appreciation of the general value and intellectual interest of the field, perhaps consideration of it as a career) as important secondary outcomes for many instructors and institutions.

But the attainment of these outcomes in an absolute sense is highly dependent on the backgrounds of the students and the specific outcomes that the instructor defines in the context of a particular course. Thus, meaningful measures of teaching quality must separate out the impact of the teacher from the many other factors that affect the attainment of educational outcomes.

Moreover, how to measure teaching quality depends not only on the definition of quality but also on the goals of the measurement. The usual goal is to encourage and guide the improvement of teaching, but the context in which this goal is to be achieved adds additional constraints.

Here I focus on the particular context of research universities, since these have established much of the culture and standards of teaching that permeate all of higher education. For an institution to encourage improvement, the measures of quality must be integrated into the incentive system. But at research universities, there is a well-established incentive system for individual faculty, departments, and institutions that is based almost entirely on comparisons and rankings of research productivity and prominence.

This system has led to the development of a detailed set of metrics to facilitate such comparisons, such as research funding, publications, measures of the significance of publications, "impact factors," professional-society awards for research contributions, etc. Although the details vary somewhat across disciplines, each has its own well-defined metrics that have widespread if not universal consensus as to their validity.

> "For an institution to encourage improvement, the measures of quality must be integrated into the incentive system.

This incentive system has been very successful. While we often hear that it is impossible to make faculty do anything, in reality nearly all faculty members are doing exactly what the incentive system is telling them they will be rewarded for doing. At research institutions, both well established and aspiring, this incentive system has been very effective at driving research productivity; it has played a large part in establishing the modern research university.

However, as has been frequently noted, the effectiveness of this system means that faculty members and institutions who take time away from research to dedicate to teaching are penalized. This tradeoff is clearly recognized. As I go around the country advocating the adoption of new science teaching practices, faculty invariably raise the concern that this might take time away from their research—and, given the system in which they work, they are right to worry.

Criteria by Which Teaching-Effectiveness Measures Should Be Judged

That said, the imbalance between the weighting of research and teaching in the university incentive system is to some extent due to the fact that the current evaluation of teaching is so inferior to the evaluation of research. If improvement in teaching is to happen, it must become a credible component in the incentive system. And to become credible, its evaluation must meet certain criteria of quality—criteria that the corresponding metrics of research quality already meet.

Validity. The most important criterion is that the measures of teaching quality must correlate with the achievement of the desired student outcomes.

Meaningful comparisons. Individual instructors must have a standard to which they can compare their performance in order to know how well they are doing and what they might do to improve. In particular, they need to be able to compare their performance with the standards of their department and institution. Department chairs must be able to compare the teaching of all the individuals in a department, and deans and provosts need to compare the performance of similar departments at the same or at different institutions.

Fairness. For a method to be fair requires that it can be used across nearly all courses/instructors with nearly equal validity. This will only be true if the correlation between the measured values of "teaching quality" and the measures of student outcomes is greater than the correlation between the measured values of "teaching quality" and other factors that are not under the instructor's control (e. g. class size, level, subject, student preparedness, institutional type).

Practicality. It must be possible to obtain the results of quality measures on an annual basis without requiring substantial investments of time and/or money.

Improvement. The measure needs to provide clear guidance to instructors not only as to how well they are doing but how they can improve.

Although some will object to placing so much importance on comparisons, the reality is that rankings are already

solidly entrenched in the current incentive system in higher education and play a large part in many institutional decisions. To have any relevance, the evaluation of teaching has to become part of that system.

Critique of Current Methods for Evaluating Teaching

I will now examine how well current methods meet these five requirements. In examining the first criterion, validity, it is particularly important to examine how effective the evaluation method is at encouraging the greater adoption of research-based teaching methods that, across hundreds of studies, have shown large and consistent improvements in STEM student outcomes compared to the traditional lecture. Cognitive psychology research on the general attainment of expertise would indicate that similar results are likely for other disciplines for which similar comparative data do not exist.

Berk (2005) discusses the many methods of evaluating college teaching that have been proposed and/or used. The dominant method for evaluating postsecondary teaching is student course/instructor evaluations, which he says are used in 97 percent of departments. A distant second is peer observation by a faculty colleague, followed by teaching portfolios.

Student course evaluations. There are literally thousands of articles on student course evaluations, and within them one can find supporting evidence for all possible opinions. Student evaluations do provide useful information: They are often used to flag instructors who have anomalously low scores in order to understand and address the reasons for the negative student opinions, which is appropriate. However, they have several critical failings with regard to the criteria listed above that should prevent them from serving as the primary method for evaluating the quality of teaching.

First, they have some fundamental limitations that transcend any details. It is impossible for a student (or anyone else) to judge the effectiveness of an instructional practice except by comparing it with others that they have already experienced. If all they have experienced are lectures, they cannot meaningfully evaluate the effectiveness of lectures relative to other practices. This prevents student evaluations from encouraging or rewarding the adoption of more effective research-based teaching methods when such practices are seldom used at an institution.

Steve Pollock [Editor's note: See Pollock's Teachable Moment in the May/June 2014 issue of *Change*] provided a dramatic example of this when he collected Colorado physics majors' opinions as to the effectiveness of clickers and peer discussion in upper-division physics courses. When students were first surveyed, they had not experienced this method of teaching in these types of courses; about 80 percent of them thought it would be a bad idea, while only 10 percent thought it would be desirable. But after the students had been taught a course in which these methods were used, the students' opinions were almost exactly reversed, with about 80 percent in favor and only 10 percent opposed.

There is another basic limitation of student evaluations that ask how much was learned in the course, which most do. People are poor at evaluating their own learning, because it is difficult to know what you do not know. The accuracy of this evaluation is also sensitive to the level of expertise of the respondent.

This is consistent with the findings of the most recent meta-analysis of studies of student evaluations I could find. Clayson (2009) reports that the correlation between student course evaluations and measures of learning (typically grades) has decreased in recent years and is now very low—in fact, negligible in courses where the grading is objective. For all these reasons, student evaluations do a poor job of meeting the first and most important criterion—correlation with desired student educational outcomes.

These shortcomings of student evaluations were confirmed at the Science Education Initiatives at the Universities of Colorado and British Columbia, which changed the teaching methods used in about 150 courses. The student course evaluations for the instructors involved were largely the same before and after they transformed their teaching methods, even though they used quite different teaching methods that usually produced measurably greater amounts of learning. Crouch and Mazur have seen similar results. Student evaluations also fail to meet the criteria for fairness and meaningful comparisons, given the numerous well-established confounding variables that are outside the instructor's control (Pascarella and Terenzini, 2005).

Many researchers who argue for the value of student evaluations do so by showing that, within a limited context, the evaluations correlate with desirable outcomes. But that is not a sufficient condition to be suitable for evaluating an instructor's teaching as a guide for improvement or as part of the incentive system. The correlation with desirable outcomes must hold over a broad range of contexts and courses and be much larger than the correlations with other factors not under the instructor's control for that range of contexts and courses. Student evaluations fall far short of meeting that condition.

To put this in more concrete terms, the data indicate that it would be nearly impossible for a physically unattractive female instructor teaching a large required introductory physics course to receive as high an evaluation as that of an attractive male instructor teaching a small fourth-year elective course for physics majors, regardless of how well either teaches.

> People are poor at evaluating their own learning, because it is difficult to know what you do not know.

Finally, student evaluations provide little guidance for improvement. Although they typically ask many distinct questions about specific aspects of the course and/or the teaching, analysis of the responses to the different questions shows that they are very highly correlated with each other.

This indicates that students are not making distinctions between the questions, so their responses provide little information on specific aspects of the teaching that might be improved. This failing is further compounded by the fact that for a given course or instructor, there is typically a distribution of strongly felt positive and negative student opinions on nearly every aspect of the teaching.

A final concern is that faculty almost universally express great cynicism about student evaluations and about the institutional commitment to teaching quality when student evaluations are the dominant measure of quality. At every institution I visit, this sentiment is voiced, along with the fear that adopting more effective research-based teaching methods will lower student evaluation scores.

Classroom observations. Faculty observations of classes for the evaluation of teaching have a different set of limitations. First, if the faculty observers have never experienced anything but lectures, they are likely to assume that lecturing is the only acceptable teaching method, no matter how much learning it does or does not produce.

Second, a faculty member who is an expert in a discipline will seldom be able to observe a class from the perspective of a learner encountering the subject for the first time. This difficulty that experts have in taking a novice perspective, and their lack of recognition of the learning difficulties faced by novices, is well documented (Ambrose et al., 2010, p. 99). This makes the judgments by faculty observers on pacing, clarity, motivation, and amount of learning achieved inherently suspect.

I have been unable to find any study showing that untrained peer-observation evaluation of teaching correlates with student learning or other outcomes. If the observation and evaluation duties are spread across multiple faculty in a department, as is almost inevitable, the evidence is that the consistency between untrained observers is also likely to be low, raising further concerns about fairness.

There are a number of classroom observation protocols for undergraduate STEM that have been developed and validated. These provide more standardized means of evaluation and hence greater consistency across multiple observers. But nearly all of these, particularly those that have shown correlations with student outcomes, require some days of training to produce reliable results.

For this and other reasons, meaningful evaluation of teaching by faculty peers is quite time consuming. The amount of learning achieved in a course, and hence the effectiveness of the teaching, involves the type and quality of homework, supporting materials, testing and feedback, etc., as well as what happens in class. Thus to properly evaluate the teaching of a colleague, an observer needs time to examine these materials, in addition to some hours of classroom observation and observation training. That much time is unlikely to be dedicated to the evaluation of a single faculty member.

Teaching portfolios. Teaching portfolios usually provide more extensive and informative materials on a faculty member's teaching, but they do not meet any of our criteria. Because they are so time consuming to create and review and so variable in what they contain and how that material is chosen and judged, they are impractical for use as a means of regularly evaluating teaching, making widespread comparisons, and measuring or guiding improvement. Given their variation, it would also be impractical to demonstrate their general correlation with measures of student learning or to compare their results across departments or institutions.

Direct measures of student learning. The most appealing way, in principle, to measure teaching quality is by directly measuring the student learning that is achieved. Unfortunately, this is not practical. To do so requires some validated measure of learning that is independent of the specific instructor, can be used on a pre-post basis to measure change rather than just input-sensitive outcomes, and is a good match to the material and learning objectives of the instructor.

This approach does have the very attractive feature that it provides an objective measure of student learning, which is why such measures have been a mainstay of science education research. However, the necessary instruments for measuring learning exist for only a very small fraction of courses. To meet criterion #2—widespread use in nearly all courses—a very large number of additional instruments would need to be developed. This would require a great deal of time and effort, so this approach to widespread evaluation of teaching will not be practical in the foreseeable future.

In conclusion, all existing methods for evaluating teaching fall far short of meeting the criteria that I have argued are required. In particular, none of these methods encourage the adoption of effective research-based teaching practices.

Several authoritative organizations have reviewed this research and called for the greater adoption of these teaching methods, including the National Research Council, the President's Council of Advisors on Science and Technology, and the American Association of Universities. In spite of the extensive research and these calls for change, polling the leadership of the three major US university associations—AAU, APLU, and AACU—failed to turn up any institution or STEM department that systematically collects data on the teaching practices currently being used in its courses.

A New Method for Gauging Teaching Effectiveness

The Teaching Practices Inventory

Here I offer a different method for evaluating teaching that does meet the above criteria, at least for the STEM disciplines. It should also work for the social sciences with some modest changes, and an analogous instrument could be developed for the humanities.

The design principle used to create this method was to first develop an instrument that could characterize as completely as possible all the teaching practices in use in nearly all STEM courses, while requiring little time and involving little subjective judgment in collecting the necessary information. Knowing the full range of teaching practices used in any given course, it is then possible to determine how often practices that research has shown consistently produce improvements in student outcomes are used when compared to possible alternatives.

The quantitative measure of the extent to which practices that correlate with improved student outcomes are used is our measure of teaching quality. Essentially, this is a backwards design to meet the five requirements listed above.

It may seem surprising to evaluate the quality of teaching by looking only at the practices used by an instructor. However, measuring practices as a proxy for difficult-to-measure ultimate outcomes is quite common when there are substantial correlations between the two.

The example most relevant to this discussion is the routine measurement of a science faculty member's research contributions for annual review. In the natural sciences, this is typically based primarily on the numbers of papers published in reputable journals and the research grants that a faculty member has had in the past one to three years, data that can be quickly and easily collected.

This system works quite well because, while having a relatively large number of grants and publications does not guarantee substantial research contributions, they tend to be well correlated. Correspondingly, a faculty member in the sciences who does not get research money and does not publish papers is very unlikely to be making significant research contributions. Using effective, research-based teaching practices as a proxy for the desired student outcomes is based on much the same concept.

This use of such a proxy is only meaningful because all the research in the past few decades has established strong correlations between the type of STEM teaching practices used and both the amount of student learning achieved and course completion rates. These correlations have been shown to hold across a large range of different instructors and institutions.

Those practices that are linked to improved learning in STEM are also consistent with empirically grounded theoretical principles for the acquisition of complex expertise. This explains the consistency of the results across disciplines, topics, and level of students, and it provides confidence that those practices will be similarly effective in situations for which there is yet no research.

The teaching practices inventory characterizes all the teaching elements of a course. The current version of the inventory was developed over a six-year period, during which it underwent numerous reviews by faculty and experts in undergraduate STEM education and several rounds of real-world testing (this is discussed in detail in Wieman and Gilbert [2014]).

Sarah Gilbert and I have now used the final version of the inventory to collect data on the teaching of more than 200 courses at the University of British Columbia in biology, computer science, earth sciences, mathematics, physics, and statistics. It has also recently been used on a limited basis by several other institutions. The inventory can be completed in less than ten minutes.

Table 1. Teaching Practices Inventory Categories

I. Course information provided *Information about the course, such as a list of the topics and organization of the course and learning goals/objectives*
II. Supporting materials provided *Materials that support learning of the course content, such as notes, videos, and targeted references or readings*
III. In-class features and activities *What is done in the classroom, including different types of activities that the instructor might do or have the students do*
IV. Assignments *The nature and frequency of homework assignments in the course*
V. Feedback and testing *Testing and grading in the course, as well as the feedback from instructor to students and from students to instructor*
VI. Other *Assorted items covering diagnostics, assessment, new methods, and student choice and reflection*
VII. The training and guidance of teaching assistants *The selection criteria and training used for course teaching assistants and how their efforts are coordinated with other aspects of the course*
VIII. Collaboration *Collaboration with other faculty, use of relevant education research literature, and use of educational materials from other sources*

> The inventory responses provide a detailed picture of how a particular course is taught.

We have always had the faculty member teaching the course fill out the inventory. Self-reporting takes less time than having another person do it, and filling out the inventory encourages reflection by faculty members on their teaching.

To reduce the risks of errors, the great majority of the responses to the items on the inventory are objective, with checkboxes or choices of numbers within ranges that simply reflect whether and how frequently something was done. Most of this information can be easily determined by looking at the course materials. Thus, a third party can complete the inventory for a course or check the faculty member's responses, if an institution felt that was needed.

The items for which that kind of objectivity is difficult are in-class activities. This is also the area where it is most difficult for a faculty member to remember accurately. To address this, we have developed the COPUS classroom observation protocol (M. Smith et al, 2013), which allows observers with little training to characterize how the instructor and the students are spending their time in class.

Scoring Rubric

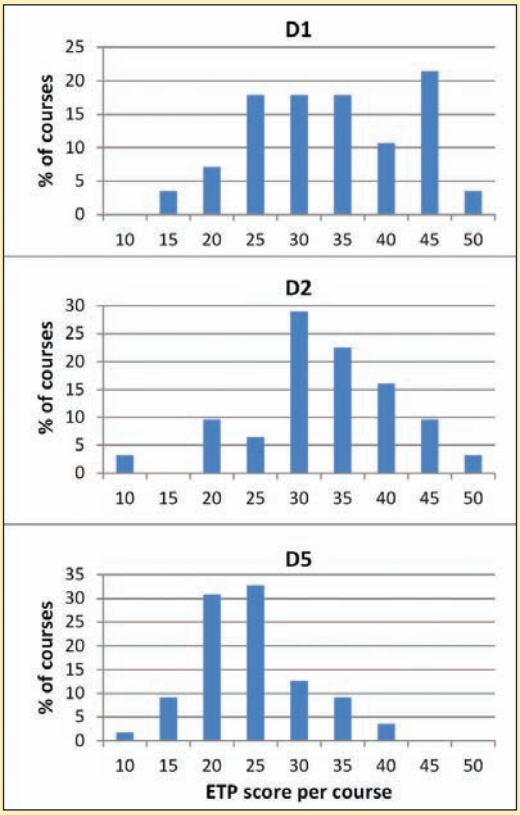
The inventory responses provide a detailed picture of how a particular course is taught, and—when data are aggregated—how a department teaches. We have created a rubric that converts the raw inventory data for a course into a measure of the extent to which practices that research has shown are most educationally effective are used. This becomes the "ETP score" ("extent of use of research-based teaching practices") for each of the eight inventory categories and for the course as a whole.

ETP points are given for each practice for which there is research showing that the practice improves learning. The distribution of points is shown on the inventory in Wieman and Gilbert (2014), along with references to the research that is the basis of the scoring.

Figure 1 contains histograms of the scores for the courses in three math and science departments of the five for which we collected extensive data and tested the inventory (from Wieman and Gilbert 2014). For instance, in D (for "department")1, the ETP scores for courses ranged from 13 to 51. The height of each bar represents the fraction of courses that have ETP scores within a range of plus or minus two around the value shown beneath the bar.

Figure 1 illustrates the ability of the inventory to identify the extent to which effective teaching practices are used by

Figure 1. Distribution of ETP Scores Across Three Different Math and Science Departments



the different faculty within a department, as well as the differences between departments. It is startling to see the range of use of effective teaching practices within a typical department—factors of 4 to 5. High-scoring courses incorporate many different, mutually beneficial practices across all categories that support and encourage student learning, while low-scoring courses have very few.

Table 2 shows the average and standard deviation (AVE [S.D.]), enrollment-weighted average (EWA), and category averages of ETP scores for three departments during one semester. The "enrollment-weighted average" is calculated by weighting the ETP score for each course by its respective enrollment. For instance, Department 1 has 28 courses with an average ETP score of 33.4—which, when considering the number of students affected, is adjusted to 39.3, because a number of the courses with quite high ETP scores have large enrollments. The columns to the right also show the average value in each of the eight inventory categories for these courses.

The sample shown in Table 3 illustrates the usefulness of inventory results for the improvement of teaching. From this table, it is easy to identify the strengths and weaknesses of the teaching of particular courses and which ones fall below departmental norms. In this case, the most effectively taught course is #10 and the least effectively taught is #2.

Figure 2 compares percent of the total possible ETP score in the courses taught in Department 3 for two different academic years.

Table 2. ETP Scores

Department	N	AVE (S.D.)	EWA	I	II	III	IV	V	VI	VII	VIII
D1	28	33.4 (9.4)	39.3	3.9	4.2	7.8	3.2	7.5	2.3	1.6	2.9
D2	31	32.6 (8.5)	33.6	3.7	4.5	6.1	3.3	8.1	1.6	2.3	2.9
D5	55	24.1 (6.5)	25.2	2.7	3.1	4.0	2.1	8.3	0.7	1.1	2.1
max possible		**67**		**6**	**7**	**15**	**6**	**13**	**10**	**4**	**6**

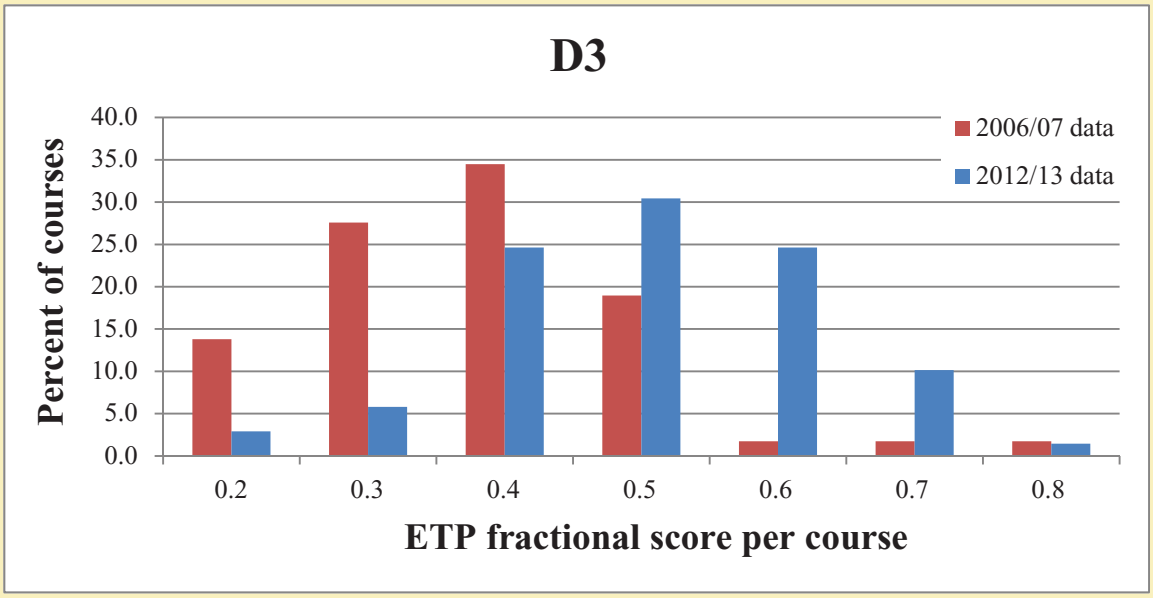
Showing average and standard deviation and enrollment weighted average for the courses in each of the three listed departments. The average ETP Score for the departments in each of the eight inventory categories described above is shown in the columns to the right.

Table 3. Category Scores

Course #	ETP total	I information provided	II supporting materials	III in-class activities	IV assignments	V feedback	VI other	VII TAs	VIII collaboration
1	28	4	4	3	4	7	1	0	5
2	10	0	1	2	2	2	1	0	2
3	35	4	3	9	2	6	4	4	3
4	33	4	4	7	4	9	0	4	1
5	32	4	5	9	0	8	1	3	2
6	22	3	2	3	1	6	0	3	4
7	18	1	4	4	2	7	0	0	0
8	32	4	5	6	3	8	2	2	2
9	19	1	4	0	2	9	1	2	0
10	46	6	6	12	4	11	1	2	4

Representative set of 10 courses from a single department.

Figure 2. Distribution of ETP Scores



Department D3 for the 2006/7 and 2012/3 academic years

Figure is from Wieman and Gilbert (2014)

> The general norms for comments and discussions posted online tend towards extreme stances and often uncivil language, and these norms have carried over to some degree to online student evaluations.

As you can see from the figure, teaching practices improved measurably in the department as it worked on improving its teaching between 2007 and 2012.

Using the Inventory

Benefits

Instructors, institutions, and students all benefit from the use of the inventory for evaluating teaching. Simply by looking at the inventory and its scoring rubric, faculty can see the range of teaching practices that are in relatively common use and what the research indicates as to which of them increase student learning, as well as how they can improve their teaching and objectively document that improvement.

Comparing their own inventory results with those of their colleagues shows them their respective strengths and weaknesses. By looking for effective practices that are the norm in the department, they can even determine which are likely to be the easiest to adopt and identify faculty who could help them use a particular practice. The inventory can free them from the capricious, frustrating, and sometimes quite mean-spirited tyranny of student evaluations—the negatives of which appear to be getting worse as institutions adopt online evaluation systems. This is likely because the general norms for comments and discussions posted online tend towards extreme stances and often uncivil language, and these norms have carried over to some degree to online student evaluations.

Departments can use inventory data to benchmark the quality of their teaching and identify both targets for improvement and the results of improvement efforts, as shown in Figure 2. It also takes little time to obtain data that can be used for institutional or accreditation reviews. Institutions can use inventory results in the same way to track their overall improvement and to compare the quality of their teaching with that of peer institutions, just as they now routinely compare research productivity and salaries across institutions.

Inventory data are also useful to students, who currently have no meaningful information by which to select courses, departments, or institutions that will provide them with the highest-quality teaching. If inventory data were available to students, they could make better-informed educational decisions and reduce the chances that they will have the all-too-common experience of encountering very poorly taught courses that have career- and life-changing consequences.

Limitations and Concerns

We have only tested this inventory in math and science courses, but we believe that it would be suitable for use in engineering and probably in the social sciences, particularly if the primary criterion is that it be superior to existing methods of evaluating teaching. Data support the general validity of the scoring rubric across this full range of disciplines, although with less specificity as to the size of the impact particular practices have. Some small changes may be needed to ensure that it adequately captures common teaching practices used in these disciplines and that the wording is readily understandable to the faculty.

But we were not successful in creating an inventory that was appropriate for use in all undergraduate STEM courses: It does not work for instructional labs, project-based courses, or seminars (which are largely student driven). We found these courses so idiosyncratic in their goals and practices that we could not characterize them in a meaningful way or evaluate their relative educational effectiveness. It is likely that there will always be a few such courses that are impossible to evaluate by any conventional measure.

Of possible concern is the reliance on faculty self-reports (a problem also with faculty self-reports on research productivity)—even though, as noted above, the inventory is designed to maximize their accuracy, as well as to allow them to be independently verified. As discussed in Wieman and Gilbert (2014), the inventory responses we have checked have been accurate, but there were no stakes attached to those responses.

The most obvious concern with the inventory data and scoring rubric is that they measure the use of particular practices, not how well those practices are being used. But the important comparison here is not with perfection, but rather with alternative methods of evaluation.

There are numerous studies showing a strong correlation between undergraduate STEM learning and the teaching methods used, independent of other characteristics of the teaching. For example, many studies have compared the same instructor using different methods (research-based

> The inventory data and scoring rubric…measure the use of particular practices, not how well those practices are being used.

active learning vs. traditional lecture). The measured differences in student outcomes with different teaching methods are usually larger than the differences in outcomes across the range of different instructors using the same method.

Looking ahead, as the use of these effective teaching practices becomes more common, researchers will be able to identify which details of the implementation of each practice impact student learning and by how much; this information can then be incorporated into updated versions of the inventory and scoring rubric to capture those details. Similarly, this future research will allow the relative weighting of items in the scoring rubric to be refined.

We have investigated the evaluation of the implementation quality of specific teaching methods across a range of courses by having a group of trained and experienced experts in science teaching attempt to carry out such assessments across a range of courses in the discipline in which they were expert. They found that doing evaluations of quality with reasonable accuracy requires familiarity with the course content, the teaching methods used, and the student population in the particular class, as well as a high level of expertise at observing and critiquing teaching. This is a combination of expertise that they didn't have, nor would anyone in a regular department, for the full range of courses to be evaluated. Thus it is doubtful that it will ever be practical to directly measure the quality of implementation on a widespread basis.

Conclusion

Current methods of evaluating teaching in colleges and universities fail to encourage, guide, or document teaching that leads to improved student learning outcomes. Here we present a measure of teaching quality that is correlated with desired student outcomes, free from large confounding variables, and that can be used to compare the quality of teaching between faculty members, departments, and institutions. In contrast with current methods for evaluating undergraduate teaching, it is appropriate to incorporate the inventory into the institutional incentive system. It also provides a way for faculty members, departments, and institutions to support claims of teaching commitment, improvement, and quality. C

Resources

- Ambrose, S., Bridges, M., DiPietro, M., Lovett, M., & Norman, M. (2010). *How learning works: Seven research-based principles for smart teaching.* San Francisco, CA: John Wiley & Sons.
- Berk, R. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17,* 48–62.
- Clayson, D. (2009). Student evaluations of teaching: Are they related to what students learn?: A meta-analysis and review of the literature. *Journal of Marketing Education, 31*(1), 16–29.
- Freeman, S., Eddy, S.L., McDonough, M., Smith, M.K., Wenderoth, M.P., Okoroafor, N., & Jordt, H. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America, 111*(23), 8410–8415. Doi: 10.1073/pnas.1319030111.
- Pascarella, E., & Terenzini, P. (2005). *How college affects students: A third decade of research.* San Francisco, CA: Jossey Bass.
- Singer, S., Nielsen, N., & Schweingruber, H. (2012). *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering.* Washington, DC: National Academies Press.
- Smith, M.K., Jones, F.H., Gilbert, S.L., & Wieman, C.E. (2013). The classroom observation protocol for undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *CBE Life Sci Educ, 12,* 618–627.
- Wieman, C. & Gilbert, S.L. (2014). The teaching practices inventory: A new tool for characterizing college and university teaching in mathematics and science. *CBE- Life Sciences Education, 13,* 552–569.