don't allow for the large cost overruns that typically occur on DOE nuclear facility construction.

Other supply options

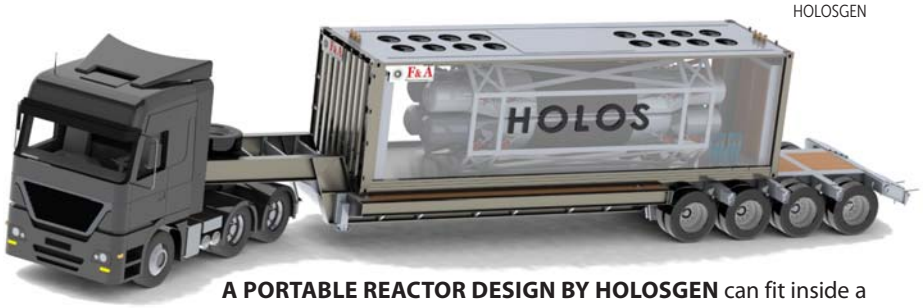
In a report last year, DOE's Nuclear Energy Advisory Committee (NEAC) was sharply critical of the agency for reallocating $23 million in FY 2019 funding from the nuclear energy research program to pay for the Centrus plant. That decision left only $10.6 million for academic research on the nuclear fuel cycle last year. The funding cut occurred just prior to the due date for academic research proposals, after "massive efforts" had been expended on proposal preparation, the committee said.

The report said other "very promising routes" could provide HALEU. DOE has set aside a portion of its surplus of HEU to be diluted to HALEU. By 2023 the agency plans to recover another 5 tons of HALEU in spent fuel from a decommissioned experimental breeder reactor in Idaho. And Congress appropriated $20 million in FY 2019 to recover HEU from spent naval reactor fuel stored in Idaho for potential conversion to HALEU. The NEAC report also suggested that 34 tons of surplus weapons-grade plutonium slated for disposal under a bilateral agreement with Russia could be diluted to provide the equivalent of 170 tons or more of US-origin HALEU. DOE plans instead to render the plutonium unusable and store it permanently at its Waste Isolation Pilot Plant in New Mexico.

Terms of the award call for Centrus to match 20% of the federal funding, the lawmakers' letter stated, in apparent contravention of 2005 legislation that requires no less than a 50–50 share.

David Kramer

HOLOSGEN



A PORTABLE REACTOR DESIGN BY HOLOSGEN can fit inside a standard shipping container and generates 13 MW of power for 12–20 years. The reactor would use high-assay, low-enriched uranium containing up to 15% ^{235}U.

Reevaluating teacher evaluations in higher education

Relying on students to rate professors is convenient, cheap, and problematic.

For decades student evaluations have been the mainstay of attempts to measure the quality of teaching at colleges and universities across the US and beyond. Now, as part of a growing focus on teaching in higher education, and because of mounting evidence of student biases, those evaluations are increasingly in the crosshairs. A smattering of institutions have begun revamping their approaches to student evaluations of teaching (SETs), and those independent efforts are fueling momentum on a national scale.

SETs have become the norm in higher education because they are convenient and cheap. The questions and scoring vary by discipline and institution, but typically before they see their final grade, students are asked to fill out a survey about the course and the instructor. Department heads or other campus officials calculate averages and often compare a given teacher's ratings to others' in the department and across the institution. The ratings inform promotion and tenure

DOUGLAS WARD/UNIVERSITY OF KANSAS

FACULTY MEMBERS BRAINSTORMED how to improve teaching and teaching evaluations last April in Boulder. Gabriela Weaver (in turquoise), Ann Austin (in purple), and Noah Finkelstein (with fingers at the board) are principal investigators on a multi-institutional, cross-disciplinary project that looks at teaching effectiveness.

decisions and are often the deciding factor in renewing teaching contracts for instructors who are not on the tenure track (see PHYSICS TODAY, November 2018, page 22).

The trouble is in what the ratings say—or don't. In 2009 the faculty union at Ryerson University in Toronto filed a grievance with the university over SETs being an unfair measure of teaching effectiveness. Last year, an arbitrator ruled in the faculty's favor: Student evaluations at Ryerson can no longer be used to assess teaching effectiveness for high-stakes decisions such as tenure and promotion.

The case could prove to be a harbinger. Traditional SETs will become illegal, predicts Carl Wieman, a physics Nobel laureate at Stanford University and a leader in science, technology, engineering, and math (STEM) education studies. "It will be hard for an institution to say they are still collecting SETs but not using them in tenure and promotion decisions," he says. University of California, Berkeley, statistics professor Philip Stark, who was an expert witness in the Ryerson case, says class-action suits are in the works. "SETs don't measure teaching effectiveness; you can't make a course better with the information that comes in. They are biased. There are all sorts of problems."

"Garbage in, garbage out"

Study after study has shown that SET responses are biased. In physics, female instructors are often rated 7–13% lower than males, notes physicist Noah Finkelstein, codirector of the Center for STEM learning at the University of Colorado Boulder (CU). Similar patterns are observed in other STEM fields. The degree of disparity varies by discipline, course, level, institution, and other factors, but across the board SETs penalize women, underrepresented minorities, nonnative English speakers, and older and physically less attractive instructors of both sexes. SET ratings are affected by the condition of the classroom, the time of day a course takes place, and other things that are outside the instructor's control, says Stark. The strongest correlation with high ratings is expectations, he adds. "If students go in thinking they will get a good grade, they give higher evaluations."

Most traditional SETs include broad questions like, "How would you rate the quality of the course overall?" and "How would you rate the quality of the instructor overall?" Such questions are coming under increasing criticism because the responses are frequently biased and unactionable—instructors don't glean ideas about how to improve their teaching. Some responses are even abusive. "That type of question offers up a vacuum to fill," says Richard Taylor, physics chair at the University of Oregon, "and encourages whatever biases students have, implicit or explicit."

Students have written, for example, "the teacher is a crybaby," and "I would rather watch my mother's head be cut off and her hair used to mop up the blood than take another class with [instructor's name]." Such comments take an emotional toll, the instructors who received them say. They also note that instructors can feel pressured to inflate grades in a bid for better ratings. (These two examples are from large STEM classes at a research-intensive university; the instructors requested anonymity because of concern about renewing their contracts.) Some departments remove incendiary comments before the instructor sees them.

Even specific questions are often misguided, argues Stark. Students are not the right people to ask about the effectiveness of a course or whether an instructor fostered an atmosphere that is consistent with campus goals for inclusion, he says. "They can't judge that. I've seen questions on whether the instructor has mastery of the material. How on Earth would a student know that?" Finkelstein agrees: "We are asking students the wrong questions and using the data badly."

Most institutions employ a numerical rating system, and it's common to evaluate teaching based only on the broad questions; some research-intensive universities ignore teaching altogether in evaluating faculty. The numerical rating approach itself is flawed. For starters, notes Wieman, students have a tendency to go down the list and check off the same score for every question. And, says Stark, "averages of categorical material are meaningless and misleading. Reporting distributions would be preferable."

Arguments about the numbers were a big part of the Ryerson grievance case. "Things went downhill when the surveys went online," says Sophie Quigley,

STUDENT EXPERIENCE SURVEY, SAMPLE QUESTIONS

(provided by the University of Oregon Office of the Provost)

The inclusiveness of this course is:

- ☐ Beneficial to my learning
- ☐ Neutral
- ☐ Needs improvement to help my learning

The opportunities for student interaction in this class are:

- ☐ Beneficial to my learning
- ☐ Neutral
- ☐ Needs improvement to help my learning

The clarity of assignment instructions and grading is:

- ☐ Beneficial to my learning
- ☐ Neutral
- ☐ Needs improvement to help my learning

The degree to which the course includes active learning is:

- ☐ Beneficial to my learning
- ☐ Neutral
- ☐ Needs improvement to help my learning

the computer science professor who filed the case on behalf of the faculty union. The university began dicing the numbers in new ways, she says. "The math was bad." For example, in some cases the averages were not even calculated properly. What's more, she notes, the student response rate took a dive, and those students who chose to respond are self selected, and may be motivated by disgruntlement with the course or instructor. Says Quigley, "It was garbage in, garbage out."

Ideally teacher evaluations could be used both for students to give voice to their opinions and for teachers to improve their teaching. "But the data don't correlate with anything you care about," says Wieman, "not learning, teaching, or good teaching methods." And, he adds, the SETs make faculty afraid to switch to more innovative teaching methods because the evidence shows that student ratings initially drop when instructors try new approaches. "Everybody knows SETs don't have validity, but they are the only evaluation people have."

Fairer approaches

Concerns about measuring teaching effectiveness, improving teaching, and mitigating bias are prompting institutions to rethink their approach to evaluating instructors. Academia has more robust ways to evaluate faculty members' research activities, including grants obtained, papers published, PhD students graduated, invited talks, and the like, says Gabriela Weaver, a chemistry professor and special assistant to the provost for educational initiatives at the University of Massachusetts Amherst. "If we want to measure how people teach, the measure should correlate with student learning," she says.

With three principal investigators at other institutions, including CU's Finkelstein, Weaver is conducting a cross-disciplinary project to test different approaches to measuring teaching effectiveness. In the NSF-funded project, called "Transforming Higher Education—Multidimensional Evaluation of Teaching," three US university campuses are implementing and studying variations on three-part teacher evaluations—the student voice, peer evaluation, and self-reflection by instructors. "We want to create a more holistic system," Weaver says.

STUDENTS ENGAGE IN ACTIVE LEARNING in an introductory physics class taught by Eric Cornell at the University of Colorado Boulder.

Meanwhile, a handful of universities have begun introducing similar approaches on their own. The University of Southern California revamped its student evaluation procedures in 2018 as part of an initiative on teaching excellence. The macro changes at USC include introducing a university-level definition of teaching excellence and new infrastructure to develop and reward it "in serious and tangible ways," says Ginger Clark, associate vice provost for faculty and academic affairs and the director of the university's Center for Excellence in Teaching. Individual departments customize their approach to the best teaching practices in their own disciplines, she adds. The university uses peer review as its primary tool for evaluating teaching, but it incorporates self-evaluations and student surveys. And the surveys, instead of focusing on the course and instructor, now hone in on the student's own experience. "Students are not trained in pedagogy, but we had been using them as our experts," says Clark. "If we are honest about teaching, we need to know what we are measuring."

Called "student learning experience evaluations," USC's new student surveys do away with global questions. Instead, they pose such questions as whether course concepts were well explained, whether the instructor encouraged discussion, whether the instructor was receptive to diverse viewpoints, and whether the criteria for the class were clear. Students are also asked how much time they spent on homework, how often they interacted with the instructor outside class, and how they participated in learning for the course. Other questions on similar surveys around the country ask whether the instructor's handwriting

GLENN ASAKAWA/UNIVERSITY OF COLORADO

is legible, whether the student could hear the instructor, and whether the student understood the textbook.

"We were concerned about bias, and also about random nastiness that didn't seem warranted," says Michael Dennin, a physicist who serves as vice provost for teaching and learning at the University of California, Irvine (UCI). The university put into practice a long-ignored policy to require additional evidence about teaching, and has revamped the surveys. "We are consistent with the national focus," Dennin says, "which is to move toward language that asks the students to assess experience in the classroom rather than to directly assess the professor." It's still early, he adds, but the shift seems to reduce bias. The UCI student experience surveys have replaced numerical ratings with categories from "strongly agree" through "strongly disagree" because psychology studies suggest that people give more thought to questions when so formulated.

The University of Oregon introduced a campus-wide overhaul to teacher evaluations this past fall. It replaced traditional SETs with self-reflection, peer review, and student feedback. As is the case at other universities at the vanguard of tuning their teacher evaluations, the questions are now designed to reflect student experience, and students fill out surveys a few weeks into a term and again at the end. The midterm feedback is seen only by instructors, says physics chair Taylor, and it can be helpful for adjusting one's teaching. The survey responses are no longer numerical ratings, and students are asked to single out something that was especially helpful and something that they would like to see changed. "We've made a complete mental model shift," says Sierra Dawson, the university's associate vice provost for academic affairs.

Burdens, rewards, and support

Using peer review to evaluate instructors is controversial. Proponents assert that with minimal training, faculty members can learn to evaluate their peers fairly and usefully, and that doing so can be a rotating service duty. And, they say, observing other instructors can be helpful for improving one's own teaching. But critics point out that no consensus exists on what makes up good teaching and that evaluators need to know the course material and be familiar with the student population to gauge level and pace. A further complication is that course preparation, office hours, mentoring, and other aspects of teaching occur outside the classroom and so far have been left out of teacher evaluations, which consider mainly lecturing.

Physics education researchers have identified practices that lead to better student outcomes in physics and "seem to be similarly effective" across fields in STEM and even the social sciences, says Wieman. Based on that research, he advocates collecting data on the practices instructors use in the classroom. He developed a rubric, which, he says, "is informative, and gives a proxy for measuring teaching effectiveness." Doing so, he says, works for large and small classes and avoids bias. In a spinoff of Wieman's approach, some departments keep track of activities in a classroom. "We send trained undergraduates into a class to note what's happening every two minutes," says CU's Finkelstein. This is not peer review, he notes, but is meant to complement other sources as a measure of what is going on in class.

University administrators recognize that any change can be difficult to implement and that, for example, expanding peer review of teaching may be seen by faculty members as a chore. "Our job is to improve teaching without taking away from research," says UCI's Dennin. In the past, he adds, negative teacher ratings have been "very relevant, but if you made the bar you were fine. And being great didn't give you a boost." Emily Miller, associate vice president for policy at the Association of American Universities, says that placing increased attention on teaching does not hurt research productivity. Researchers who are working to improve teaching are "as effective at getting grants and research dollars as they were before," she says. "And they may become more competitive for getting top graduate students."

Still, it's not just reward and punishment; the other piece for getting buy-in is offering support. To that end, UCI and other campuses offer assistance to departments and instructors with self-evaluations, definitions of excellent teaching, and more.

In 2018 the National Academies of Sciences, Engineering, and Medicine launched an ongoing roundtable on systemic change in undergraduate STEM education. "Radical things are happening on the landscape of higher education, with new technologies, changing student demographics, new models of certification for jobs," says Heidi Schweingruber, director of the National Academies board on science education. The roundtable is looking at how to catalyze improvements in instruction. "There will be implications for tenure and promotion, but we haven't gone deeply into it yet." One thing that has become clear from the roundtable, she adds, is that evaluations of teaching are a "potential lever for change."

Change is always stressful, says Taylor, but pilot studies in a few University of Oregon departments suggest that the new approach will be beneficial to instructors. "There is angst because people are unsettled." Within a couple of years, he says, the new three-pronged evaluation system "will become the norm, and it will be a better norm."

Toni Feder PT