

Report to The Teagle Foundation

on a Listening on Assessment

Peter T. Struck

February 1, 2007

On January 18, 2007, at the American Academy of Arts and Sciences in Cambridge, Massachusetts, the Teagle Foundation convened about 20 administrators, faculty, and graduate students from prestigious research I universities for an open discussion on the uses of assessment—systematic measuring to identify specific learning gains—to invigorate undergraduate teaching. The meeting carried forward to the research universities an investigation that Teagle has been developing in other Listenings with liberal arts colleges.

CALL FOR ACTION

The stage for discussion was set by a presentation that amplified and elaborated themes drawn from Derek Bok's *Our Underachieving Colleges*. Participants heard reasons why now is the right time to make a priority out of taking more deliberate steps measurably to improve undergraduate learning. First, we are learning more from the fields of cognitive science and the social sciences about how people learn. Second, we now have better assessment tools to measure what people have learned and meaningfully compare results of pedagogical experimentation. Third, the world of higher education is facing outside pressure, that could become intervention, from accreditors, state agencies, and from the federal government, to do just this kind of thing. Finally, as a large and socially responsible institution, higher education stands to improve its stewardship of public and private resources by experimenting more systematically to identify strengths and weaknesses. The case was also made that this embrace of new efforts to improve undergraduate learning was inevitable because the outside pressure is serious, but more importantly because of what was identified as a deep and disturbing inconsistency, regarding the topic of undergraduate education, between faculties' deepest beliefs, aims and aspirations, and how faculties currently act. It was suggested that this is an inherently unstable situation and once it is pointed out, it will correct itself.

PEDAGOGY OR ACCOUNTABILITY?

Participants quickly identified as salient a binary theme that has been part of similar discussions in other settings: assessment motivated by pedagogy on the one hand, and by accountability on the other. First, assessment as an internal academic conversation tends

to merge quickly with a discussion of teaching—a version of the assessment discussion that was energetically engaged by participants throughout the day and produced most of the recorded observations that follow. The conversation was propelled by consideration of ways to encourage better teaching, with the topic of assessment as a subset of that broader issue. Many specific examples of pedagogical innovations were adduced as well as institutional strategies to implement them successfully.

Assessment as a public policy topic, on the other hand, grows out of a discourse of "accountability." The conversation about accountability tends to come from outside the academy and its strongest voices may or may not be perceived as benevolent from within. By design, the discussion of this Listening spent little time directed specifically toward the notion of accountability itself, but a set of ideas that might be said to lie within its penumbra did come up, and it is useful to make them explicit. Participants' vigorous embrace of the conversation about pedagogy had its mirror image in a tepid response to notions and language that instrumentalized students and teachers and that attempted to quantify an enrichment that many will be resolutely committed to seeing as deeply qualitative. If one endorsed the idea that, say, a truly successful liberal arts education is transformative or inspires wonder, the language of inputs and outputs and "value added" leaves one cold. Such language may ring more to the tune of "workforce development" than of education, and participants voiced concern that in some important sense it might miss the point.

While the dichotomy between accountability and pedagogy is, in a pure sense, a false one, in the give and take of the larger public conversation about higher education reform, the two motivations to measure learning may run afoul of one another. One participant, for example, spoke of the need cautiously to balance attention to the goal of accountability enforced by rigorous assessment methods with the goal of encouraging an open environment of pedagogical discussion, experimentation, and innovation. Running forward with the first will have the effect of suppressing the second.

WHOSE TEST IS IT?

A prominent sub-theme, which might be called the ownership question, emerged over the course of the day. For whose use is the information of an assessment being gathered? Two prominent currently available assessments take as their unit of analysis the institution as a whole. According to their standard protocols, NSSE and the CLA both aggregate results and prepare them for a given institution's central administration. While such an approach has the benefit of spurring change at the top, it does not produce similar ferment at the level of individual students and professors. (This rule was proved by an exception. When one university handed the CLA over to departments to self-assess their effectiveness in developing critical thinking, they were spurred to innovation and curricular reform. Their different approaches are shared to encourage experimentation and innovation.)

A common theme in the discussions was a kind of centrifugal urge to move the stakes and claims of assessment away from the center, and out to the capillary level. The theme

could, nearly without remainder, be mapped onto the discussion of accountability vs. pedagogy: more centralized ownership of results correlating with a stronger accountability urge and more local results with a pedagogical one. Participants expressed interest in initiatives that would encourage individual instructors to engage in what were termed by one participant as "low-stakes / high yield" assessment. This formulation references frequent assessments that maximize useful pedagogical information (for both students and faculty) and at the same introduce minimal intrusiveness into the learning environment.

In addition to the greater promise for learning gains, such a strategy also had the best chance of practical success. It was pointed out that faculty members are, after all, human beings and as such will be no more quick to embrace self-scrutiny than anyone else, and tenured faculty at prestigious universities may even be less quick. This is especially true if change is forced from the top. On the other hand, since faculty, one assumes, are by nature curious and disposed to tinkering, participants predicted professors would be likely to embrace initiatives if they could truly become a kind of research problem of how to teach more effectively.

This discussion of the importance of centrifugal initiatives brought up an interesting conundrum that peppered the diverse conversations of the day. Top administrators understand well that their faculties have a certain disposition to resist imposition of large scale initiatives from the top, and that such initiatives will risk being seen as bureaucratic paperwork, of which, all agreed, there is already enough. Fewer doubts were raised by participants about attempts (from the top) to nurture bottom-up, small-scale initiatives among the faculty, provided these focused on pedagogy. The question then becomes what are the most effective ways to nurture bottom-up change?

CURRENT STATE OF TEACHING ON CAMPUSES

While not all participants embraced the notion of a crisis, no one at the meeting argued that the current state of undergraduate education was entirely acceptable. Against this background, participants located many opportunities for improvement. From a broad perspective, they seemed to share a sense that while their institutions had effectively produced messages strongly supportive of research, they could strengthen the institutional structures that endorsed strong teaching. Several participants pointed out ways that their institutions were unintentionally sending infelicitous messages that reduced the value of teaching. Two specific examples came up. The common practice of offering time out of the classroom as a reward introduces some difficulty, and even the phrase itself of "teaching relief," embedded in our discourse and a common part the incentive structure for the most highly-valued faculty, has a pernicious character (a quality that becomes clearer when trying to imagine the inverse award of "research relief"). Second, most participants acknowledged that much more could be done with centers for teaching and learning. Irrespective of the merits of their services and activities, participants expressed concern that faculty still perceived these centers as sites of remediation.

Graduate students, perhaps the most astute readers of such institutional messages, confirmed that it quickly became obvious to them that successful academics looked for ways to avoid teaching. Graduate students further observed that, in sharp contrast to their research, teaching requirements focused on quantity, not quality of teaching. They all mentioned that a lack of systematic teacher training (it is currently "like triage," according to one participant) stood in sharp contrast to the detailed and careful tasking their mentors placed on their development as researchers, and reinforced a sense of teaching as a second-tier activity.

WHAT WE ARE LEARNING ABOUT LEARNING

As a kind of proof of concept, the participants were treated to a concise summary of recent specific gains in cognitive science and their provocative relevance to methods of teaching. Advances were shared in four main areas of research: retention, context specificity, implicit learning, and stereotyping bias stigma.

For reasons that have no relation to long-term educational goals, we test students at the end of a learning cycle and give little concern to whether students retain anything over the longer term, even six months down the road. For most undergraduates, who will not go on in our fields, loss of content is profound. Cognitive scientists are now coming to understand the differences between the strengths of our memories' encoding and retrieval powers that reinforce the case for active learning teaching. We could also do much more to be attentive to longer-term educational outcomes. Cognitive scientists have found that it is actually much harder to generalize from a context than we tend to think. So the truism about teaching concepts with concrete examples to flesh them out is indeed true. But "context" can refer not just to context regarding content, but also regarding context in time, and here a counter-intuitive result emerges. Teachers typically see advantages accrue to teaching one concept at a time, drilled to mastery via examples before moving to the next concept. This procedure, known as a "block trial," does indeed produce short term gains (of the kind that show up on midterms) but it also produces meager long term retention. Conversely, a teaching approach that uses a "mixed trial," where teachers disaggregate their materials and interleave multiple topics, while less efficient in the short term, produces significant long-term retention gains. This research suggests that when teachers engage in blocking, because of its identifiable short-term benefit, it is actually working against their students longer term learning. Blocked learning disappears, interleaved learning sticks.

Cognitive scientists have also discovered that most of what we know we are not conscious of and are not able to articulate. For understandable reasons, teachers tend to focus on what people can say about what they know—explicit learning. But we are missing out on a huge and consequential area of knowledge, which cognitive scientists have named "implicit learning." This kind of learning has to do with skills and habits of mind. Physical skills like bike-riding provide a useful analogy. Such knowledge is internalized seamlessly into our thinking and steers our intellects in important ways. We have hardly begun to figure out ways of using a classroom environment to instill faculties and capacities into the implicit system. As it stands we are ceding this territory to other

broad cultural spheres, like pop culture and television, which are extremely successful at encoding our implicit knowledge.

Many educators also stand to gain from a fuller appreciation of the ways in which stereotyping bias and stigma are operative in the classroom. A few examples of recent research made the point. In one now well-known investigation, young women of Asian ethnicity were given tests in math. Before each test, they filled out questionnaires that asked innocuous questions, but that made either their ethnicity or their gender salient. When prompted to think of themselves as Asian they scored measurably better than when their femaleness was prompted. In a second example, MRI evidence shows that African American faces provoked a vigilance response in observers. But if the researcher asked an observer to focus on specifics in the faces, as simple as looking for a mustache, that response was mitigated. In a third experiment, dyadic cross-race interactions produced a degradation in whites' performances on generic tests. The researcher formulated the hypothesis that conscious brain capacity was being used up on attempts not to appear biased.

Finally, one participant added that cognitive science has been producing evidence that the traditional distinction between cognition and emotion is not as clear cut as some might think. These experiments underscored the need for an awareness of what we would call an emotional component to learning, and may underscore the importance of understanding teaching as the development of a relationship.

INTERVENTIONS THAT STAND TO IMPROVE STUDENT LEARNING:

Graduate Education

While the focus of the day remained on undergraduate learning, all participants—administrators, faculty, and graduate students alike—agreed that a focus on graduate students held out particular promise. Dispositions toward the profession will be to a large extent formed during those years, and all saw possible improvement through more systematic training for graduate student teachers.

In general departments could be encouraged to instill the idea that the formation of scholars includes the forming of the technical, intellectual, moral, and inspirational aspects of teaching. A few specifics were also mentioned. An intensive program, perhaps during the summer month before the first teaching experience would provide an opportunity to introduce students to best practices and to instill in them a sense of institutional mission for teaching. Students in cohorts could work together, assess one another's teaching, and be assessed by trainers and undergraduates recruited for the program. Graduate students could keep, and be evaluated on, a process notebook to keep track of what works and what doesn't. Teaching mentors among faculty might be assigned to cohorts of incoming graduate students, to meet for regular workshops and discussions of teaching methods. Finally one participant suggested competitive, prestigious teaching fellowships, aimed for final-year social science and humanities

students and, so as not to interfere with post-doctoral positions, for earlier-year students in the hard sciences.

Undergraduate education

Some of the most compelling of the day's discussions offered up examples of specific classroom interventions that have yielded learning gains. For a list of these results, see Appendix. Here, some general remarks. One participant suggested a benefit in re-considering teaching work as promoting development (epistemological, analytical, etc.) rather than producing expertise in a particular field. Setting broad and explicit goals at an institutional level that articulate what an institution as a whole strives to instill in its students—for example, an openness to doubt, to conviction, and to revision—reinforces with individual professors and students a sense of coherence to the overall education.

Another participant identified two different sources of evidence from which scholars might draw to improve teaching. First, scholars might use the evidence gathered by specialists in fields that have a direct bearing on how people learn. Relatively recent research in cognitive science (see above) and the social sciences offer real opportunities for improving pedagogy—some mention was made of the promise of information science as well. This was relatively new ground for many present. A second strategy tries to encourage scholars to use their own classrooms as experimental laboratories to develop and test their own hypotheses about learning, which has the additional benefit of asking scholars to do exactly what they do best. No approach will be more efficient at nurturing a local culture of experimentation and assessment than asking professors to "find the best way." Simply setting up the problem produces the incentive to develop local, tailored, and effective means to measure results.

STRATEGIES AND TACTICS

Given the administrative experience of the participants, a particularly rich vein of discussion yielded observations regarding implementation of overall policy. All participants agreed on the importance of fostering a culture of teaching and learning at all levels of the institution. Ad hoc initiatives, involving assessment or not, disconnected from a stronger overall institutional mission, have less chance of success than those that are built into one. Several components of such a strategy emerged from the day's conversation.

Most important, recognition for innovative teaching should be part of the institution's reward structure. In the current state of affairs, large salary increases do not typically accrue to teaching stars. As it stands, teaching tends to be something from which the most celebrated are relieved, not something for which they are rewarded. Of course, such a change of reward structure would have an effect on the competition among elite institutions to attract faculty. One participant suggested inter-institutional discussion between large and prestigious universities, convened around the health of the profession as a whole, might be necessary to set workable standards. Further, with respect to the specific topic of assessment, a push from the top to make good teaching a core mission

will create an incentive among faculty to develop standards themselves to measure what counts as good teaching. One thinks here of the profession's success in coming up with standards and measures (imperfect but nonetheless widely accepted) for sizing up the quality of research—on which a whole raft of professional rewards depend. While good teaching may be as difficult to measure, one assumes, it would likely not be more difficult.

Of course, attempts to tinker with the calculus by which prestige is allocated run distinctive risks, which were partially addressed. One noted the benefit of having leaders on the faculty endorse the mission of teaching. Administrative prodding will be much more effective when joined by the active involvement of a Nobel Prize winner or two. In general, initiatives and language that seek explicitly to express aspirations for excellence in teaching in a comparative relation with research excellence will be less productive than a language that pushes simply for engagement, innovation, and invigoration of teaching.

Several participants pointed to the need for administrators to create incentives to share good teaching ideas and to institutionalize this sharing. They pointed out that currently, teaching advances tend to be pursued piecemeal by individual instructors, and when one instructor makes a useful gain there is little "bounce" across the institution (as one participant put it). Institutions stand to benefit by facilitating the dissemination of good pedagogical ideas across all academic units, and setting up systems to encourage their adoption.

One participant mentioned a procedure that is within every top administrators' grasp. When presidents' and provosts' committees review tenure and hiring cases, it was observed, they rarely raise even a single question about teaching. If even one case were sent back to the faculty (not necessarily declined) for insufficient evidence of good teaching, it would be an efficient way to make a point that evidence of good teaching was simply expected to be part of any appointee's portfolio. It would also spur a discussion among faculty about the best way to gather evidence of truly good teaching.

Participants shared several of what could fall into the "lessons-learned" category in their efforts to spur innovation at the department level. Participants generally endorsed the strategy of rewarding and reinforcing success rather than shaming failure. The notion of "competitive emulation" promised greater success than a high-stakes carrot or stick approach. This takes place via a process of identifying puzzles, perhaps pointing out a few possible solutions, and then kicking it over to the departments to come up with better ones. Two administrators reported limited success with blunt interventions regarding traditional measures of success in graduate programs—for example, time to degree and placement rates. Departments tend to feel such measures overlook the particularities of their fields. Also, while they have the clear attraction of being easily measurable, such qualities are admittedly only rough proxies for the overall health of a program. One administrator reported greater success after discovering that departments whose students moved through the Ph.D. exams as a cohort tended to finish their degrees earlier. In most cases, simply pointing this out to departments started a process where local solutions were adopted to shorten time to degree. One administrator reported the benefit of simply

making a point to attend departmental meetings to talk seriously about teaching. Just showing a face made a point about the administration's commitment to the issue (and provided a valuable chance to elicit faculty's reciprocation).

APPENDIX: SPECIFIC EXAMPLES OF TEACHING IMPROVEMENTS

One-minute essays. One participant reported real gains from a simple exercise of asking for one minute essays at the end of classes. Students are asked to respond to two questions: what was the most important thing you learned today and what are you most confused about. These regular mini-evaluations provided timely information for the instructor to keep the course on track.

Hand-held remotes. Some professors of large lecture courses have found benefit from the creative use of hand-held remote control devices for students. Professors engage students during lecture by soliciting responses that a local computer tabulates and displays on an overhead projector. The exercise served multiple purposes. It allowed the professor instant information on whether main points were sinking in, it strengthened student's grasp and retention of salient ideas, and since data was stored, quickly built into a valuable source of longitudinal information about student's learning, under different teaching strategies, from year to year.

Measuring writing skills. One university discovered a problem with how its science majors were doing in developing their writing ability, a fact which was clarified as a specific intervention to measure it found that the students' writing actually deteriorated during their four years. This finding spurred the relevant departments into a period of experimentation which resulted in dramatic improvements.

Project-based learning. After observing that graduate students will teach themselves materials and skills they need to complete projects, one participant reported success in applying a similar template to undergraduates.

Exit interviews. One university's math department knew its students were mastering technical concepts but wondered whether they were well-rounded enough to put them into larger contexts. They did clinical interviews with undergraduates asking them open-ended questions—for example, "How would you explain what an integral is to someone who doesn't know?" These interviews, which were video-taped and shared among faculty, revealed sufficient depth, but lack of breadth and so stimulated faculty ideas for teaching. The faculty decided as a group to adjust their teaching to nurture the impulse to treat uncertainty as a starting point for hypothesis formation and inference generation, and not as an end point of conversation.

Writing portfolios. The faculty of Carleton College some years ago identified a need to improve students' writing and adopted a college-wide writing portfolio requirement. When faced with the need to evaluate the portfolios the faculty found themselves engaged in an intense and intellectually rewarding cross-disciplinary discussion of standards and measures of clarity and originality of thought and expression. Faculty there reported a re-invigorated attention to their goals and expectations regarding their own students.

Revised course evaluations. One university improved its student course evaluations through a committee, with students, faculty, and administration included, that tried to bring the questionnaire in line with the institution's core educational goals. Questions asked now ask how much courses improved students' abilities to analyze, think critically, write, etc. Not only did these new evaluations provide more useful information for instructors, they also reinforced a message for students about what the university as a whole considers as important to their overall education.

Investigating and nurturing study habits. Several participants mentioned the case of a professor of mathematics at Berkeley who found that his African American students were underperforming his Asian American students and set out to find out why. He discovered that Asian American students tended to study in groups while African American students tended to study alone. Learning math, he hypothesized, is especially facilitated by group study. When he encouraged group study among African American students their results improved dramatically.