Sci Eng Ethics (2012) 18:351–367
DOI 10.1007/s11948-010-9250-2

Assessing Graduate Student Progress in Engineering Ethics

Michael Davis · Alan Feinerman

Received: 19 July 2010 / Accepted: 8 November 2010 / Published online: 21 November 2010

Abstract Under a grant from the National Science Foundation, the authors (and others) undertook to integrate ethics into graduate engineering classes at three universities—and to assess success in a way allowing comparison across classes (and institutions). This paper describes the attempt to carry out that assessment. Standard methods of assessment turned out to demand too much class time. Under pressure from instructors, the authors developed an alternative method that is both specific in content to individual classes and allows comparison across classes. Results are statistically significant for ethical sensitivity and knowledge. They show measurable improvement in a single semester.

Keywords Ethical judgment · Ethical sensitivity · Ethics knowledge · Assessment · Engineering ethics · Engineering graduate education

> "Assessment is not so much a can of worms as a sea of snakes. The charts are bad; the weather, foul."—Anonymous

Within a single course, assessing student progress in engineering ethics, research ethics, or any similar field is (in principle at least) not much harder than assessing progress in calculus, organic chemistry, or macro-economics.[1] When one or more

[1] The philosopher-author of this article wrote this sentence without the "in principle"; the engineer-author added it to warn other engineers that, in practice, it can be hard—at least for engineers without experience of this sort of grading. As with any new form of grading, there is a "learning curve".

M. Davis (✉)
Humanities Department, Illinois Institute of Technology, 5300 S. Shore Drive #57, Chicago, IL 60615, USA
e-mail: davism@iit.edu

A. Feinerman
Department of Electrical and Computer Engineering, University of Illinois at Chicago, 851 South Morgan St. (M/C 154), Chicago, IL 60607, USA

instructors control the content of a course, they need only prepare tests to measure how much students have learned of what was taught. In a course in engineering ethics, for example, the instructor can tell whether students identify engineering ethics problems of the sorts discussed, whether they draw on information presented, whether they apply methods of resolution they have practiced, and whether their solutions are competent or not. Instructors need only score accordingly.

Assessing ethics learned becomes hard when the assessment is across different courses, departments, or educational institutions. What is especially difficult is assessment when there is little or no control of what in particular is taught across a wide range of courses or institutions. That especially difficult problem is the subject of this article. The solution offered here, though novel, has an analogue in ordinary comparison of achievement across courses (for example, "class standing"). The evidence offered for success is preliminary but intriguing. We offer that solution here for the usual reasons: to add to what is available in the literature; to invite critical discussion; and to entice others into trying our solution under conditions more favorable to its vindication. We recount *how* we reached that solution in part because the solution is easier to appreciate once one understands the problems it had to solve; but we also recount those problems to help others avoid the problems. Too many reports make assessment sound easy ("plug and chug"). For any assessment of ethics learning across the graduate curriculum in engineering—and, no doubt, in many of the sciences as well, assessment has several practical impediments not yet reported in the literature.

Background

When we speak of "ethics", we refer to those (morally permissible) standards of conduct that apply to members of a group simply because they are members of that group (and to the conduct those standards make appropriate). Engineering ethics is for engineers because they are engineers; research ethics is for researchers because they are researchers; and so on. We are not here concerned with ordinary morality as such or moral philosophy as such (two other senses of "ethics"). When we speak of teaching *ethics*, we mean ethics in this special-standards sense. When we speak of *teaching* ethics, we mean at least one of the following:

(a) improving ethical **sensitivity** (the ability to recognize problems covered by the relevant standards),
(b) increasing ethical **knowledge** (appropriate terms, relevant standards, related institutional practices, such as "hot lines", decision procedures, and other ethical resources),
(c) enhancing ethical **judgment** (the ability to make competent choices of the appropriate sort for the appropriate reasons more often than chance or common sense), and
(d) reinforcing ethical **commitment** (the likelihood that students will act on what they have learned).

There is, of course, no way for ordinary academic methods to tell whether students will later use what they learned (d). But that is no surprise. Even in calculus, organic chemistry, and macro-economics, there is no test to tell whether students will use what they learn. The working assumption, derived from common sense, is that students are significantly more likely to use what they know than what they do not know. We should not expect more of ethics assessment than of assessment in other subjects.[2]

When we talk of *assessing* how much ethics has been taught, we will be concerned only with the first three aspects of teaching ethics (a–c).

Educators have developed three major approaches to teaching ethics (in this sense) within the formal curriculum: (1) freestanding courses, such as Engineering Ethics; (2) modules, that is, large-scale insertions of ethics instruction into technical courses (for example, an hour-long discussion of research misconduct or screening of a pedagogical movie such as *Gilbane Gold* or *Ethics and Water*); and (3) micro-insertion, the small-scale insertion of ethics instruction into technical courses, resulting in a dozen or so "ethics mini-lessons" during a semester, each lasting only a few minutes.[3]

In Autumn 2006, Illinois Institute of Technology (IIT) received a three-year grant from the National Science Foundation (NSF) to develop the third approach, micro-insertion, as a way of integrating ethics into engineering's *graduate* curriculum. We were to train graduate faculty in engineering to develop and use micro-insertion. (For more about micro-insertion, Davis and Riley 2008, and Riley et al. 2009.) Among the innovations of the project were: training graduate *students* to develop micro-insertions to make it easier for faculty to prepare assignments for class; developing a website, the Ethics In-Basket (http://hum.iit.edu/ethics-in-basket/), to which anyone might add problems they developed (as well as download for their own use problems developed as part of the project); and creating the infrastructure to make that website (more or less) permanent. Our chief hypothesis was that micro-insertion could teach engineering ethics to graduate students measurably better than ordinary "ethics free" technical courses. A crucial component to the project was, then, assessment of the ethics learned.

Our original plan of assessment was relatively conventional. One element was to be a student self-report we had used before. We have described this mode of assessment (and results) for earlier projects elsewhere (Davis 2006). While a self-report can tell us that students noticed the ethics (or not), thought they learned something useful (or not), and approved (or disapproved) of the integration of ethics into a technical course, a self-report cannot tell us how much, if anything, the students actually learned. The new elements of our assessment plan were to tell us that. We wanted to measure, as directly as possible, the effectiveness of the ethics teaching.

[2] The literature on in-course assessment of ethics learned is not large but growing quickly. Beside other work sited here, see, for example: Bebeau (2002a, b, 2005), Mumford et al. (2006), Kligyte et al. (2008).

[3] Outside the formal curriculum, there are several other options available, such as handing out the NSPE Code of Ethics as part of orientation materials, integrating ethics discussions into orientation sessions, special departmental colloquia or seminars on issues in engineering ethics, and voluntary events, such as outside speakers on ethics.

We wanted to measure improved ethical judgment, of course. At the time, there was no standardized test of *ethical* judgment as such, much less one specific to engineering.[4] We proposed to use what seemed a reasonable surrogate, the Defined Issues Test (DIT-2), a standardized measure of moral development. Since moral judgment seems to improve with moral development, the DIT-2 would (we thought) provide a reasonable measure of *moral* judgment. Any unusual improvement in moral judgment during the 15 weeks of a single class in which ethics was taught would, we thought, probably consist of (or correspond to) improvement in the relevant ethical judgment.

At the beginning of the semester, and again at the end, students in each class were to take the DIT-2. The difference between pre- and post-test would provide a measure of how much, if at all, micro-insertion raised the students' ethical judgment. While it may seem unlikely that a few micro-insertions would have any effect measurable by the DIT-2, there is in fact evidence that even a 35-minute film (without subsequent classroom discussion) can have a significant effect (Loui 2006). That surprising result suggested that several micro-insertions over a semester in a single course might have a similar effect. If we could confirm (or disconfirm) that hypothesis, we would have learned something important. We (or others) would also then be in a position (in a later study) to compare the relative effect on judgment of the three methods of including ethics in the graduate engineering curriculum.

We also wanted to measure ethical sensitivity and knowledge. One standardized test that seems to do that for engineering is the ethics component of the Fundamentals of Engineering Exam (National Council of Examiners for Engineering and Surveying). Unfortunately, the questions on that exam were too specific for our purposes, too focused on problems that licensed Professional Engineers (PEs) face, and too difficult for graduate students who have not studied for the PE exam. We therefore proposed to develop our own exam to assess ethical sensitivity and knowledge.

Because Muriel Bebeau had done something similar for dental students, we initially adopted her approach (Bebeau and Thoma 1999). We developed a list, appropriate for graduate students in engineering, of what Bebeau calls "intermediate concepts" (such as data integrity), that is, ethical terms that graduate students in engineering should recognize and understand. Because these ethical terms must be appropriate for *graduate* students, many of them concern research, making the list of concepts useable for teaching other graduate students but not for teaching most engineers (those not in research).[5]

[4] A test of ethical development in engineering has recently become available. Borenstein et al. (2010). If, as we believe, ethical judgment improves with ethical development, the title of their article is not misleading.

[5] Right now, this is our complete list:
- Accessibility (designing with disabilities in mind)
- Animal subjects research
- Authorship and credit (co-authorship, faculty and students)
- Publication (presentation: when, what, and how?)
- National security, engineering research, and secrecy
- Collaborative research
- Computational research (problems specific to use of computers)

Once we had our list of intermediate concepts, we could (we thought) develop a test, that is, a set of short problems (much like those used in micro-insertion) in which those issues appeared. We could give the test to students near the beginning of the term in a class which was to have micro-insertions and again near the end. We could do the same with an equal number of similar "control" students (students in sections of the same course that are not using micro-insertion). In this way, we could (we thought) build on the method that Bebeau developed for dentistry and that others have successfully applied to social work, journalism, and other fields outside science and engineering (IOM 2002). We would simply be extending her method to graduate students in engineering. This pre- and post-testing for intermediate concepts would give a measure of micro-insertion's contribution to ethical sensitivity (the ability to see issues) and ethical knowledge (at a minimum, the ability to label issues with appropriate terms).

That, basically, is what we *proposed* to do. As it turned out, our proposals were impractical for reasons neither we nor any reviewer realized. One important finding of our research is that there are serious *practical* impediments to research of this kind, not the least of which is the small size of graduate classes in engineering and the infrequency with which those classes are taught. (The next section describes three other impediments.) Another important finding is that panels evaluating research proposals concerned with ethical assessment probably should contain more reviewers with experience of assessment in graduate courses in engineering (or the natural sciences) in classrooms *not* their own.

Practical Limits on Assessment

There were at least three separate impediments to the assessment plan described above (beside the two already mentioned). We might categorize them as: (1) time, (2) relevance, and (3) comparability.

The first impediment, time, was a total surprise. The DIT-2 requires at least 1 hour to administer; the substantive test we were planning to use in addition would, we thought, require about the same amount of time.[6] Allowing for both a pre- and

Footnote 5 continued

- Conflicts of interest
- Cultural differences (between disciplines as well as between countries)
- Data management (access to data, data storage, and security)
- Confidentiality (personal information and technical data)
- Human subjects research in engineering fields
- Peer review
- Research misconduct (fabrication, falsification, and incomplete disclosure of data)
- Obtaining research, employment, or contracts (credentials, promises, state of work, etc.)
- Responsibilities of mentors and trainees
- Treating colleagues fairly (responding to discrimination)
- Responsibility for products (testing, field data, etc.)
- Whistle blowing (and less drastic responses to wrongdoing).

[6] The self-assessment would only have taken 15 min once at the end of the semester but was dropped while trying to find room for the new tests.

post-test for each class, we were asking faculty to devote at least 4 hours of class time during a term to testing that they would not otherwise perform. The faculty we had recruited for the project refused to take that much time from the substance of their course. They had (they said) too much technical material to cover; the curriculum was just too full. They were willing to give no more than 15 min of class time at the beginning of the term and another 15 min at the end. We had faculty at three different schools—Howard, IIT, and University of Illinois-Chicago (UIC). All three groups—without any opportunity to communicate with counterparts at the other two schools—reacted in almost exactly the same way, even to the number of minutes they set as their upper limit. Since we had had to work hard to recruit these faculty, we did not try to replace them with others who would be more liberal in their use of class time. (We did not even try to find faculty who would be willing to loan their classes for use in the "control testing" we were then planning).

We considered testing online instead of in-class. We abandoned that idea because the experience of other researchers, especially a group at Georgia Tech, was that the percentage of a class taking a test online (or completing it) would be substantially lower than the percentage if the test were taken in class. Even when classes were larger than we had available, such as a typical undergraduate engineering class at Georgia Tech, the rate of online response could be low enough to make the results of testing more or less meaningless (Borenstein et al. 2010, p. 395). Given that we were already dealing with small classes, generally ten grad students or less, a low response rate would have made results statistically useless.

The faculty's main concern was class time. But it soon became clear that they had a second important concern, "relevance". The tests we were proposing to administer had only an indirect connection with the official subject of the course. The faculty did not think they should (or legally could) require students to take pre- and post-tests most questions of which had little or nothing directly to do with the course. They felt that they might be using student time in a way that would be, well, unethical (a violation of academic standards). We never put this question to an IRB, but most IRB members we have raised the question with seem to agree. We would have been forcing students to undergo tests not for their benefit.

The faculty would not have raised that objection to similar testing in a course, such as Introduction to the Profession, that has the stated purpose (among others) of introducing students to engineering ethics. Helping students assess their ethical sensitivity, knowledge, or judgment would be wholly legitimate in such a course (they thought). But how do you explain to students in a course like Computer Design or Process Modeling that you are going to use almost a twelfth of the term's class time for ethics assessment (most of which would not be relevant to the course)? The tests would probably be unpopular with students—and justifiably so. Their time (they would say) is too valuable to be taken up with such "non-technical" matters largely foreign to the course description.

These two impediments were enough to sink our plans. But there was a third. To give comparable results, a standardized test of sensitivity or knowledge has to be general enough to provide information about ethics learned in *any* graduate engineering course (or at least most of them). But it also has to be specific enough to provide information that is useful for assessing a particular class's small-scale

contribution to ethical sensitivity and knowledge. Even asking about the Code of Ethics of the National Society of Professional Engineers (NSPE), though relatively general, is not general enough for a graduate course in Computer Design where the IEEE Code of Ethics would be more appropriate. In any case, the ethics taught in Computer Design might be much too specific for a general test to pick up, for example, sensitivity to the issue of honesty in reporting one computer as "twice as fast" as another when there are several criteria for the computer's speed and different criteria yield different relative speeds.

We were not the only ones to run up against this impediment. NSF has granted several million dollars for projects to improve ethics teaching of graduate students in engineering. All those projects included assessment. When the Principal Investigators (PIs) on those projects were called together to report progress (January 8–9, 2009), those who were trying to do the sort of assessment we were interested in were all having the same problem we were. They faced what seemed to be a natural law governing tests for sensitivity and knowledge (but not judgment): *The more general the test, and therefore the more useful for comparing across courses, the less able it is to register much about the ethics that students learned in a particular course and, therefore, the more likely to register "nothing learned"; the more specific the test, and therefore the more useful for registering what students learned in a particular course, the less useful for comparison across courses*. There seemed to be no middle ground for a test both general enough to produce comparable results across a wide range of courses and specific enough to measure what was actually learned in a particular course.

In addition, tests that even tried to be general enough tended to be quite long—with most questions being irrelevant to most courses. The PIs who did *not* have this problem with assessment were using their own classes for research. They could use a test designed specifically for their own class, though they could not compare their results against classes with a different syllabus. But even they seemed to have given up the idea of administering the DIT-2 in class (citing time and relevance as reasons).

One way to avoid this third impediment (incomparability), or at least to moderate its effect substantially, is to define the body of knowledge first, structure the curriculum to teach it, and only then develop a test to assess what was learned. Knowing what will be covered overall allows for considerable economy in test design. That, in fact, seems to be what made it possible for Bebeau to design useful tests of ethical sensitivity and knowledge for her dental school.

That, however, was not a way open to us. We could not control the curriculum at any of the three schools in which our research was to be performed. Indeed, given how accreditation in engineering now works, no one outside the engineering department (or program) in question could have such control. Each department is free, within wide limits, to structure its curriculum as it sees fit—provided it makes clear what it is doing and how what it is doing satisfies general criteria for what students should know by graduation (ABET 2009). The test we planned to develop would (if we succeeded) have been useful precisely because it did not depend on control of curriculum. We had, it seemed, reached a dead-end.

Unwilling to give up, we called in a consultant, a member of IIT's Institute of Psychology. She not only confirmed the dead-end but described her own attempts to get around it in order to assess ethics learning in IIT's interdisciplinary design classes, each consisting of less than a dozen students, mostly undergraduates. Each semester, each such class is supposed to solve a *new* practical design problem. While each class is therefore likely to be quite different in substance from most others, there is a list of outcomes all should achieve. Improved ethical sensitivity, knowledge, and judgment are among those outcomes. Our consultant had actually worked out a preliminary test of fourteen multiple-choice questions (in three versions) and tried it out in design classes with which she was associated. Her conclusion was that the test measured nothing. Whatever ethics was learned in those classes, if any, was too specific for her general test to measure.

Perhaps seeing the disappointment on the faces of our research team, she went on to describe other means of assessment we might try, such as interviews (in which students were asked to describe ethics problems they had faced), portfolios (of the ethics problems dealt with in homework assignments), and the like. (For more about these standard options in assessment, see Suskie 2004.) The problem with most of the methods she suggested, apart from the enormous investment of researcher time needed to carry them through and our need to have graduate students willing to do the extra work of being assessed in this way, was what American engineers like to call "comparing apples and oranges" (that is, comparing incommensurables).[7] Being highly sensitive to what was learned in one design class, these methods do not provide much information that can be directly compared to what has been learned in any other design class. How does one compare learning something about conflict of interest in one class with learning something about safety in another? Again and again we returned to the problem of finding a procedure that would allow results in one class to be compared reliably with results in another. Sometime during this discussion, someone, tiring of the many references to "apples and oranges", asked, "Why can't we just count fruit?" The question sent the discussion off in a new direction, with someone else noting that the *ratio* of apples before and after some had been eaten *was* (mathematically) comparable to the *ratio* of oranges once some of them had been eaten. The same would be true of a ratio of scores on a single course's pre-test (the test given before exposure to the micro-insertions of ethics) and post-test (the test given afterward) when compared to the same ratio from another course. Seemingly out of nowhere, we had a new plan.

Each instructor would develop his or her own pre- and post-test. This would allow for the ethics questions to be integrated into the exams in just the way that the ethics learning had been integrated with the technical. The instructors would be doing what they usually do on exams, seeing how much of what was taught was learned. Since the ethics taught had been integral to the course (as it is supposed to be for micro-insertion), the ethics of using class time for such testing was resolved. There need be no additional testing, only the usual number of exams with subject matter and grading adjusted slightly—all appropriate to the class in question.

[7] The Danes seem to have an especially nice expression to make this point: "Which is higher, the Round Tower or the volume of a thunderclap?".

Thinking this plan over, we realized that it was not as radical as it seemed at first. We use grades to compare student learning across an enormous range of courses (for example, for "class standing" or "Dean's List"). Within a course, a grade tells us a good deal about the substance of what was achieved. But the grade itself is a pure number (just as a ratio is), allowing comparison across courses and even (much more roughly) across institutions. If grades in general allow for such comparison, why should the ratio of pre- and post-test scores not do the same?

That seemed a good question, one to be tested by doing the assessment that way and seeing what could be learned. The bane of ethics assessment has always been the null result. The great achievement of the DIT-2 (and its predecessors) was to find something (seemingly important) that could be measured. There was no guarantee that our method would produce any measurable result. We therefore undertook two "pilots" to see whether we could measure anything in this way with the intention of using the first pilot to discover (and fix) problems in the research design. If the pilot assessments did not show we were reliably measuring something ethical, we would not need a control to assure us that what went on in class explained that something. We therefore put off the control assessment until we knew there would be some point to it.

First Pilot

Alan Feinerman, a co-PI, generally teaches two courses at UIC each Spring, ECE 449 (Microdevices and Micromachining Technology—a nanofabrication laboratory course) and ECE 541 (Microelectronic Fabrication Techniques—which examines the science behind nanofabrication processes). ECE 449 seemed the best choice for a pilot for at least three reasons. First, most of the graduate engineering courses we were dealing with (at IIT and Howard) turned out to have such variable enrollments that their instructors could not guarantee that they would be taught as scheduled (which was generally once a year). ECE 449 seemed to have enough students every year to be a "sure thing". Second, ECE 449 contained as large a number of graduate students as we had available at any of our three institutions. Third, Feinerman had himself become interested in ethics assessment.

There was one more reason for choosing ECE 449. That course (unlike Feinerman's ECE 541) had a substantial number of (senior) *under*graduates as well. Their presence would allow for the testing of the assessment method on undergraduate as well as on graduate students, if we chose, and perhaps even for comparison of undergraduate and graduate improvement.

After we obtained IRB approval from both UIC and IIT, Feinerman performed his first pre- and post-test in Spring 2008, a second round of testing in Spring 2009 (after some restructuring), and the control in Spring 2010. The 2008 pre-test was embedded in the first regular test of the semester (T1, February 19); the post-test in the fourth test of the semester (T4, May 6, the last). T1 had nine problems, three of which concerned ethics in whole or in part. For example, question T1-2, that is, the second question on T1, was (designed to be) entirely about ethics:

> List one **advantage** *and* one **disadvantage** MEMS/NEMS has for society (that means you, your relatives, their friends, …)?

Knowing at least one social advantage and at least one social disadvantage of technological breakthroughs in micro-electrical mechanical systems (MEMS) or nano-electrical mechanical systems (NEMS) concerns engineering ethics insofar as engineers have a professional obligation to look after the public health, safety, and welfare and knowing about such advantages and disadvantages is necessary to do that (and is therefore a form of ethical knowledge). Neither Feinerman's lectures nor assigned readings for ECE 449 had yet discussed the social advantages or disadvantages of MEMS/NEMS. The corresponding question on T4 was:

> List one **advantage** *and* one **disadvantage** that a micro-fluidic device has for society (that means you, your relatives, their friends, …)?

The only difference between this T4 question and its T1 counterpart (above) is that students should (probably would) have learned something about the advantages and disadvantages of micro-fluidic devices during the semester (for example, their use in pumping insulin).

The difference between an ethics question and others on T1 or T4 is in the answers that would be appropriate (for example, because the question concerned "advantage for society" rather than fluid flow). In this respect at least, Feinerman had thoroughly integrated the ethics pre-test into the course's ordinary assessment process. Students saw nothing odd in these question since they seemed to test student knowledge of the technology (and, in fact, did test that too).

Scoring such questions was to be simple. No correct item would earn a score of 0; one correct item, 5; and at least two correct items, 10. Simple scoring is important for at least three reasons. First, it makes learning to score ethics questions (relatively) easy (or, at least, less daunting) for anyone familiar with the technical material but not with grading ethics. A more complex set of rubrics (though providing a better assessment overall) would take a professor of engineering longer to learn to use reliably (and is more daunting). (See, for example, the rubrics used in: Sindelar et al. 2003) Like most engineers, Feinerman had neither the time nor the patience for learning a complex grading scheme; he also did not have the money or time to train a graduate student to do the grading for him. Second, the simple scoring meant that scoring the tests would not be nearly as time consuming as using the more complex rubrics would be. Third, the simplicity of scoring meant that the grading of questions would be (relatively) objective. There would be relatively few judgment calls.

The results of the first pilot were suggestive but not statistically significant. Of the three pairs of questions, two of the three scores for graduate students showed improvement (from average of 8.3 to 9.2 on one and from 9.2 to 10 on the other). But on a third pair of questions, change went the other way (from 9.2 to 7.5)—owing to one student whose score went from 10 on T1 to 0 on T4. One student has a significant impact when the total pool of grad students is six. The averages for the ten undergraduates showed no such irregularity. See Table 1.

Table 1 ECE 449, Spring 2008

	T1-1	T4-1	T1-2	T4-3	T1-3	T4-6
U1	–	10	–	5	–	10
U2	5	10	5	5	3	0
U3	5	5	5	5	5	5
U4	10	10	3	5	5	10
U5	10	10	10	5	10	10
U6	5	10	0	10	10	5
U7	10	10	10	10	10	10
U8	5	10	5	5	10	5
U9	5	10	–	5	3	10
U10	10	10	10	10	10	10
Mean	7.2	9.5	6.0	6.5	7.3	7.5
STD	2.6	1.6	3.7	2.4	3.2	3.5
G1	10	10	–	10	10	10
G2	10	10	–	5	10	10
G3	5	5	10	10	10	10
G4	10	0	–	10	5	10
G5	10	10	10	10	10	10
G6	10	10	5	10	10	–
Mean	9.2	7.5	8.3	9.2	9.2	10
STD	2.0	4.2	2.9	2.0	2.0	0.0
Mean	6.9	7.8	1.14	Undergrad		
STD	3.1	2.8		Undergrad		
Mean	9.0	8.8	0.98	Grad		
STD	2.1	2.8		Grad		

"T" stands for "test"; the number immediately following is the test's number; the number after the hyphen is the question's number. "U1" is the first undergrad; "G1" is the first grad student. Each grad is identified by numerals 01 through 10

We now had to decide how to present the ratio. We quickly agreed *not* to present it as a fraction; the denominators would generally be different and comparing fractions with different denominators is not easy for most people. We would carry out the division each ratio represents. We then had to decide whether the pre-test score or the post-test score would be on the bottom. We decided to let the pre-test score be on the bottom. Because (we hoped) the post-test score would generally be higher than the pre-test score, putting the pre-test on the bottom would (generally) give us a result greater than 1, a number that could (in principle) be quite large rather than approach in increasingly smaller steps toward 1. (We thought that advantage was worth the small risk of the pre-test score being zero.)[8] Following

[8] There is also an ethical issue here, though probably one that is merely academic. Both methods of displaying the ratio as a single number are, in principle, misleading. We would make large improvements seem small if we put the pre-test score on top but make the large improvement seem more dramatic than it is in fact if we put the pre-test score on the bottom. The issue is merely academic insofar as the observed

these decisions, we did the arithmetic for our first pilot. The mean "improvement" for the grad students was 0.98 (the sum of the T4 means divided by the sum of the T1 means), an overall *loss* of ethical sensitivity or knowledge. The undergraduate result was better: 1.14. But neither result was statistically significant. The standard deviation on pre and post-tests for both grad and undergraduate students was well above 3.

These results illustrate a problem of dealing with small numbers. The negative result for the graduate students was due entirely to one student getting 10 on one question on the pre-test and 0 on the corresponding question on the post-test. All other grad students got the same score on that T4 question (one designed to be easy) as on the corresponding T1 (10 in all but one instance)—and all the undergrads but one also got 10 on it (with that one scoring 5 on both). Four undergrads improved over the first test; none did worse. The temptation, then, is to treat the grad student's 0 on that one post-test question (after a 10 on the pre-test) as an outlier (perhaps he misread the question). Ignoring that one outlier, the graduate students' overall improvement score is the same as the undergraduate: 1.1.

Second Pilot and the Control

The results of this first attempt at assessment were humbling—for at least three reasons, all related to grading. First, students occasionally came up with answers that, though insightful in their way, managed to ignore what had been taught (for example, they identified true social disadvantages not covered in the course). Feinerman felt he had to give them credit (and that credit is represented in the table). Generally, according to Feinerman, students with more practical experience in engineering (for example, 10 years in the field) were more likely to produce such unexpected answers than those with less experience.[9] Second, but related to the first problem, Feinerman came to believe that using different questions on T1 and T4 added substantially to problems of comparison because one question might produce more unexpected responses than another. We therefore decided to use the same questions on pre- and post-test the next time rather than (as in this first pilot) similar questions that might turn out to be not as similar as supposed. Third, we decided that the questions, though easy to grade (10, 5, 0), did not provide enough information about what students had learned. More open-ended or complex questions would provide more information (without making the grading more complicated).

Footnote 8 continued
change is probably never going to be large enough to give a false impression, however we chose to define the ratio. But note: a reviewer for this journal referred us to Hake (1998) which uses a more complicated method to get a number, one that seemingly escapes both the risk of zero denominator and our ethical dilemma: (Posttest Score–Pretest Score)/(Maximum Possible Score–Pretest Score).

[9] This observation is, of course, possible because the class is small (and a lab rather than lecture) Feinerman was able to learn a good deal about each student in the course of the semester. A more formal study would gather this sort of information in advance. This observation is not meant to prove anything, merely to suggest an explanation worth further investigation.

Though humbling in these ways, the first data set is nonetheless suggestive, much more suggestive than the usual statistical measures allow. Except for that one graduate student's slip on one question, all the students in the class, grad as well as undergrad, either improved their ethics score from T1 to T4 or remained the same. While micro-insertion seems to have had only a small effect (.1), it did seem to have some, enough at least to measure, even if (given the small numbers involved) the result was not statistically significant. We had reason to be hopeful—and to try an improved pilot the following year.

One improvement in the Spring 2009 pilot was to use the same set of ethics-assessment questions in the pre- and post-test (reducing unexpected variability). The ethics-assessment questions were also somewhat different. Feinerman came to think that explicitly asking about "social" advantages and disadvantage was too heavy-handed and made the ethics questions stick out too much. So, he tried to do a better job of integrating the ethics into the technical. Here, for example, is the first ethics-assessment question in the pre-test (and post-test) for Spring 2009:

(a) Why are researchers investigating nanotechnology to correct environmental pollution?
(b) List advantage(s) of using nanotechnology for environmental pollution remediation.
(c) List disadvantage(s) of using nanotechnology for environmental pollution remediation.
(d) Give an example of an acceptable and an unacceptable use of nanotechnology for environmental pollution remediation.

The first three parts (a–c) require technical information (more or less). The fourth part requires a specifically ethical response (assuming "acceptable" to be a reasonable prompt for "ethical"). The middle two parts are explicitly open-ended, allowing students to list as many advantages or disadvantages as they can (some of which could be ethical). The pre-test had six such questions (of varying difficulty). The post-test was identical. The pre- and post-tests were given as free-standing quizzes to between 75 and 100% of the total class (depending on attendance). The pre-test near the middle of the semester (April 1, 2009), the post-test near the end (May 1, 2009). The students were informed that the pre- and post-test were for their own benefit and would not directly count toward their grade but that participants would have an extra question dropped from two tests that did count toward their grade. This inducement led to nearly unanimous participation in the voluntary tests. The ethics tests concerned nano topics that were covered after the pre-test and that were selected because they were areas of research by UIC faculty: nano particle remediation of ground pollution, drug delivery by nanotechnology, infra-red detectors, quantum dots, and nano-magnets.

Feinerman chose to separate the pre- and post-test from his regular sequence of tests for two reasons. First, he wanted to have more questions than would be possible if he integrated the questions into the regular tests (six ethics questions rather than three). The larger number of ethics questions made it likely that the overall standard deviation would be smaller. Second, he had come to think his students would find the ethics test questions enough like his other questions that he

need not bury them in a regular test. He did not use them as part of the final grade because he did not want to have to grade the post-test before the end of the term (and he thought his students would take the tests seriously even without grading as an incentive).

The results of this second pilot were much more satisfactory than the first. See Table 2. The standard deviation was much smaller than on the first pilot: 0.59 (for undergrads as well as for grads). The measured improvement was also much greater: 1.40 for grad students (and 1.60 for undergrads).

The control was accomplished during Spring 2010 with pre- and post-tests administered to between 58 and 100% of the class. The pre- and post-test were, in most respects, like those in the second pilot. The only exception was that Feinerman, being on Sabbatical that semester, did not teach the course (though he did administer and grade the tests). The ECE 449 instructor that semester did not

Table 2 ECE 449, Spring 2009

		1	2	3	4	5	6	Sum	1	2	3	4	5	6	Sum	Ratio
1	U01	2	6	0	0	0	7	15	5	7	0	5	6	8	31	2.07
2	U02	6	5	0	0	6	7	24	4	5	3	1	5	5	23	0.96
3	U03	0	1	0	0	2	5	8	3	4	1	0	2	4	14	1.75
4	U04	9	7	4	0	2	8	30	4	7	0	4	4	7	26	0.87
5	U05	3	3	0	0	5	5	16								
6	U06	3	5	2	1	0	3	14	5	5	0	5	3	7	25	1.79
7	U07	2	5	0	0	1	2	10	5	3	0	7	1	1	17	1.70
8	U08	4	6	0	0	0	2	12	0	5	3	1	4	1	14	1.17
9	U09	1	0	4	2	0	2	9	5	3	4	0	3	5	20	2.22
10	U10	4	2	0	0	0	7	13	4	6	0	0	5	7	22	1.69
11	U11	7	6	1	1	4	10	29	9	8	6	3	8	7	41	1.41
12	U12	3	3	0	0	0	5	11	7	6	5	4	5	5	32	2.91
	Avg	3.7	4.1	0.9	0.3	1.7	5.3	15.9	4.6	5.4	2.0	2.7	4.2	5.2	24.1	1.68
	Std	2.6	2.2	1.6	0.7	2.2	2.6	7.6	2.2	1.6	2.3	2.5	1.9	2.4	8.2	0.59
13	G01	6	8	4	1	4	7	30	4	9	4	3	6	9	35	1.17
14	G02	11	9	7	4	11	8	50	9	14	7	9	9	11	59	1.18
15	G03	1	7	0	0	0	8	16	0	3	0	0	0	7	10	0.63
16	G04	6	5	0	0	0	3	14	6	4	3	0	2	5	20	1.43
17	G05	9	9	2	0	2	0	22	10	10	7	3	5	5	40	1.82
18	G06	8	9	0	4	5	7	33	6	7	4	4	6	5	32	0.97
19	G07	0	5	0	0	1	3	9	4	3	0	3	5	3	18	2.00
20	G08	8	11	0	0	5	9	33	5	5	3	1	3	8	25	0.76
21	G09	5	3	0	1	0	4	13	6	7	5	4	5	6	33	2.54
22	G10	11	11	1	0	3	6	32	12	12	4	0	11	9	48	1.50
	Avg	6.5	7.7	1.4	1.0	3.1	5.5	25.2	6.2	7.4	3.7	2.7	5.2	6.8	32.0	1.40
	Std	3.7	2.7	2.4	1.6	3.4	2.9	13	3.4	3.8	2.4	2.8	3.2	2.4	15	0.59

The numbers along the top indicate questions. "U" indicates an undergrad; "G", a grad student. Each student is identified by numerals 01 through 12

discuss the nano-topics on the pre- and post-tests either. He taught the course much as Feinerman would have taught it before he began to integrate ethics. It was therefore as clean of ethics as a graduate course in engineering might reasonably be. The students were not offered any inducement to take the tests—but participation of those present was again nearly unanimous. Each test was administered part-way through a regular lecture.

The pre-test was administered on March 15; the post-test, on April 26. The number of students taking the tests were about the same as the year preceding (ten grad students and seven undergrads). Feinerman graded the six ethics questions on each of the two tests just as he had done the year before. This time, however, the students showed virtually no ethical progress overall (about .04 as against .4 or .6 for the second pilot). The standard deviation was about the same as the second pilot (.4 for grad students, .6 for undergrads). The difference between teaching some ethics in a single graduate engineering class and teaching none was, it seemed, both measurable and significant. See Table 3.

Hidden in the control's immobile average are some disturbing individual changes, however. Note, for example, that student g03 scored 4 on pre-test question 1 but 1 on the post-test. Student g03 seemed to have forgotten three-fourths of the ethically pertinent information he or she knew earlier in the term. This is an extreme

Table 3 ECE 449, Spring 2010

	Pre	Pre	Pre	Pre	Pre	Pre	Sum	Post	Post	Post	Post	Post	Post	Sum	Ratio
	1	2	3	4	5	6		1	2	3	4	5	6		
u01	5	5	1	3	5	6	25	5	5	3	2	3	7	25	1.00
u02	0	3	0	0	3	7	13	5	6	0	0	0	6	17	1.31
u03	3	6	0	0	0	5	14	2	4	0	0	0	5	11	0.79
u04	1	7	0	0	0	2	10								
u05	0	4	1	1	3	4	13	1	2	0	1	3	3	10	0.77
u06	0	1	0	1	2	2	6	4	4	0	0	3	1	12	2.00
u07	5	4	0	0	0	6	15	3	2	0	0	0	1	6	0.40
Avg	1.88	4.00	0.63	1.13	2.25	4.75	13.7	3.00	3.57	0.86	1.00	2.00	4.14	13.5	1.044
Std	2.17	2.00	1.06	1.55	2.12	1.91	5.82	1.73	1.62	1.46	1.53	2.00	2.48	6.66	0.56
g01	3	11	0	8	1	5	28	7	8	0	5	1	5	26	0.93
g02	8	8	2	4	4	3	29	4	7	5	3	4	0	23	0.79
g03	4	4	0	0	3	2	13	1	3	0	0	2	2	8	0.62
g04	7	4	4	0	0	4	19								
g05	1	3	0	2	2	2	10	3	5	1	1	1	3	14	1.40
g06	3	2	0	1	2	3	11	6	5	0	2	4	4	21	1.91
g07	0	8	0	2	0	0	10	0	7	0	2	0	0	9	0.90
g08	5	7	3	1	6	6	28	5	6	5	0	0	7	23	0.82
g09	1	4	0	6	1	3	15	3	7	0	1	1	5	17	1.13
g10	7	5	0	1	0	5	18	3	5	0	2	0	5	15	0.83

Numbers under "Pre" or "Post" are question numbers. "U" indicates an undergrad; "G", a grad student. Each student is identified by numerals 01 through 10

case, but there are a fair number of others that are similar, if less extreme. This trend is cancelled overall by improvement in the scores of other students (and, in some cases, by improved scores of the same student on other questions). What is going on in this control? Is this simply random variation largely masked in 2008 and 2009 by the general trend of improvement (but recall that one student in 2008 who, on one question, mysteriously went from pre-test 10 to post-test 0)? Or do we have two trends more or less cancelling each other out, for example, a significant amount of ethical learning in some being cancelled by boredom in others who have largely lost interest in answering questions they answered earlier in the term and have no better idea how to answer now? Exit interviews might have answered these questions had we known they would arise and had we had the trained staff to do such interviews. These are now questions that future researchers should look into.[10]

Conclusion

These results are, of course, preliminary. We are dealing with small numbers of students. A class with one or two different students might have produced markedly different scores. There is certainly a need to reproduce this experiment on a larger scale—with methods more sophisticated than ours. Greater care should to be taken to exclude bias we may have introduced by allowing one person, the course instructor, to do all the grading. It would be interesting to track ethical sensitivity separately from ethical knowledge. Perhaps this sort of testing works better with one.

While we believe these results are important for assessment of teaching ethical sensitivity and knowledge, we do not want to overstate our accomplishments. What we have presented is what engineers call a "proof of concept": we now have reason to believe that our experiment of comparing pre- and post-test scores as ratios provides a reliable way to measure improvement in ethical sensitivity and knowledge across widely different courses, departments, and institutions—or, at least, seems to provide a way to measure such improvement worth further inquiry. We seem to have found a way to assess the effect of teaching ethics capable of establishing significant improvement in ethical sensitivity and knowledge without the need to develop a standardized test. We have taken one step toward that holy grail: comparing methods of teaching ethical sensitivity and knowledge across classes, departments, and institutions.

[10] We did not use the paired T test to check for statistical significance at the time of manuscript submission due to the minimal statistics training of one of the co-authors (Feinerman) and the virtual absence of statistical training in the other (Davis). In response to one reviewer's suggestion, we did. Feinerman used PASW 18 to run the paired T test. The resulting difference in means for the undergraduates and graduates was statistically significant in 2009 (with $p = .004$ and .053 respectively). Looking at the difference in means in 2010, the control year, the difference was not statistically significant, with $p = .722$ and .710 for undergraduates and graduates respectively. If the sum of the initial (T1) and final (T4) test responses are analyzed in 2008, the results are not statistically significant, but the difference is greater than in the control year with $p = .162$ and .296 for undergraduates and graduates respectively. The paired T test confirms that micro-insertion does teach the students ethics.

Acknowledgments Work on this paper was funded in part by a grant from the National Science Foundation (EEC-0629416). We should like to thank S&EE's editor and five reviewers for their extensive comments on earlier versions of this article.

References

ABET, Inc. (2009). Criteria for accrediting engineering programs, 2010–2011. http://www.abet.org/Linked%20Documents-UPDATE/Criteria%20and%20PP/E001%2010-11%20EAC%20Criteria%201-27-10.pdf.

Bebeau, M. J. (2002a). The defining issues test and the four component model: Contributions to professional education. *Journal of Moral Education, 31*(3), 271–295.

Bebeau, M. J. (2002b). Outcome measures for assessing integrity in the research environment (Appendix B), in integrity in scientific research: Creating an environment that promotes responsible conduct. National Academy Press, Washington, D.C. (available on NAP website: http://www.nap.edu/books/0309084792/html).

Bebeau, M. J. (2005). Evidence-based ethics education. *Summons, The Journal for Medical and Dental Defence Union of Scotland (Summer)*, 13–15.

Bebeau, M. J., & Thoma, S. J. (1999). Intermediate concepts and the connection to moral education. *Educational Psychology Review, 11*, 343–360.

Borenstein, J., Drake, M. J., Kirkman, R., & Swann, J. (2010). The engineering and science issues test (ESIT): A discipline-specific approach to assessing moral judgment. *Science and Engineering Ethics, 16*, 387–407.

Davis, M. (2006). Integrating ethics into technical courses: Micro-insertion. *Science and Engineering Ethics 12*, 717–730, esp. 726–727.

Davis, M., & Riley, K. (2008). Ethics across graduate engineering curriculum. *Teaching Ethics, 8*(Fall), 25–42.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64–74.

Institute of Medicine (IOM). (2002). *Integrity in scientific research*. Washington, DC: National Academics Press.

Kligyte, V., Marcy, R. T., Sevier, S. T., Godfrey, E. S., & Mumford, M. D. (2008). A qualitative approach to responsible conduct of research (RCR) training development: Identification of metacognitive strategies. *Science and Engineering Ethics, 14*, 3–31.

Loui, M. C. (2006). Assessment of engineering ethics video: *Incident at Morales*. *Journal of Engineering Education, 95*, 85–91.

Mumford, M. D., Devenport, L. D., Brown, R. P., Connelly, S., Murphy, S. T., et al. (2006). Validation of ethical decision making measures: Evidence for a new set of measures. *Ethics and Behavior, 16*, 319–345.

Riley, K., Davis, M., Jackson, A. C., & Maciukenas, J. (2009). 'Ethics in the Details': Communication engineering ethics via micro-insertion. *IEEE Transactions on Professional Communication, 52*, 95–108.

Sindelar, M., Shuman, L., Besterfield-Sacre, M., Miller, R., & Mitcham, C., et al. (2003). Assessing engineering students' abilities to resolve ethical dilemmas. In *Proc. 33rd annual frontiers in education 3 (November 5–8)*. S2A 25–31.

Suskie, L. (2004). *Assessing student learning: A common sense guide*. San Francisco: Joessey-Bass.