

**The Engineering and Science Issues Test (ESIT):
A Discipline-Specific Tool for Assessing Moral Judgment**

Jason Borenstein

Office of the Vice Provost for Research and School of Public Policy, Georgia Institute of
Technology

Matthew J. Drake

A.J. Palumbo School of Business Administration, Duquesne University

Robert Kirkman

School of Public Policy, Georgia Institute of Technology

Julie L. Swann*

Stewart School of Industrial Engineering, Georgia Institute of Technology

Abstract: We developed a new tool to assess ethics pedagogy in science and engineering. The Engineering and Science Issues Test (ESIT), measures moral judgment in a manner similar to the Defining Issues Test 2 but is built around technical dilemmas in science and engineering. We used a quasi-experimental approach with pre- and post-tests, and we compared the results those of a control group with no overt ethics instruction. Our findings are that several (but not all) stand-alone classes showed a significant improvement compared to the control group when the metric includes multiple stages of moral development. We also found the written test had a higher response rate and sensitivity to pedagogy than the electronic version. We do not find significant differences on pre-test scores with respect to age, education level, gender or political leanings, but we do on whether subjects were native English speakers and or had previous ethics instruction.

Keywords: engineering ethics, science ethics, assessment, moral judgment, ethics education

I. Introduction

* Corresponding author, jswann@isye.gatech.edu, 755 Ferst Drive, Atlanta, GA 30332-0205, (404) 894-2300.

The importance of ethics education for students in the biomedical sciences has been well-recognized. Yet the same recognition regarding students in engineering and related domains of scientific research is a rather recent phenomenon, driven in large part by changes in the Accreditation Board for Engineering and Technology, Inc. (ABET) accreditation criteria. Ethics education in technical fields is at an early and formative stage, and there is much potential work to be done.

To this point, there have in general been three approaches to ethics education for students preparing for careers in the technical professions. [1] The first is to require students to take a semester-long, stand-alone course in ethics, often taught by a philosopher; this may or may not be a course in engineering ethics as such. The second is to incorporate ethics modules into existing courses within the engineering curriculum. Introductory courses and senior design workshops are often selected for this purpose. A third approach combines the first two: require students to take a course in ethics while at the same time introducing ethics materials into engineering courses.

In choosing among these options, it would be helpful to know which is most likely to be effective in helping students to become responsible professionals. While everyone likely has their intuitive preferences and their hunches, there have not yet been many formal studies examining this question within the context of science and engineering education. A meta-analysis of abstracts given at the American Society for Engineering Education noted that most existing efforts toward ethics education in science and engineering has focused on curriculum development [2], though program assessment is at least acknowledged within the ABET Engineering Criteria 2000 as a significant challenge for programs to address [3].

Our previous investigation into the efficacy of different approaches to ethics education in science and engineering yielded results that were inconclusive as to the effect of the pedagogy [4]. We used the second edition of the Defining Issues Test (DIT-2) to measure general moral reasoning in a quasi-experimental study of engineering ethics education at Georgia Tech. Three groups of undergraduate students completed the DIT-2 both at the beginning and at the end of a semester (Spring 2004): students in a large lecture section of an engineering ethics course, students in an engineering course that included an ethics module, and, as a control, students in an engineering course with no overt ethics content. This investigation was the first published

account of using the DIT-2 for engineering, and not all experiments with the DIT-2 have used a control group.

The DIT-2 is a widely-used instrument for assessing moral judgment, that is, the ability to apply general moral principles to particular situations, which is widely acknowledged to be an important component of ethics education [5-7]. There are decades' worth of data from the DIT-2 and its predecessor, the DIT. The design of the DIT-2 and its interpretation are rooted in Lawrence Kohlberg's theory of moral development. According to this theory, individuals pass through a series of self-contained stages, each of which has its own distinctive cognitive structure.[8] Kohlberg's approach is hierarchical and unabashedly Kantian: simple avoidance of punishment reflects the lowest level of moral development, while principled reasoning about justice is the highest level. The developers of the DIT-2 have introduced some refinements to the theory, discussed in the next section, and now cast moral development in terms of three conceptual schemata that may overlap in the thinking of any one individual: the preconventional schema, the conventional schema, and the postconventional schema.

In assessing the impact of a particular educational intervention, comparing the results of a pre-test and a post-test using the DIT-2 promises to reveal the degree to which students have moved away from preconventional thinking and toward postconventional thinking, especially if their results can be compared to those from a control group. This, at least, was our hope as we began our previous study.

Results from that study did not turn out as expected. Taken on its own, the full course in professional ethics seemed to have a small impact on the moral reasoning of the students, as measured by the DIT-2, while the engineering course with the ethics module did not. When compared to the control group however, even this small impact turns out to be negligible. In short, we found no statistically significant effect on the moral reasoning of students who received either form of ethics instruction [4].

A number of rival hypotheses could account for these results. There is always the possibility that ethics education has no tangible impact on moral development in any mode (H_0). This would indeed be unfortunate, especially from our point of view as educators. We advocate reserving judgment on H_0 until other hypotheses have been tested.

Another possibility (H_1) is that we were trying to measure the wrong thing. This could reflect a problem with the theory behind the test. The DIT-2 is a test of moral judgment, which

is only one aspect of moral experience and conduct [9], and perhaps not the one most readily addressed by ethics education. H_1 is the subject of another investigation in which we are engaged: the development of a discipline-specific test of ethical sensitivity.

Yet another possibility (H_2) is that a measure of moral reasoning more precisely tailored to the practical context of engineers and researchers would be more sensitive to the effects of professional ethics education. The dilemmas included in the DIT-2 are very general in nature, drawn from the domains of what might be called personal and social ethics. It should be possible, we hypothesized, to develop a new version of the instrument with dilemmas that are much more like those encountered by engineers and researchers in their professional lives.

The developers of the DIT-2 (along with other experts on ethics education) have long recognized the usefulness and need for profession-specific assessment tools and encouraged the development of more of these [9, 10] but few exist to date. Some exceptions include a version of the Defining Issues Test created for lawyers and doctors, and several versions for dentists (see Bebeau [9] for a summary).

In this paper, we report the first results from our effort to test H_2 . We developed a new instrument called the Engineering and Science Issues Test (ESIT), modeled on the DIT-2 but tailored to science and engineering, and we administered it to 277 students at Georgia Tech during the academic year 2005-2006. We took a quasi-experimental approach very similar to the one used in our previous study.

Two different modes of implementation were used during the course of the study, electronic and written. In principle, an electronic version of the test is desirable for a variety reasons, including the fact that it makes the test easier to administer with large groups of subjects. However, an electronic version of the test may offer a set of challenges beyond a paper version of the test in getting thoughtful serious answers.

In this paper we describe the original DIT-2 test and the ESIT test that we developed. We provide details of the experimental study, and we measure the test for general validity as a measure of the impact of ethics education on moral judgment. We also use the test to assess whether there is an impact on context-specific moral judgment from several courses that count toward ethics education requirements at Georgia Tech. Finally, we identify several areas of future development and analysis that will continue to improve assessment in ethics education, especially in science and engineering.

II. Methods

A. Background on the DIT-2

James Rest and his collaborators departed somewhat from Kohlberg's methods in developing the original DIT. While Kohlberg set a production task for his subjects, recording in detail their responses to moral dilemmas, the DIT sets a recognition task that more or less amounts to a multiple-choice test. The answer sheet to the DIT-2 can be scanned electronically, making it possible to gather and process data from a large number of subjects in a single study. The reliability and validity of the DIT were established by thirty years of testing, during which time it was found to be particularly sensitive to the impact of ethics education programs. [11, 12]

More importantly, by the time they developed the DIT-2 Rest and his colleagues had introduced significant theoretical refinements to the Kohlbergian framework. Rather than a linear series of stages, they have come to think in terms of three distinct moral schemata, all of which may be active in the mind of an individual at the same time. The preconventional schema is characterized by narrow personal interest, the conventional schema by an appeal to duty and to maintenance of the existing social order, and the postconventional schema by the search for moral ideals on which a social order ideally ought to be based. [11, 13] Note that the postconventional schema is no longer confined to a strictly Kantian conception of justice: it can include any moral outlook that offers a critical perspective on the social status quo.

The DIT-2 itself consists of a set of five moral dilemmas, each stated in a single paragraph. Each dilemma is accompanied by twelve questions that raise issues about the dilemma, a mix of preconventional, conventional, and postconventional considerations. So, for example, a dilemma about a starving man who contemplates stealing from a rich man in his village is accompanied questions about the fear of getting caught, upholding community laws, and the conflict between laws and the basic claims of a society's members [14]. There is also a scattering of nonsense questions throughout the test, in part as a way of identifying data from subjects who are not taking the test seriously.

Subjects are to rate each of the questions on its importance as an issue raised by the dilemma (from 1=great importance to 5=no importance), after which they are to rank the four most important issues. Responses are analyzed to determine the degree to which

postconventional thinking is prevalent (P-score, see below) and the degree to which postconventional thinking is present and preconventional thinking is absent (N2 score, see below).

B. Developing the Engineering and Science Issues Test

Three of us (Borenstein, Kirkman, and Swann) collaborated in the development of the new instrument, the ESIT. Together we developed six case studies of about one paragraph in length, each of which is intended to reflect an ethically problematic situation that a scientist or engineer might reasonably expect to confront in professional practice. We created one case study more than the number of cases in the DIT-2, in order to leave open the possibility of removing one later if we find that it is not effective at measuring moral reasoning. Here is one example of a case from the ESIT:

Engineer Jameson owns stock in RJ Industries, which is a vendor for Jameson's employer, Modernity, Inc., a large manufacturing company. Jameson's division has been requested by management to cut one vendor: either RJ Industries or Pandora Products, Inc. Pandora Products makes a component that is slightly higher in quality and slightly more expensive than that made by RJ Industries. Management and the other engineers in her division do not know that Jameson has a financial interest in one of the two vendors. Jameson is unsure whether she should participate in the decision.

As with the DIT-2, each case description in the ESIT is followed by a set of twelve questions representing different issues that may or may not have been raised within the case. For example, one of the questions associated with the above case is, "Will Jameson's decision potentially cause harm to the public?"

Each of us worked independently to create a subset of the questions for each case, with examples from each of the three cognitive schemata recognized in neo-Kohlbergian theory: preconventional, conventional, and postconventional. The twelve questions associated with each case are divided roughly equally among these three categories. We also scattered a total of 6 "nonsense" questions throughout the test in order to assess a student's ability to discern whether an issue is not ethical in nature or is not relevant to the case under consideration.

Once we developed the bank of questions, we worked independently to assess the questions and determine appropriate rating for each one: preconventional, conventional, postconventional, or “nonsense.” We then compared our respective ratings for each question, kept those on which we reached consensus, and revised or replaced questions about which we disagreed. This was an iterative process at the end of which we all agreed on the ratings for all of the questions included in the ESIT. We agreed, for example, that the question about public harm in the Jameson case would be an instance of postconventional reasoning.

In the final step, we randomized both the order of the cases and the order of their associated questions in order using a standard random number generator to minimize the potential for schema bias in the construction of the test itself.

C. Taking the ESIT and Research Design

As with the DIT-2, subjects taking the ESIT are presented with a recognition task. They are asked first to rate each question in terms of the importance of the issue it raises, from 1 (great importance) to 5 (no importance). Subjects are then asked to rank the four issues they consider to be most important for making a decision in the case. The data collected from subjects can then be analyzed in the same way as data from the DIT-2, scoring the prevalence of postconventional thinking and the non-prevalence of preconventional thinking. The test includes a brief set of instructions at the beginning along with an example to show rating and ranking.

The ESIT was administered to some students electronically via the Internet and to others in a written form. In almost all classes, students received class credit for participating in the experiment, with an alternate assignment possible if they did not want to take the survey.

In the current study, we followed the same procedure we used in our previous study: each participating student filled out the ESIT twice, once at the beginning of the school term and once at the end. Students were assigned an identification number that combined information about the course in which the student was enrolled as well as randomized digits to protect their identity. This number enabled us to conduct a matched-pairs design by comparing each student’s pre- and post-test answers. For the purposes of validating the test, pre-test scores were used.

The students included in the study had the ability to enroll in particular courses on their own, which entails that they were not randomly assigned to the courses. Thus, the design of the study is quasi-experimental in nature. It should be noted that engineering students, whose majors

may require them to fulfill ethics or humanities requirements, frequently take the courses included in the survey, which might constitute a form of selection bias.

We excluded tests from the analysis if they met one of several criteria. These were designed to make sure the subject took the test seriously and/or understood the instructions. For example, survey results were excluded from the analysis if a significant portion of the survey remained incomplete or if the answers seemed to reflect a lack of thought as measured by the nonsense score. Incomplete surveys were more frequently a problem with regard to the electronic version of the survey, which is discussed below in more detail. Several guidelines are similar to those used for DIT-2 analysis; specifically, we excluded any tests with the following characteristics:

- Failed to complete 24 rating questions (equivalent to 2 dilemmas)
- Failed to complete 9 ranking questions (approximately 2 dilemmas)
- Received a “nonsense” score of 11 or more points, where a nonsense item ranked as the most important issue incremented the score by 4, a nonsense item ranked the 2nd most important issue incremented the score by 3, and so on; the score was summed over all nonsense items ranked

Applying the above metrics to the students’ responses in the order they are presented, we omitted the following numbers of electronic and written tests:

- 30 electronic pre-tests
- 25 electronic post-tests
- 7 written pre-tests
- 2 written post-tests

The majority of these excluded responses were identified because the respondents failed to complete enough ranking questions. One possible explanation for this omission is that the students may not have read the survey instructions closely enough to understand their task in ranking the questions. We noticed this same difficulty in our previous experience with the DIT-2. In the electronic form of the test, this issue can be exacerbated since students who take the test electronically did not have real-time access to an instructor in order to clarify the instructions.

We used two main measures of moral judgment on the ESIT test. The first of these is similar to the P-score in the DIT-2, which measures the percentage of postconventional reasoning weighted by the importance given in the ranking scores compared to the maximum

score possible. The index also normalizes for rankings that are omitted. The second measure used is similar to the N2-score in the DIT-2, which accounts for the postconventional thinking that is present and the pre-conventional thinking that is absent. The specific calculations that we are using are as follows:

- $P\text{-SCORE} = (4 * \text{the number of post-conventional issues ranked first} + 3 * \text{the number of post-conventional issues ranked second} + 2 * \text{the number of post-conventional issues ranked third} + 1 * \text{the number of post-conventional issues ranked fourth}) / (60 - 4 * \text{the number of first rankings omitted} - 3 * \text{the number of second rankings omitted} - 2 * \text{the number of third rankings omitted} - 1 * \text{the number of fourth rankings omitted})$
- $N2\text{-SCORE} = P\text{-SCORE} - 3 * (\text{average rating on pre-conventional issues} - \text{average rating on post-conventional issues}) / \text{standard deviation of pre- and post-conventional issues}$

For more detail on the original indices in the DIT-2 and the rules used to exclude surveys, see [15, 16].

D. Study Population

Five undergraduate courses and one graduate course at Georgia Tech were included in this study.¹ Students from the experimental group were enrolled in “Ethics and the Technical Professions” (ETP), “Ethical Theories” (ETh), or “Science, Technology & Human Values” (STHV) during the time when they were participants in the study. These are three-credit, semester-length courses taught by full-time faculty in the School of Public Policy. Students in the control group were not enrolled in a course with a significant ethics component during the time when the study was being conducted. The ESIT survey was administered in the same manner to these students as it was to students in the experimental group. Table 1 shows the number of responses by class and test type. The total number of students who completed both the pre and post tests was 277 overall.

The electronic version of the ESIT was administered during the fall semester of 2005. The students from participating classes were sent an electronic link that they had to access within

¹ Note that approval was obtained from Georgia Tech’s Institutional Review Board (IRB) for human subjects research prior to the beginning of the study.

a defined time period of a few days. The expectation was that students would take the survey during non-class time. The link was given to the students on two separate occasions during the course of the semester, near the beginning and close to the end. The fall 2005 administration of the survey included four sections of ETh, one section of ETP, and one section of STHV in the experimental group. The control group consisted of six sections of an undergraduate introduction to industrial engineering course, one graduate course in probabilistic models, and an undergraduate course in U.S. government.²

The written version of the ESIT was administered during the spring and summer semesters of 2006. In most classes, the survey was given to the students to complete during regular class time. The written version was used to increase the response rate of students in the participating courses and because there was some concern that electronic surveys may not be taken as seriously as written versions; this is also evidenced by the larger number of electronic tests excluded based on thresholds described above. The pre-test and complete experiment response rates for the electronically administered test were very low, but fortunately, the written test produced a significantly higher response rate (77.8% in the written exam versus only 32.8% in the electronic exam). The spring 2006 written test session included two sections of ETh and one section of STHV in the experimental group, and the control group for the written test was one undergraduate section of statistics in the summer of 2006.

Table 1 shows that a total of 104 students took the pre-test but dropped out of the experiment before taking the post-test. We analyzed the pre-test scores of this group of students compared with the 277 participants who completed the entire study and survived the exclusion scrutiny described in Section II.C. The difference in the pre-test scores for both the P-score and N2 measures were significant at the 1% level, which identifies a distinction between the moral reasoning development of the students who dropped out of the study and those who finished it. Further analysis obtained by examining the electronic and written tests separately showed that this bias exists only for the electronic version of the test. The electronic test included 90 of the 104 students who dropped out of the experiment, and both measures showed significant differences at the 5% level; neither the P-score nor the N2 score showed significance for the written test. One explanation for this result is that the lack of monitoring of the students in the

² The U.S. Government course is the only one that did not offer the students credit for participating, and it is also the only class with a high number of freshmen, both of which may explain why there are few subjects from this class who completed the full experiment.

electronic test made it easier for students to drop out of the study. We speculate that the conditions of the electronic test would in effect select for those students who already have a degree of self-discipline or moral motivation that may be correlated with moral reasoning.

Electronic ESIT (Fall 2005)

Enrollment / Responses	ETP	ETH	STHV	Control	Total
Students enrolled (N)	22	116	169	175	482
Pre-test only (N)	3	14	33	40	90
Pre-response rate (%)	68.2	54.3	49.7	49.1	51.5
Completed experiment (N)	12	49	51	46	158
Final response rate (%)	54.5	42.2	30.2	26.3	32.8

Written ESIT (Spring / Summer 2006)

Enrollment / Responses	ETH	STHV	Control	Total	BOTH TESTS
Students enrolled (N)	74	33	46	153	635
Pre-test only (N)	4	4	6	14	104
Pre-response rate (%)	90.5	84.8	82.6	86.9	60.0
Completed experiment (N)	63	24	32	119	277
Final response rate (%)	85.1	72.7	69.6	77.8	43.6

Table 1: Study population for ESIT experiments

Clearly, in addition to differences in the form of the test, students’ moral development over the course of the semester is likely to depend in part on the particular circumstances under which ethics instruction took place. The experimental group included students enrolled in a number of different courses, with considerable variation in stated objectives, course design, reading assignments, and forms of evaluation. The primary textbook used in the “Ethics and the Technical Professions” courses is Harris, Pritchard, and Rabins, *Engineering Ethics: Concepts and Cases*, third edition [17]. The course focuses heavily on case studies to identify and examine ethical issues in engineering practice. The textbooks used in the sections of the “Ethical Theories” course included in the study are Rachels’ *The Elements of Moral Philosophy*, fourth edition [18], and Mappes and Zembaty’s *Social Ethics: Morality and Social Policy*, sixth edition [19]. The content of the course is roughly divided between the theoretical dimensions of ethics and contemporary ethical issues. The textbooks and materials used in “Science, Technology & Human Values” vary depending on the semester and on which professor is instructing the course.

Table 2 breaks down the students' demographic statistics for the entire study by course grouping. The vast majority of the participants in both the experimental and control groups were engineering majors. Most participants were either junior- or senior-level undergraduate students, and the few graduate students in the study were a part of the control group. There were also approximately four times as many males in the study as females. The majority of students were native English speakers, although approximately 11% of the respondents spread across the experimental and control groups reported a primary language other than English; the graduate control class had a higher percentage of non-native English speakers. In fact, 30.8% of the students in the control classes as a whole reported that their first language was something other than English; whereas, only 5.5% of the students in the experimental classes were non-native English speakers.

Demographic Group	ETP	ETH	STHV	Control	Total
Majors					
Engineering	10	93	66	72	241
Science and computer science	2	5	5	3	15
Management and humanities	0	8	3	1	12
No response	0	6	1	2	9
Educational Level					
Freshman	0	6	2	4	12
Sophomore	2	11	3	14	30
Junior	4	26	13	37	80
Senior	6	68	57	8	139
Graduate	0	0	0	14	14
No response	0	1	0	1	2
Sex					
Male	12	85	62	57	216
Female	0	25	13	19	57
No response	0	2	0	2	4
Native Language					
English	12	104	71	57	244
Language other than English	0	7	4	20	31
No response	0	1	0	1	2
All	12	112	75	78	277

Table 2: Demographic breakdown by class

III. Results

A. Validity and Reliability of the ESIT

There are several measures to determine validity of a test or measure in moral education. As identified by Rest, et al. [11], these include whether the test is impacted by ethics education, whether the measure improves with age or education level, and whether subjects show significant improvement in a longitudinal study. The sample size of 277 respondents is too small to conduct a comprehensive validity check on the ESIT instrument. However, we were able to examine several characteristics of the data points in this study that are relevant to the validity of the instrument.³

One of the demographic questions on the ESIT asked respondents to report the amount of their prior exposure to ethics instruction that they had coming into the current course. They could choose any, all, or none of the following options: (i.) Dedicated ethics course for the technical professions; (ii.) General ethics or philosophy course; (iii.) Some ethics content in other courses; and (iv.) Other. If the P and N2 scores computed from the ESIT are valid in measuring moral reasoning ability, we would expect that students with previous exposure to ethics would score higher on the pre-test than those who have never had any kind of formal training.

Only 50 students, roughly 1/6 of the study, had no prior exposure to ethics instruction. Two-sample t-tests conducted on the pre-test P and N2 scores identified significant differences between the average scores of those students with prior ethics experience and those with no experience. (See Table 3 for the detailed results of these tests.) This provides some support for the validity and reliability of the ESIT instrument.

It is clear from the results of other analysis, however, that more testing is necessary to confirm the validity of the ESIT. The developers of the DIT and DIT-2 found that scores tended to be higher for respondents who were older rather than younger and had achieved a higher general level of education. In our study, correlations between the ESIT pre-test scores and either age or educational level were not significantly different from zero. The fact that age was not correlated with either of the score was not surprising, however, since most of the respondents were between 19 and 25 years old; the DIT exams, in contrast, were administered to a broader

³ The developers of the DIT and DIT-2 assessment tools for general moral reasoning accumulated results on the validity of their instrument through several decades of administering the test to respondents of all ages and professions. We plan to add to these results as the test continues to be used by ourselves and others.

range of age groups. Figure 1 depicts the spread of respondents' educational levels in each course. Clearly most of the students were junior- or senior-level undergraduate students, so we need additional responses from students at other levels of education to make more definitive statements about the correlations between respondents' educational levels and their ESIT scores.

Experienced Students vs. No Experience	N	Pre-test P	Pre-test N2	P Diff p Value
Students with any reported ethics experience	227	0.512 (0.009)	2.940 (0.096)	0.005***
Students with no reported ethics experience	50	0.449 (0.020)	2.340 (0.240)	N2 Diff p Value 0.023**

Table 3: Comparison on ESIT pre-test means (standard errors) of students with and without prior ethics experience. [*,**,***] denotes statistical significance at the [10%, 5%, 1%] level.

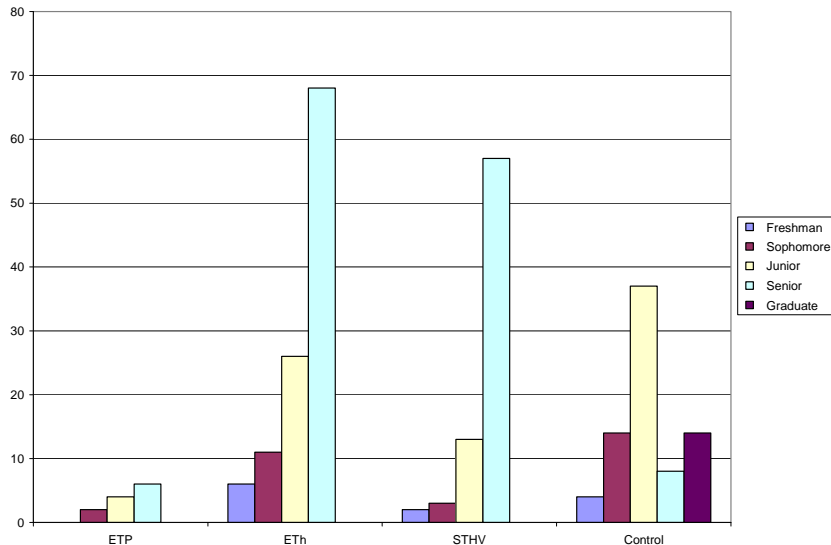


Figure 1: Distribution of education level by class type

B. Using the ESIT to Measure the Effect of Ethics Instruction

To analyze the effect of a semester of ethics instruction on the development of the students' moral reasoning ability, we measured the initial scores on the ESIT (pre), and scores at the end of the semester (post), and the change in students' scores on the ESIT from the pre-test to the post-test. We looked at both the P and N2 scores for each of these measures. The most important measure of the effect of ethics instruction is the change in scores that was achieved from the beginning of the semester to the end. We compared the change in scores of students in the experimental groups to that of students in the control groups in order to ensure that what we

are measuring is the effect of pedagogy rather than other effects such as the effect of retaking the test.

Separating the students who completed the entire study into the experimental and control groups, we conducted two-sample *t*-tests on the differences in each score for the two groups; the results of these tests are provided in Table 4 across several groupings. We mark the results in the table that are significant at the 10% level or better. We also found that if we excluded fewer written tests according to the rules above that the same comparisons remained significant but at a stronger level.

One observation that holds in all cases is that the N2 score exhibited statistically significant improvement due to ethics instruction in many of the *t*-tests that we conducted; on the other hand, the P-score did not show significant differences between the experimental and control groups. The developers of the original DIT and DIT-2 tests prefer to use the N2 score because it makes use of both the respondents' relative ranking of the post-conventional issues as well as the respondents' ability to rate the issues corresponding to lower stages of Kohlbergian development lower than the higher stages; the limitation of the P-score is that it only considers the ranking data. As the ESIT continues to be used, the data will enable a more complete understanding of whether the P-score will be useful for measuring moral reasoning in science and engineering.

The overall results show that there is a significant improvement in the overall population of those who had ethics education as compared to those who did not, as measured by the N2 score on the ESIT. However, further analysis shows that the effect is significant for the subjects who took the written test (as compared to the written control) but not for those who took the electronic version (as compared to the electronic control). This suggests that the written test may provide more validity overall, although this could be due to several reasons including that subjects might dedicate more time to the written version or that the people who complete the entire electronic version are pre-selecting for those who are more amenable to ethics instruction.

Another interesting comparison is looking at the curriculum and whether that has an impact of outcomes in moral reasoning. The ETP and ETh curricula are most similar in terms of content, therefore we grouped them for this analysis. We find that the ETP and ETh classes have a positive impact on moral reasoning (measured by N2) compared to the control group, while the STHV curriculum did not. This could be due to several factors. One possibility is that the STHV

curriculum is not as effective at changing moral reasoning, since it is focused more broadly on the social dimensions of technology rather than on ethical decision making *per se*. The STHV class in the experiment was also a much larger class than the other experimental classes, so this could also be a factor. In addition, it was taught by a different instructor than ETP and ETh. However, we conjecture that the primary effect may be due to the differences in content of the course. This suggests that some ethics course curricula could be more effective than others at achieving particular outcomes of moral education.

Other initial results that could relate to curriculum are found in comparing the ETP and ETh classes individually to the controls; in this we found that t-tests for the ETP class as compared to the control was significant but the other classes were not. This is especially interesting since the one section of the ETP class took the electronic version of the test. It should be noted this class had only 22 students, which could also be a reason for the improvement in scores since smaller class sizes generally allow for the kind of discussion that is more likely to lead to moral development [20]. In general there are a number of variables relating to the size of the course, the course curriculum, and the instructional techniques used in the course that subsequent studies will add to the data set, enabling us to test the impact of curricula or class size more formally.

Thus, the overall results of the ESIT test shown in Table 4 suggest that ethics instruction has a positive impact on students' moral reasoning ability in science and engineering since there is a significant difference between the N2 scores of the experimental group and the control group. This is interesting in light of our previous results that general moral reasoning scores did not show an impact from the ethics education for science and engineering at Georgia Tech [4].

Overall Experimental vs. Control Group	N	Pre-test P	Post-test P	P-Score Diff	Pre-test N2	Post-test N2	N2 Score Diff	P Diff p Value
Overall experimental	199	0.508 (0.009)	0.531 (0.010)	0.022 (0.009)	2.964 (0.102)	3.341 (0.108)	0.380 (0.082)	0.648
Overall control	78	0.480 (0.015)	0.495 (0.017)	0.015 (0.012)	2.489 (0.186)	2.605 (0.192)	0.115 (0.110)	N2 Diff p Value 0.053*

ETP and ETh Curriculum vs. Control Group	N	Pre-test P	Post-test P	P-Score Diff	Pre-test N2	Post-test N2	N2 Score Diff	P Diff p Value
ETP and ETh experimental	124	0.516 (0.012)	0.549 (0.013)	0.033 (0.011)	3.153 (0.119)	3.512 (0.124)	0.360 (0.092)	0.280
Overall control	78	0.480 (0.015)	0.495 (0.017)	0.015 (0.012)	2.489 (0.186)	2.605 (0.192)	0.115 (0.110)	N2 Diff p Value 0.086*

STHV Curriculum vs. Control Group	N	Pre-test P	Post-test P	P-Score Diff	Pre-test N2	Post-test N2	N2 Score Diff	P Diff p Value
STHV experimental	75	0.497 (0.014)	0.501 (0.016)	0.004 (0.015)	2.651 (0.183)	3.059 (0.197)	0.410 (0.160)	0.573
Overall control	78	0.480 (0.015)	0.495 (0.017)	0.015 (0.012)	2.489 (0.186)	2.605 (0.192)	0.115 (0.110)	N2 Diff p Value 0.123

Electronic Experimental vs. Electronic Control	N	Pre-test P	Post-test P	P-Score Diff	Pre-test N2	Post-test N2	N2 Score Diff	P Diff p Value
Electronic experimental	112	0.498 (0.012)	0.520 (0.014)	0.021 (0.012)	2.693 (0.141)	2.392 (0.247)	0.380 (0.110)	0.289
Electronic control	46	0.485 (0.017)	0.485 (0.021)	0.000 (0.016)	3.076 (0.154)	2.554 (0.255)	0.160 (0.160)	N2 Diff p Value 0.255

Written Experimental vs. Written Control	N	Pre-test P	Post-test P	P-Score Diff	Pre-test N2	Post-test N2	N2 Score Diff	P Diff p Value
Written experimental	87	0.522 (0.014)	0.545 (0.015)	0.023 (0.014)	3.312 (0.140)	3.682 (0.140)	0.370 (0.120)	0.555
Written control	32	0.473 (0.028)	0.510 (0.029)	0.037 (0.019)	2.630 (0.284)	2.677 (0.296)	0.047 (0.130)	N2 Diff p Value 0.075*

Table 4: Comparisons of pre, post, and (post – pre) moral reasoning scores between experimental groups and control groups on the ESIT. Number (N) is indicated, along with mean (standard error). [*,**,***] denotes statistical significance at the [10%, 5%, 1%] level.

C. Analysis by Demographics, Political Leanings, and Native Language

Some researchers have found a relationship with measures of moral reasoning on the DIT-2 and other factors such as age, gender or major. For the ESIT, dividing the students into groups based on their age, major, gender, and educational level yielded no significant results with respect to the P and N2 pre-test to post-test improvement scores. The correlations between the improvement in each ESIT score and both age and educational level were not significant even at the 10% level. This suggests that age and educational experience of the students does not impact the effectiveness of ethics instruction on moral reasoning development. However, we did find that the correlations were significant between the raw pre-test and post-test scores with age and educational level, but these correlations were negative. However, as we note elsewhere, we had very few students outside of a narrow age range in the study, and this factor is confounded with the fact that many of them, especially at the graduate level, are non-native English speakers (see discussion below).

An analysis of the respondents' improvement scores with respect to additional demographic variables did not yield any significant results. Two-sample t-tests for the difference in P and N2 scores between engineering students and non-engineering students were not significant at the 10% level either, but the power of this test is limited since only 34 non-engineering majors completed the entire study. A final set of t-tests found no significant difference in the improvement in ESIT scores between male and female respondents, which provides some evidence that the ESIT does not contain gender bias.

There have also been some links found with measures on the DIT-2 and political leanings: some researchers have claimed that the DIT-2 has in effect a left leaning bias, serving as a gauge of the subjects' affinity for liberal political positions [11]. The demographic section of questions on the ESIT provides an opportunity for respondents to report their self-perceived political affiliation. The possible responses are as follows: (i.) Very liberal; (ii.) Somewhat liberal; (iii.) Neither liberal nor conservative; (iv.) Somewhat conservative; and (v.) Very conservative. Correlations between the respondents' self-reported level of conservatism and all pre-test, post-test, and improvement scores were not significantly different from zero. This is the same result we found in our previous study using the DIT-2, and the initial results suggest that the ESIT is measuring something other than political leanings [1].

Analyzing the students based on their native languages, we found a significant difference in the overall results for native English speakers compared with non-native English speakers. Table 5 shows a significant difference between the pre-test N2 scores of the two groups at the 1% level. This could be due to several reasons, including potential cultural bias of the test, differences in beliefs about ethics across cultures, or simply a greater understanding of what is being asked for by native-English speakers.

Examining each group of students to see the effect of ethics instruction produced a similar result. The improvement in N2 scores for native English speakers in the experimental group was significantly different from the improvement scores for native English speakers in the control group at the 1% level. This difference, however, was not significant among the non-native English speakers. The significant difference in N2 improvement scores for the English-speaking students could either be a result of the fact that the moral reasoning development of native English speakers is more responsive to classroom ethics instruction in English or that the ESIT is able to detect moral reasoning development in native English speakers more easily.

Benchmark studies of the DIT-2 have generally used native-English speakers for benchmark comparisons. However, addressing cultural issues is certainly important in science and engineering, where many people come from different cultures and languages, especially at the more advanced levels of education. Clearly, more testing is required to verify any suggested cultural bias in the ESIT or identify differences needed in ethics education to address cultural differences.

Overall English vs. Non-English Group	N	Pre-test P	Post-test P	P-Score Diff	Pre-test N2	Post-test N2	N2 Score Diff
Overall English group	244	0.503 (0.008)	0.523 (0.009)	0.020 (0.008)	2.910 (0.094)	3.200 (0.099)	0.300 (0.072)
Overall non-English group	31	0.489 (0.026)	0.509 (0.029)	0.020 (0.018)	2.210 (0.320)	2.660 (0.350)	0.455 (0.150)
p-values for differences between the two groups		0.609	0.653	0.989	0.041**	0.147	0.355

Experimental vs. Control for English Group	N	Pre-test P	Post-test P	P-Score Diff	Pre-test N2	Post-test N2	N2 Score Diff	P Diff p Value
Experimental English group	187	0.508 (0.010)	0.531 (0.011)	0.023 (0.010)	2.994 (0.106)	3.376 (0.109)	0.380 (0.086)	0.507
Control English group	57	0.485 (0.018)	0.496 (0.018)	0.011 (0.015)	2.626 (0.200)	2.637 (0.213)	0.011 (0.120)	N2 Diff p Value 0.015**

Experimental vs. Control for Non-English Group	N	Pre-test P	Post-test P	P-Score Diff	Pre-test N2	Post-test N2	N2 Score Diff	P Diff p Value
Experimental non-English group	11	0.529 (0.039)	0.531 (0.031)	0.002 (0.033)	2.558 (0.422)	2.976 (0.562)	0.419 (0.270)	0.472
Control non-English group	20	0.466 (0.033)	0.497 (0.041)	0.031 (0.022)	2.018 (0.433)	2.493 (0.451)	0.475 (0.190)	N2 Diff p Value 0.867

Table 5: Analysis of means (standard errors) of native English speakers versus non-native English speakers. [*,**,***] denotes statistical significance at the [10%, 5%, 1%] level.

IV. Conclusions and Future Research

This study is our second step in an ongoing effort to assess ethics education in science and engineering. In the first, we found that the instruction did not have a significant impact on general moral reasoning (as measured by the DIT-2) as compared to the control group, but it was not clear if we were trying to measure the wrong outcome or if the test was not finely tuned enough to the outcome that we were trying to measure. In this paper we have described our efforts to test hypothesis H₂, which says that the initial test did not provide a measure sensitive enough to the effects of ethics education specific to the science and engineering disciplines. Overall, the results of our study with the ESIT would tend to support H₂, which in no way excludes the possibility that H₁ might also have some validity, that is, the hypothesis that ethics education more profoundly effects components of ethical decision making other than moral

judgment, such as ethical sensitivity. H_1 is the focus of a separate research project, now underway.⁴ We do take our results as counting against H_0 , the hypothesis that ethics education has no measurable impact.

The instrument we developed, the Engineering and Science Issues Test (ESIT), provides a set of moral dilemmas, which asks students to rate and rank a set of questions that raise ethical issues that may bear on an ethical decision about the dilemma. We found the ESIT has some validity in measuring moral education, as measured by pre-tests of those who have and have not had ethics education previously. The validity of the ESIT requires extensive testing in a variety of institutional contexts; and with a much broader range of participants (in terms of age, background, etc.); the validity will be further assessed as it continues to be used by ourselves and others.

We found that the ESIT did find improvement in moral reasoning due to the ethics education at Georgia Tech when measured by the increase in postconventional reasoning and the decrease in preconventional reasoning (N2 score). Interestingly, measuring the postconventional reasoning alone did not show an effect from ethics instruction, which is different from our results using the DIT-2. This is an area that bears further study, including refinements of the measures as well as creation of other ones.

We found that the written form of the test showed more improvement than did the electronic form of the test, though this could be partly due to self-selection of those completing the electronic version of the test. Given the written version of the test, a much larger percentage of students surveyed completed the test; thus, it is more likely to be effective at measuring outcomes. We found that some courses may be better than others at improving moral reasoning, although this is another area that would benefit from further study to understand the factors that lead to better outcomes.

We welcome other researchers to make use of the ESIT in order to assess the impact on ethics education on students thus adding to the database of knowledge of moral reasoning outcomes in science and engineering.⁵ This will help to increase understanding of the validity of the test, contribute to further refinements of the test, and generate understanding of factors in ethics education that help to achieve particular outcomes. We recommend several specific areas

⁴ We are developing and testing an assessment instrument called the Test of Ethical Sensitivity in Science and Engineering (TESSE). Please contact the authors for further information.

⁵ Please contact the authors for the most recent version of ESIT or to be on a distribution list for future tests.

for further experimentation including studies to further tease out the effect of different curricula content on educational interventions (including ethics in discipline-specific classes, ethics education delivered over the internet, or graduate research ethics) and impact of the size of the class. Other specific areas that could include more data include the effect of the test as it relates to education level, age, or across different cultures and religious beliefs as is done for the DIT-2.

Ethics education is important, in science and engineering as well as many other fields. In some professions, there has been extensive analysis done to understand what kinds of ethics education is effective in fostering responsibility and good decision making on the part of professionals. Scientists and engineers face complex situations that call for an ethical response, in graduate research as much as in industrial practice. This was recognized by ABET when it strengthened its ethics requirements in its revised accreditation criteria. The need for more attention to ethics education in science and engineering is also highlighted by very public examples of moral failings on the part of scientists and engineers. We hope that the work we have presented here is a useful contribution to an understanding of what it is that makes ethics education effective in such contexts, though we fully recognize that there is much more work to be done.

Acknowledgments

This research was funded in part by a grant from the College of Engineering Undergraduate Initiative at the Georgia Institute of Technology, and in part by a Focused Research Program grant from the Office of the Vice Provost for Research at the Georgia Institute of Technology. In addition, Dr. Swann was supported in part by NSF DMI-0348532.

We would like to acknowledge Dr. Harry Sharp for his help with developing the scanning form for the test and converting tests to raw data and Mr. Andy Haleblian for help in creating the electronic version of the ESIT.

References

- [1] Herkert, J.R., "Engineering ethics education in the USA: content, pedagogy and curriculum", *European Journal of Engineering Education* Vol. 25, No. 4, 2000, pp. 303-313.
- [2] Haws, D.R., "Ethics Instruction in Engineering Education: A (Mini) Meta-Analysis", *Journal of Engineering Education* Vol. 90, 2001, pp. 223-229.
- [3] Herkert, J.R., "ABET's Engineering Criteria 2000 and engineering ethics: Where do we go from here?", *International Conference on Ethics in Engineering and Computer Science*: <http://www.onlineethics.org>, 1999.
- [4] Drake, M., P. Griffin, R. Kirkman, and J. Swann, "Engineering Ethical Curricula: Assessment and Comparison of Two Approaches", *Journal of Engineering Education* Vol. 94, 2005, pp. 223-231.
- [5] Harris, C.E., M. Davis, M.S. Pritchard, and M.J. Rabins, "Engineering ethics: what? why? how? and when?" *J. Engineering Education* Vol. 85, 1996, pp. 93-96.
- [6] Newberry, B., "The Dilemma of Ethics in Engineering Education", *Sci. Eng. Ethics* Vol. 10, 2004, pp. 343-351.
- [7] Narvaez, D., and J. Rest, "The four components of acting morally", *Moral behavior and moral development: An introduction*, New York: McGraw-Hill, 1995.
- [8] Kohlberg, L., "Stage and Sequence: The Cognitive-Developmental Approach to Socialization", *The Psychology of Moral Development: The Nature and Validity of Moral Stages*, San Francisco: Harper and Row, 1984, pp. 7-169.
- [9] Bebeau, M.J., "The Defining Issues Test and the Four Component Model: Contributions to professional education", *J. Moral Educ.* Vol. 31, No. 3, 2002, pp. 271-295.
- [10] Rest, J., and D. Narvaez, (Eds), *Moral development in the professions: Psychology and applied ethics*, Hillsdale, NJ: Lawrence Erlbaum Associates, 1994.
- [11] Rest, J., D. Narvaez, M.J. Bebeau, and S.J. Thoma, *Postconventional Moral Thinking: A Neo-Kohlbergian Approach*, Mahwah, NJ: L. Erlbaum Associates, 1999.
- [12] Self, D.J., and E.M. Ellison, "Teaching Engineering Ethics: Assessment of Its Influence on Moral Reasoning Skills", *Journal of Engineering Education* Vol. 87, No. 1, 1998, pp. 29-34.
- [13] Narvaez, D., and T. Bock, "Moral schemas and tacit judgement or how the Defining Issues Test is supported by cognitive science", *J. Moral Educ.* Vol. 31, No. 3, 2002, pp. 297-314.
- [14] Rest, J., and D. Narvaez, "DIT-2: Defining Issues Test", Minneapolis-St. Paul: University of Minnesota, 1998.
- [15] Rest, J., D. Narvaez, S.J. Thoma, and M.J. Bebeau, "DIT2: Devising and Testing a Revised Instrument of Moral Judgment", *Journal of Educational Psychology* Vol. 91, No. 4, 1999, pp. 644-659.

- [16] Rest, J., S.J. Thoma, D. Narvaes, and M.J. Bebeau, "Alchemy and Beyond: Indexing the Defining Issues Test", *Journal of Educational Psychology* Vol. 89, No. 3, 1997, pp. 498-507.
- [17] Harris, J., Charles E., M.S. Pritchard, and M.J. Rabins, *Engineering Ethics: Concepts and Cases*, 3rd ed., Belmont, CA: Wadsworth, 2005.
- [18] Rachels, J., *The Elements of Moral Philosophy*, 4th ed., New York: McGraw-Hill, 2002
- [19] Mappes, T.A., and J.S. Zembaty, *Social Ethics: Morality and Social Policy* 6th ed., New York: McGraw-Hill, 2002.
- [20] Schlaefli, A., J. Rest, and S.J. Thoma, "Does Moral Education Improve Moral Judgment? A Meta-analysis of Intervention Studies Using the Defining Issues Test", *Review of Educational Research* Vol. 55, No. 3, 1985, pp. 319-352.

Biographical Sketches

Jason Borenstein is the Director of Graduate Research Ethics Programs at Georgia Tech and the Editor of the Journal of Philosophy, Science & Law. While working for Georgia Tech, he has instructed courses on topics such as ethical theories, engineering ethics, biotechnology & ethics, and science & values in the policy process. He is also a member of the CITI-RCR Developers Group, a committee dedicated to creating online educational resources in the area of research ethics. Dr. Borenstein received his doctoral degree in philosophy from the University of Miami (FL) in May of 2001. Address: School of Public Policy, Georgia Institute of Technology, 685 Cherry Street, Atlanta, GA 30332-0345; email: borenstein@gatech.edu.

Matthew J. Drake is an Assistant Professor of Supply Chain Management in the A.J. Palumbo School of Business Administration at Duquesne University. He received a B.S.B.A. from Duquesne University and an M.S. and Ph.D. in Industrial Engineering from the Georgia Institute of Technology. His main research interests are in economic models for supply chain and revenue management and the ethics of supply chain collaboration. Address: A.J. Palumbo School of Business Administration, Duquesne University, 925 Rockwell Hall, Pittsburgh, PA 15282-0180; telephone: 412-396-1959; fax: 412-396-1797; email: drake987@duq.edu.

Robert Kirkman is an Assistant Professor of Philosophy in the School of Public Policy and Director of the Center for Ethics and Technology at Georgia Tech. He received his PhD in Philosophy from Stony Brook University. His primary field of research is environmental philosophy with a particular interest in ethical issues spurred by metropolitan growth. He teaches courses in engineering ethics, environmental ethics, and political theory. Address: School of Public Policy, Georgia Institute of Technology, 685 Cherry Street, Atlanta, GA 30332-0345; telephone: 404-385-4258; email: robert.kirkman@gatech.edu.

Julie L. Swann is an Assistant Professor in the Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology who received her Ph.D. in Industrial Engineering and Management Science from Northwestern University. Her research interests are in integrating methods from optimization and economics to solve problems in supply chain management and healthcare policy. Her teaching interests include these areas as well as improving ethics in engineering curricula. (Correspondence regarding the paper should be addressed to this author.) Address: Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0205; telephone: 404-385-3054; fax: 404-894-2301; email: jswann@isye.gatech.edu.