Do ethics classes teach ethics?

Howard J. Curzer
Texas Tech University, USA

Sabrina Sattler
Texas Tech University Health Sciences Center, USA

Devin G. DuPree
Green River Community College, USA

K. Rachelle Smith-Genthôs
Texas Tech University, USA

Abstract
The ethics assessment industry is currently dominated by the second version of the Defining Issues Test (DIT2). In this article, we describe an alternative assessment instrument called the Sphere-Specific Moral Reasoning and Theory Survey (SMARTS), which measures the respondent's level of moral development in several respects. We describe eight difficulties that an instrument must overcome in order to assess ethics classes successfully. We argue that the DIT2 fails to solve these problems, and that the SMARTS succeeds. The SMARTS was administered as pre-test and post-test during several semesters to ethics and non-ethics classes. Ethics students improved significantly more than non-ethics students in both moral theory choice and moral reasoning. Thus, ethics classes do indeed teach ethics.

Keywords
DIT, ethics assessment, moral development, moral reasoning, moral theory, SMARTS

All who are not maimed as regards their potentiality for virtue may win it by a certain kind of study and care. (Aristotle, NE 1099b18-20)

Corresponding author:
Howard J. Curzer, Department of Philosophy, Texas Tech University, Lubbock, TX, 79409-3092 USA.
Email: howard.curzer@ttu.edu

Driven by the demands of accreditation agencies as well as by the eagerness of many (though hardly all) faculty members to make use of outcome-based performance measures to improve their teaching, 'assessment' is fast becoming the most popular word in academia.[1] Administrators are desperately seeking useable, accurate, economical assessment instruments for every discipline. Meanwhile, ethics is emerging from its long hibernation and becoming sexy again. Awakened by heightened concern about student plagiarism, ethical lapses in various professions, research misconduct, and political sleaziness, ethics is turning away from the sterile, meta-ethical slumbers of the mid-twentieth century, and turning toward practical questions. Applied ethics centers are proliferating and spawning conferences; the National Institutes of Health (NIH), the National Science Foundation (NSF), and various professional societies are requiring ethics education; ethics questions are appearing on licensing exams; and so on. Together, the rise of assessment and the attempt to improve students' morals create a demand for instruments capable of assessing the contributions to moral development made by different sorts of ethics instruction.

Difficulties

Yet many of our colleagues smile indulgently when they discover that we are attempting to assess the impact of ethics classes on students' moral development. Why does this project seem quixotic to many academics? Some may be confused about the nature of our project. Thinking that our endpoint is behavior change, they are rightly skeptical of a survey's ability to measure changes in behavior. But our actual goal is more modest – to measure changes in moral thinking. Of course, there is the usual skepticism about assessing any class, and about the very existence of moral truths. But even academics who (a) accept that non-ethics classes are assessable and (b) accept that ethics is not illusory often (c) believe that ethics classes are not assessable. What are the *peculiar* difficulties of assessing ethics classes?

1. *Theory is not enough*. People using every moral theory[2] are found on both sides of every moral question. In many situations, a single moral theory is used to argue both that a certain option is morally required and that it is not required. Worse yet, some people with good moral theories reliably make bad decisions. Clearly, something in addition to theory is necessary to get to the right decision. Therefore, determining a person's moral theory is only part of the task of determining whether that person is morally well developed.

2. *Moral development is uneven*. People are not equally ethical across the board; they are better at dealing with some moral challenges than others. Despite philosophers' obsession with the unity of virtue, we all know benevolent yet intemperate people, courageous yet unjust people, and so on. This unevenness does not apply only to virtues and vices. Many people are far from virtuous with respect to some aspects of human life and even farther from virtuous with respect to other aspects. This poses a problem because measuring moral development in all aspects of a person's life is impractical. Assessment instruments cannot measure across a reasonably complete range of aspects of a person's life. Thus, assessment instruments will inevitably misevaluate many people.

3. *Experts disagree in theory*. Several moral theories are in current use, and ethicists squabble about how to rank these theories, and by extension how to rank their adherents. For example, claiming either that utilitarians are morally superior to virtue ethicists or the reverse would be controversial, to say the least. Yet without consensus about which theories are better and worse, how can progress be measured?

4. *Experts disagree in practice*. Similarly, ethicists often disagree about what is right in a particular situation. Sometimes this is because (3) ethicists endorse different moral theories, but other times it is because (1) ethicists come to different conclusions about what is right using the same theory. Yet other ethicists claim to arrive at decisions without applying theories at all. How are respondents' answers to be evaluated when there is no clear right answer?

5. *People are inconsistent*. Rather than using the same moral theory day in and day out for years until they make a comparatively quick, dramatic leap forward to a better moral theory, almost everyone regularly uses several different moral theories on different days or even at different times of the day for moral decision-making. (Not only do people use different theories in different spheres of human life, as indicated by (2) above, they also use different theories within the same sphere.) Insofar as any moral theory may be attributed to people, it seems that no one is at just one level of moral development. Thus, any measuring instrument that purports to determine each person's (single) level of moral development is misguided. Such an instrument has an inappropriate aim based upon an obviously false assumption.

6. *People have issues*. Not only do people use different theories, they also have views about various hot-buttons issues which do not reflect their overall moral theories. Prejudices, stances associated with authority figures, residues of wonderful or traumatic experiences, and other firmly held beliefs may persist alongside overall theories which are incompatible with them. Assessment instruments will be misled insofar as their questions hit on such emotionally charged anomalies among the moral beliefs of the respondents.

7. *People are ignorant*. Many respondents have a very limited range of experiences, especially about the issues debated at the national level, and they are loaded with patently false beliefs about the world. Assessment instruments which ask about situations involving these beliefs will often elicit poor choices, and will therefore down-rank the respondents' levels of moral development. But the respondents' choices are evidence of ignorance about the world rather than indications of retarded moral development. Thus, assessment instruments are likely to come to mistaken conclusions about respondents.

8. *The inner is not the outer*. As is well known, actual decision-making involves implicit moral theories, intuitions, and/or emotions which differ from the moral theories people use when explaining or justifying their actions (Cushman et al., 2010). For the purpose of measuring moral development, the implicit responses are the important ones when they determine action and differ from the explicit theories. Straightforwardly asking about reasons for action will elicit the explicit theories rather than the implicit responses. But how are implicit responses to be accessed and assessed?

Our stalking horse: The DIT2

The ethics assessment industry is currently dominated by the second version of the Defining Issues Test (DIT2) plus various instruments modeled upon it (Narvaez, 2005; Rest et al., 1999a, 1999b, 2000; Thoma, 2002; Walker, 2002). The goal of the DIT2 is to determine where to locate the respondent on the moral development path. The DIT2 consists of questions about scenarios similar to the well-known Heinz dilemma. Respondents are presented with several questions related to the scenario, and asked to rate their importance on a Likert-type scale.

Like any widely used survey, the DIT2 has its critics (Nucci, 2002; Puka, 2002). It has been accused of bias against conservatives, for example (Emler et al., 1983). Our own concerns about the DIT2 differ from the criticisms currently on the table. They arise from the eight difficulties of assessing ethics classes listed above.

1. *Theory is not enough*. The psychological theory underlying the DIT2 is called the Four-Component Model (FCM) because it takes decision-making to be a product of four components: moral judgment, sensation, motivation, and character. While these components are obviously interrelated, the FCM separates them because they do not typically develop at equal rates within a single individual. Sometimes a person perceives well, but judges badly, for example. The DIT2 focuses on the judgment component. It purports to measure a person's level of judgment. We are skeptical. Moral judgment is (or includes) the application of moral theories to perceived concrete situations using a reasoning process. Admittedly, these items are deeply interrelated, yet they can vary independently. The FCM acknowledges the independence of judgment and sensation; we think it should also acknowledge the independence of theory choice and reasoning. Some people do not correctly complete the practical syllogism because they have implausible premises; they correctly apply sophomoric moral theories. Other people fail because they reason badly from plausible premises; they apply sophisticated moral theories in inept ways. Like Kohlberg who criticized the DIT for lumping together two items which vary independently (content and structure), we criticize the DIT2 for lumping together theory choice and reasoning. Either the DIT2 measures a jumble of two independent constructs, or it ignores moral reasoning, and only actually measures the theory choices of respondents (Bebeau, 2002; Puka, 2002).[3]

2. *Moral development is uneven*. People develop at different rates in different spheres of human life. But although the DIT2 consists of multiple questions concerning several scenarios, it generates only one rating per respondent. That is, it locates the respondent at a single level of moral development, across all spheres. It attempts to measure only overall moral development. Thus, the DIT2 does not take account of the common phenomenon that a person's ability to think morally develops at different rates within different spheres of human life. In this respect, too, the DIT2 is insufficiently fine-grained.

3. *Experts disagree in theory*. The DIT2 takes stands on which theories are superior to which other theories. Very roughly speaking, one might make the following comparison.

Moral theories	Neo-Kohlbergian schemas
Divine command	
Normative egoism	Personal interest (transitional)
Care ethics	Personal interest (consolidated)
Normative Relativism	Maintaining norms (transitional)
Normative Relativism	Maintaining norms (consolidated)
Social contract	Post-conventional (transitional)
Deontology	Post-conventional (consolidated)

But once this comparison is made, it becomes clear that some of the DIT2's rankings are highly controversial among ethicists. For example, there is no consensus on the claim that normative egoism is worse than normative relativism. Of course, this is a mark against the DIT2; surveys should not hold respondents to a standard which a substantial percentage of experts reject.

4. *Experts disagree in practice*. Because the DIT2 ranks theories, it takes a stand by extension on what acts are right in various situations. The right answers are the ones endorsed by the best theory. But again many of these particular claims – claims about what agents should do in particular situations – are highly controversial among ethicists. Just as there is no consensus on what Heinz should do, so there is nothing close to consensus on what should be done in any of the DIT2's scenarios. This lack of consensus is another mark against the DIT2.

5. *People are inconsistent*. The FCM which underlies the DIT2 arranges moral theories in a hierarchy of increasing sophistication. According to the FCM, people do not use one theory for years and then discard it in favor of another, like ascending a ladder. Rather, people use several different theories at each point on the moral development path. Progress consists in using more and more sophisticated theories more and more often. The DIT2 accepts the view that almost everyone deploys a hodge-podge of incompatible moral theories for implicit decision-making. But because the FCM ranks these theories one above the next, it seems to imply that as individuals move from decision to decision they bounce back and forth among several different levels of moral development. According to the FCM, morally advanced people use better theories more often than morally underdeveloped people, but all use all sorts of theories to some extent. However, someone who moves rapidly from theory to theory is better described as confused or lacking commitment to any moral theory at all rather than someone at a particular point on a moral development path. This suggests that there is a problem with the FCM's picture of people using different moral theories in quick

succession. Moreover, as we all know from personal experience and from talking with other people, moral theories are very firmly held (though inconsistently applied, as mentioned below). Like other fundamental beliefs, a person's underlying moral theory changes very rarely, and with great difficulty and pain. People typically hold fast to their fundamental normative beliefs despite evidence and arguments, stories and tears, and so on. But this observation is also difficult to reconcile with the FCM's picture of people using different moral theories in quick succession.

6. *People have issues*. The DIT2 uses only five scenarios (the DIT used only six scenarios, the DIT2's short version uses only three), and these scenarios concern issues about which people have strong feelings and firm beliefs. These are hot-button issues for many people. Thus, there is a high likelihood that many respondents will give outlier answers to a significant part of the DIT2. That is, the DIT2's attempt to measure moral development is at least partially confounded by unrepresentative, idiosyncratic responses to isolated items.

7. *People are ignorant*. The DIT2 focuses on public policy issues. But except for a small coterie of VIPs, the moral choices actually confronting people concern the personal issues of everyday life. Most of us deal with micro-ethics rather than macro-ethics. Moreover, many people are uninformed or misinformed about questions of public policy. Because they lack an accurate grasp of the relevant facts, the positions people take on public policy may often be influenced by ideological and political factors that are unrelated to the moral theories and reasoning they use to make daily decisions in their own lives. Finally, the current polarization of politics and ideological cast of public discourse tend to decouple people's stands on public policy from the moral theories and reasoning they use to make decisions in their own lives. So by focusing on public policy issues, the DIT2 is not asking the right sort of questions.

8. *The inner is not the outer*. Everyone agrees that moral decision-making is at least partially a function of implicit responses. Decisions do not simply issue from theories explicitly endorsed by agents. The DIT2 addresses the problem of accessing and assessing these implicit responses by making brief, cryptic references to various theories in its multiple-choice answers. The rationale behind this 'fragment strategy' is that respondents who are already thinking about a scenario along the lines of a certain theory will recognize the reference, but other respondents will just pass over that answer (Narvaez and Bock, 2002). We believe that the fragment strategy fails because cryptic references typically evoke all sorts of lines of thought – meaningful and otherwise. Rather than accessing the actual, implicit responses of the respondents, we believe that the fragment strategy of the DIT2 is just generating noise.

Our alternative instrument: SMARTS

We are in the process of developing an instrument to measure moral development which will avoid the eight difficulties of assessing ethics classes. Our instrument, the Sphere-Specific Moral Reasoning and Theory Survey (SMARTS), improves upon the DIT2 with

respect to all eight obstacles. We do not claim to have solved all of the problems; we only claim to do better than the DIT2.[4]

1. *Theory is not enough*. SMARTS consists of 20, two-part questions about various scenarios. The first part of each question asks whether some act is right or wrong. If the respondent answers 'right', then he or she is offered seven explanations for *why* it is right. Three are applications of sophisticated moral theories; three are applications of sophomoric moral theories; one is an appeal to authority. If the respondent answers 'wrong', then he or she is also offered seven explanations for *why* it is wrong based on these theories. Thus, each moral theory is deployed in two ways in each question. One answer justifies a choice via plausible utilitarian reasoning; another justifies the opposite choice via dubious utilitarian reasoning. Similarly for the other moral theories. For example,

Sam and Betty are unmarried college students. They do not love each other; they are just friends. But they do enjoy having sex with each other, occasionally. Each of them understands that the other is making no commitment; nothing more than physical pleasure is involved. Neither is currently in another relationship, and they are always careful to use birth control. Are they doing something morally wrong?

    a. Yes
    b. No

If your answer is 'Yes', *why* do you think that they are doing something morally wrong?

1. Because God prohibits adultery. (A) (N)
2. Because they would each get more pleasure from sex with someone they love. (E) (B)
3. Because sex should be enjoyed at the right time, in the right way, with the right person, just like other sensual pleasures. Sex with someone you don't love is not 'with the right person'. (V) (G)
4. Because sex ought to be for deepening a relationship rather than just superficial pleasure. (C) (G)
5. Because if everyone acted like this, then the institution of marriage would disappear, and people would lose the option to marry. (D) (B)
6. Because our society frowns on sex outside of marriage. (R) (G)
7. Because even if Sam and Betty avoid problems, other couples who follow their lead will end up hurting each other or third parties. Sam and Betty are increasing grief by setting a bad example. (U) (B)

If your answer is 'No', *why* do you think that they are *not* doing anything morally wrong?

1. Because Sam and Betty are enjoying themselves, and there are no drawbacks for either of them. (E) (G)
2. Because this sort of relationship is widely condoned in our society, as our books, movies, TV, and songs illustrate. (R) (B)

3. Because sex should be enjoyed in moderation, just like other sensual pleasures. Having sex with each other enables Sam and Betty to have a moderate amount of sex. (V) (B)

4. Because Sam and Betty are following their natural instincts, and what they are doing is not against the law. (A) (N)

5. Because both Sam and Betty are adults making free, rational choices and are neither deceiving nor coercing anyone. (D) (G)

6. Because friends should share activities and secrets, and sex is the ultimate bonding experience. (C) (B)

7. Because Sam and Betty are adding a bit of pleasure to the world by pleasing each other. (U) (G)

   (G)=good, (B)=bad, (N)=no theory, (A)=appeal to authority, (E)=egoism, (R)=relativism, (C)=care ethics, (U)=utilitarianism, (D)=deontology, (V)=virtue ethics.

Thus, SMARTS measures moral theory choice (which we shall call 'theorizing') and moral reasoning independently in the following way. When respondents choose among the proffered explanations, they display tendencies to choose either sophisticated or sophomoric theories. They also display tendencies to choose either right or wrong applications of whatever moral theories they choose. Respondents who use sophisticated moral theories but nevertheless choose what their theories would take to be the wrong answers have a deficiency of reasoning, but not of theorizing. Conversely, respondents who use sophomoric moral theories but nevertheless choose what their theories would take to be the right answers have a deficiency of theorizing, but not of reasoning. Thus, SMARTS treats theorizing and reasoning as two different constructs. We do not measure every aspect of moral decision-making, but by allowing for the possibility that people can apply sophomoric theories well or sophisticated theories badly, SMARTS measures more than the DIT2.

2. *Moral development is uneven*. Following Aristotle, we take human life to consist in collections of overlapping, yet different spheres, each concerned with a different package of goods, passions, and actions. Courage consists in thinking, feeling, and acting well in situations involving physical risk; temperance consists in thinking, feeling, and acting well in situations involving sensual pleasure; and so on (Nussbaum, 1988). Some people are good at reasoning about sensual pleasure, but bad at reasoning about physical risk, others the reverse. Each question of SMARTS concerns a scenario which falls into one of the following three spheres of human life: (a) sensual pleasure, (b) deception and cheating, (c) helping others. From a virtue perspective, these are the spheres governed by temperance, justice, and benevolence. Thus, SMARTS tests the moral theorizing of respondents, independently in these three spheres, and similarly for the respondents' moral reasoning. Thus, SMARTS allows for the possibility that a respondent is a morally sophisticated theorizer with respect to temperance, but a naïve theorizer with respect to justice, for example. Again, although we do not measure every aspect of human life, we do separate out several important aspects. Unlike the DIT2, we

do not attempt to measure all aspects together, but instead allow for uneven moral development with respect to different spheres of human life.

3. *Experts disagree in theory*. SMARTS takes utilitarianism, deontology, and virtue ethics to be sophisticated moral theories; egoism, relativism, and the care ethics to be sophomoric moral theories; and the appeal to authority as something other than a moral theory. Unlike the DIT2 ranking, the SMARTS ranking is uncontroversial among ethicists.[5] Almost every introductory ethics textbook published in the last few decades ranks these theories as we do, with the following caveats. First, care ethics is not directly disparaged, but partiality is disparaged, and care ethics is a sort of partiality. Second, the appeal to the authority of God is typically listed as a moral theory – Divine Command theory. However, Divine Command theory instructs its adherents to follow God's orders rather than to use a moral theory as a guide to making one's own decisions.

4. *Experts disagree in practice*. SMARTS offers two answers for each theory in each scenario: one of these is a right answer and the other is a wrong answer *according to that theory*. However, SMARTS takes only minimal stands on what choices are right, *simpliciter*, in different situations. If the three sophisticated theories disagree about what the agent should do, SMARTS does not specify which choice is correct. SMARTS endorses (or discards) a choice only when all three sophisticated theories agree in accepting (or rejecting) that choice. Thus, SMARTS ranks respondents' theorizing and reasoning without taking a stand on what is right in any case where reasonable people disagree. It takes a stand for (or against) a choice only when all sophisticated people would be for (or against) that choice. In the current version of SMARTS, all three sophisticated theories agree in only 2 of the 20 scenarios.

5. *People are inconsistent*. What is needed is a model which (a) captures the insight that each person uses several different theories over short periods of time (b) without denying that people hold tenaciously to their theories when challenged. We maintain that people *appear* to deploy several theories for decision-making for two reasons. People use different moral theories in different spheres of human life, as mentioned above. But an individual often seems to use different moral theories within the same sphere. How is this fact to be explained? People in our society are familiar with many different moral theories.[6] Theoretically, these theories cannot be combined because they have incompatible assumptions, methodologies, and so on. Yet refined versions of each theory use complex epicycles to incorporate the insights of the others. Virtue ethicists might use utilitarian rules as the bases for constructing lists of virtues, for example. In light of these epicycles, a reasonable approximation would be to use each of the three sophisticated moral theories in different sorts of situations. One might describe this as a single, multifaceted, technically self-contradictory, yet sophisticated moral theory, namely utilitarianism–deontology–virtue–ethics. Thus, reasonably sophisticated respondents may seem to flit from one theory to another while actually holding a single, compound, internally inconsistent theory throughout.

6. *People have issues*. While the DIT2 uses only five scenarios, SMARTS uses 20 scenarios, each of which concerns a somewhat different issue. Admittedly, this

makes hitting the respondents' buttons more likely. On the bright side, however, using so many scenarios dilutes the effect of any button-hits to the point of negligibility. The larger number of scenarios in SMARTS enhances the likelihood that responses will reliably reveal a pattern of moral thinking, rather than merely transient features unique to hot-button topics.

7. *People are ignorant*. SMARTS uses scenarios about everyday life. By sticking to micro-ethical scenarios about which the respondent is likely to have some personal (or at least some vicarious) experience, SMARTS avoids inappropriately down-ranking people because of their ignorance about the world. It avoids confounding ethics measurement with knowledge of macro-ethical issues of public policy.

8. *The inner is not the outer*. We eschew the DIT2's fragmentation strategy. The explanations for choices within scenarios presented to respondents by SMARTS briefly, but completely express moral justifications for choices. Instead of the fragmentation strategy, SMARTS brings the implicit responses of the respondents to consciousness by separating the response to each scenario into two stages. First, SMARTS asks respondents what should be done in a scenario. Actually using the implicit theory to make a choice brings the implicit response closer to consciousness. Only after the respondents answer the toss-up question does SMARTS offer alternative rationales for the answer. Hopefully, one of these rationales will dovetail with the respondent's fully-activated, implicit response.

We began by listing eight reasons for skepticism about assessments of ethics classes. We have argued that the DIT2 fails to address these worries adequately, and that our competing instrument, SMARTS, does a better, though not perfect job.

Administration of the survey

SMARTS has been under development off and on for 5 years. Questions which yielded idiosyncratic answers were replaced in successive incarnations of SMARTS. Ethicists were informally consulted at several points to ensure that the 280 explanations of why choices in the scenarios would be right or wrong accurately represented the seven moral theories SMARTS utilizes. A panel of philosophy graduate students sorted the 20 questions into the three spheres of human life, and the 280 explanations into the seven moral theories to confirm that the questions and explanations were correctly categorized. Focus groups of graduate students and undergraduates also helped to refine the instrument. Wordsmiths reviewed the entire test in order to eliminate misleading or button-pushing phrases. Versions of SMARTS have been administered to college classes near the beginning and end of nine semesters. Incomplete, incorrectly completed, and unmatched surveys were excluded. We have used only matched-pairs of correctly completed pre- and post-tests.

For each question, the participant was given a moral theory score based on the level of theory represented by their chosen response (sophisticated theory = 2, sophomoric theory = 1, no theory = 0). The participant was also given a moral reasoning score based on the use of theory as represented by their chosen response (good use of theory = 2, bad use of theory = 1, no use of theory = 0). Total moral theory and reasoning scores were computed by summing the moral theory and reasoning scores from each question.

Table 1. Scale development.

Version	Semester	Matched-pairs	Schools	Class types	Alpha theory		Alpha reasoning	
1	S09	34	TTU	Ethics				
2	F09	59	TTU	Ethics				
3	S10, F10, S11	210	TTU	Ethics, Logic, Beginning	.33		.26	
4	F11, S12	305	TTU, Drury Husson	Ethics, 11 non-ethics courses				
5	F12	100	TTU	Ethics				
6	S13	225	TTU, Drury Husson, Stout	Ethics, 9 non-ethics courses	.66	.64	.65	.61

TTU: Texas Tech University.

As shown in Table 1, Versions 1 and 2 of SMARTS were administered to 'Introduction to Ethics' classes. Version 3 of SMARTS was administered not only to several 'Introduction to Ethics' classes, but also to 'Beginning Philosophy' classes and 'Logic' classes. All three sorts of classes enroll similar student populations and are taught by philosophers, but 'Beginning Philosophy' classes present non-ethical theories and 'Logic' classes teach non-moral reasoning. Thus, 'Beginning Philosophy' students and 'Logic' students make suitably matched control groups for the 'Introduction to Ethics' classes. Version 4 of SMARTS was administered in a multi-center, multi-class-type trial. Version 5 of SMARTS was an expanded version, primarily aimed at improving the reliability of the instrument. In the present article, we present data obtained via Version 6 of SMARTS which was administered at the beginning and end of the Spring 2013 semester at four universities to a broad range of classes.

Reliability of SMARTS

Reliability is a measure of consistency or repeatability of a survey. An important form of reliability is *internal reliability* (whether the questions are all measuring the same trait) as measured by Cronbach's alpha (Babbie, 2011; Carmines and Zeller, 1991; Dane, 2011).

As shown in Table 2, the six subscales of SMARTS all have relatively low levels of internal reliability, but the overall scales of theorizing and reasoning are close to the generally acceptable level of 0.70. Overall reliability has increased substantially since Version 3 (S10–S11) where theory and reasoning scores were .33 and .26, respectively. The results discussed below suggest that although SMARTS overall scales and subscales are not yet as reliable as we would like, SMARTS is able to show the effect of interventions such as ethics instruction.

We hypothesized that people's moral theorizing and reasoning develop independently of each other, and independently within the spheres of human life. Thus, we expected

Table 2. Internal reliability measured by Cronbach's alpha.

	Theory		Reasoning	
Measure	Pre	Post	Pre	Post
Overall scales	.66	.64	.65	.61
Benevolence	.42	.43	.18	.33
Justice	.52	.52	.58	.48
Temperance	.40	.44	.55	.23

Table 3. Correlations between difference scores (post–pre) for ethics courses.

	Total theory	Total reasoning	Temperance theory	Temperance reasoning	Justice theory	Justice reasoning	Benevolence Theory
Total theory	–						
Total reasoning	.66*	–					
Temperance theory	.65*	.42*	–				
Temperance reasoning	.52*	.68*	.64*	–			
Justice theory	.56*	.30*	.08	.07	–		
Justice reasoning	.42*	.63*	.17*	.17*	.45*	–	
Benevolence theory	.63*	.50*	.105	.23*	.12	.13	–
Benevolence reasoning	.32*	.61*	-.001	.14*	.05	.043	.60*

*p < .05.

that ethics classes would produce different results in respondents' theorizing and reasoning within each of the three spheres. We measured these differences.

As shown in Table 3, all subscale difference scores are significantly correlated with overall theory and reasoning difference scores. However, while difference scores for some subscales exhibit a robust relationship, others are not related to one another, which could suggest that improvement in the subscales, while related to the overall scale, is not necessarily related to improvement in other subscales. These relationships between the overall scales and subscales provide preliminary support for the hypothesis of independent moral development.

Results of SMARTS

A central question of this article is whether the intervention of an ethics class improves thinking about ethics. We tested this through a comparison of the pre- and post-test results of matched-pairs of surveys with respect to theorizing and reasoning in each of

Table 4. Changes in scores.

	Course mean differences (*N* of matched tests)		
	Non-ethics (44)	Ethics course (183)	Significance
Theory	−.25	.94	.044*
Reasoning	−1.09	.38	.015*
Temperance theory	−.14	.69	.013*
Temperance reasoning	−.36	.13	.085
Justice theory	−.18	.5	.039*
Justice reasoning	−.61	.28	.004*
Benevolence theory	.16	−.04	
Benevolence reasoning	−.11	−.04	

*$*p < .05$.

three spheres. Do the mean scores on the subscales increase significantly more in ethics classes than in non-ethics classes?

As shown in Table 4, independent samples *t*-tests revealed that participants' overall theorizing and reasoning scores improved during the semester in ethics classes. The overall difference score measures, and the measures in the justice and temperance spheres (though not the benevolence sphere) were marginally to significantly higher in ethics classes when compared to non-ethics courses.

Although the reliability of the scales and subscales of SMARTS could improve, SMARTS is able to show the effect of interventions. These statistically significant changes in theory and reasoning scores are unlikely to be due to chance. SMARTS is measuring something. The fact that students in ethics classes improve in moral theorizing and reasoning more on average than students in the control group indicates that this improvement is produced by instruction in ethics. To answer the question of this article's title, *ethics classes do teach ethics.*

Discussion

We initially described eight problems peculiar to the assessment of ethics classes, and argued that the DIT2 fails to overcome these obstacles. When introducing the SMARTS, we emphasized its ability to succeed where the DIT2 fails. The results from our Spring 2013 administration of the SMARTS demonstrate the SMARTS' ability to reveal ethical development when an intervention – in this case an ethics course – has been implemented. Some ethics instructors maintain that ethics courses improve the moral thinking of students. By requiring ethics instruction, accreditation agencies and professional organizations implicitly concur. Our study provides modest backing for their claims. Overall,

students in ethics classes experienced mild gains in moral reasoning and somewhat greater gains in theorizing. Similar gains were not seen in the control group. Indeed, students in the control group did not improve in moral theorizing or reasoning at all. We have substantiated the claim that ethics classes mildly improve the moral thinking of students. This is no surprise; studies using the DIT2 have found similar results. Since we used a different instrument and measured different aspects of ethics, our study adds an increment of independent confirmation to the existing literature, and some further specificity.

Caveats: Although we have shown that moral thinking can be taught, we have certainly not settled the longstanding question of whether virtue can be taught (Plato, 1980), for we have not shown that ethics classes improve moral passions, perceptions, or behavior. We have also shown nothing about the effectiveness of web-based ethics instruction or ethics modules imbedded within non-ethics courses. As usual, further research is needed.

The fact that these improvements are small should not be surprising or disheartening. Since almost half of the students started the class at the highest level of theorizing and more than a third started at the highest level of moral reasoning, possible improvement was limited. Moreover, by the time students get to college, they have been exposed to moral education (in many cases 'indoctrination' might be a better term) from parents, peers, churches, teachers, Hollywood, and other sources for at least 18 years. The fact that ethics instructors can make any positive impact at all in a one-semester course is an impressive and promising achievement.

	Problems	DIT2 inadequacies	SMARTS solutions
1	*Theory is not enough.*		
	Since something in addition to theory is necessary to get to a decision, determining a person's moral theory does not settle the issue of whether that person is morally well developed.	Judgment is the application of moral theory to perceived concrete situations using a reasoning process. The DIT2 does not separately measure the two independent constructs: moral theory choice and moral reasoning.	SMARTS measures moral theory choice and moral reasoning independently.
2	*Moral development is uneven.*		
	People are not equally ethical across the board, Since assessment instruments can be sensitive only to some aspects of each person's psychology, they will inevitably misevaluate people.	The DIT2 does not take account of the common phenomenon that a person's ability to think morally develops at different rates within different spheres of human life.	SMARTS independently tests the moral theorizing and reasoning of respondents in three different spheres.
3	*Experts disagree in theory.*		
	Different moral theories are endorsed by different ethicists. Yet without consensus about which theories are better and worse, how can progress be measured?	Some of the DIT2's rankings are highly controversial among ethicists. Surveys should not hold respondents to a standard which many experts reject.	SMARTS ranks moral theories in a coarse-grained way which is uncontroversial among ethicists.

	Problems	DIT2 inadequacies	SMARTS solutions
4	*Experts disagree in practice.*		
	There is much disagreement about what is right in a particular situation. How are respondents' answers to be evaluated when there is no clear right answer?	By ranking theories, the DIT2 takes stands on what acts are right in various situations. The right acts are the ones endorsed by the best theory. But many of these claims about what agents should do are highly controversial among ethicists.	SMARTS takes no stand on what choices are right in cases where reasonable people disagree.
5	*People are inconsistent.*		
	Almost everyone regularly uses several different moral theories. No one is at just one level of moral development. Any measuring instrument that purports to determine a person's (single) level of moral development is misguided.	According to the FCM, morally advanced people use better theories more often than morally underdeveloped people, but all people use all sorts of theories to some extent. But someone who flits from theory to theory is better described as confused or lacking commitment to any moral theory at all.	SMARTS tests the moral theorizing of respondents in three different spheres. SMARTS also takes several moral theories to be equally sophisticated, so moving among them does not indicate a change of moral development level.
6	*People have issues.*		
	People have hot-buttons beliefs which do not reflect their moral theories. Assessment instruments will be misled insofar as their questions hit on such emotionally charged anomalies.	The DIT2 uses only five scenarios and these concern hot-button issues. Many respondents will give outlier answers to a significant part of the DIT2.	SMARTS dilutes the effect of hot-buttons issues by using twenty scenarios.
7	*People are ignorant.*		
	Many people have very limited experience about macro-ethical issues, and patently false beliefs about the world. Assessment instruments which ask about situations will therefore down-rank the respondent's level of moral development inappropriately.	The DIT2 focuses on macro-ethical issues. But people often lack an accurate grasp of the facts relevant to these issues. The moral choices actually confronting most people concern the personal issues of everyday life.	SMARTS asks about micro-ethical issues.
8	*The inner is not the outer.*		
	People typically use an implicit moral theory for decision-making which differs from the one they use when explaining or justifying their actions. But how are implicit theories to be accessed and assessed?	The DIT2 uses the 'fragment strategy' to access implicit theories. The rationale is that respondents who are already using a certain theory will recognize a cryptic reference, but other respondents will just pass over that answer. However, cryptic references typically evoke all sorts of lines of thought – meaningful and otherwise.	SMARTS accesses implicit theories by asking respondents what should be done in a situation before offering alternative rationales.

Acknowledgements

Funding

Notes

1. Our guess is that it is outstripping title holders of previous years such as 'sustainability', 'interdisciplinary', 'narrative', 'paradigm', 'hermeneutics', and 'semiotics'.
2. By 'moral theory', we shall mean a set of general principles (or other standards of conduct) that specify which actions are morally right (wrong) and explain why they are right (wrong). Since the focus of this article is the assessment of ethics classes, the theories we have in mind are those typically discussed in ethics classes (e.g. appeal to authority, egoism, relativism, care ethics, utilitarianism, deontology, virtue ethics).
3. The Four-Component Model (FCM) uses the term 'schema' at crucial points; it does not use the term 'moral theory'. Although schemas and moral theories are not quite the same, we will background the differences and use only the term 'moral theory' for the sake of consistency and cross-disciplinary communication.
4. We have not yet compared the results directly because the Sphere-Specific Moral Reasoning and Theory Survey (SMARTS) is still under development, but we plan to run the second version of the Defining Issues Test (DIT2) and the next version of SMARTS alongside each other in Fall 2014.
5. Is our classification of care ethics as a sophomoric moral theory unfeminist? Some feminists would certainly contest our classification, but many would applaud (Card, 1990, 1996; Davion, 1993; Spelman, 1991).
6. We suspect that the reason for the widespread familiarity with so many theories is that as people mature and move about within our pluralistic society, they are exposed to, and internalize many different moral theories from many sources.

References

Babbie E (2011) *The Basics of Social Research*, 5th edn. Belmont, CA: Wadsworth.

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

Card C (1990) Caring and evil. *Hypatia* 5: 101–108.

Card C (1996) *The Unnatural Lottery: Character and Moral Luck*. Philadelphia, PA: Temple University Press.

Carmines EG and Zeller RA (1991) *Reliability and Validity Assessment*. Newbury Park, CA: SAGE.

Cushman F, Young L and Greene JD (2010) Our multi-system moral psychology: Towards a consensus view. In: J Doris, G Harman, S Nichols, et al. (eds) *The Oxford Handbook of Moral Psychology*, pp. 47–71. Oxford: Oxford University Press.

Dane F (2011) *Evaluating Research: Methodology for People Who Need to Read Research*. Newbury Park, CA: SAGE.

Davion V (1993) Autonomy, integrity, and care. *Social Theory and Practice* 19: 161–182.

Emler N, Renwik S and Malone B (1983) The relationship between moral reasoning and political orientation. *Journal of Personality and Social Psychology* 45: 1073–1080.

Narvaez D (2005) The Neo-Kohlbergian tradition and beyond: Schemas, expertise and character. In: C Pope-Edwards and G Carlo (eds) *Nebraska Symposium on Motivation*, vol. 51, pp. 119–164. Lincoln, NE: University of Nebraska Press.

Narvaez D and Bock T (2002) Moral schemas and tacit judgment or how the Defining Issues Test is supported by cognitive science. *Journal of Moral Education* 31: 297–314.

Nucci L (2002) Goethe's Faust revisited: Lessons from DIT research. *Journal of Moral Education* 31: 315–324.

Nussbaum M (1988) Non-relative virtues. *Midwest Studies in Philosophy* 13: 32–53.

Plato (1980) *Meno* (trans. GMA Grube). Indianapolis, IN: Hackett Publishing.

Puka B (2002) The DIT and the 'dark side' of development. *Journal of Moral Education* 31: 339–352.

Rest J, Narvaez D, Bebeau M, et al. (1999a) A Neo-Kohlbergian approach: The DIT and schema theory. *Educational Psychology Review* 11: 291–324.

Rest J, Narvaez D, Bebeau M, et al. (1999b) *Postconventional Moral Thinking: A Neo-Kohlbergian Approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Rest J, Narvaez D, Bebeau M, et al. (2000) A Neo-Kohlbergian approach to morality research. *Journal of Moral Education* 29: 381–396.

Spelman E (1991) The virtue of feeling and the feeling of virtue. In: C Card (ed.) *Feminist Ethics*, pp. 213–232. Lawrence, KS: University of Kansas Press.

Thoma S (2002) An overview of the Minnesota approach to research in moral development. *Journal of Moral Education* 31: 225–245.

Walker L (2002) The model and the measure: An appraisal of the Minnesota approach to moral development. *Journal of Moral Education* 31: 353–367.

Author biographies

Howard J. Curzer is a Professor of Philosophy at Texas Tech University. His publications include a monograph entitled *Aristotle and the Virtues* (OUP, 2012), a teaching anthology entitled *Ethical Theory and Moral Problems* (Wadsworth Press, 1999), and various articles on ancient philosophy, contemporary virtue ethics, and applied ethics. His work has been supported by grants from the NSF and NEH.

Sabrina Sattler is the Senior Director for Institutional Assessment in the Office of Institutional Planning and Assessment at Texas Tech University Health Sciences Center. Her work involves survey development and administration, leading focus groups, and quantitative as well as qualitative data analysis.

Devin G. DuPree is the Institutional Assessment Coordinator at Green River Community College in Auburn, WA, and a doctoral candidate in the Marriage and Family Therapy program at Texas Tech University. His research relates to the cross-disciplinary study of interpersonal behavior and education, including topics such as couple/relationship education and education related to morals, ethics, and other interpersonal behavior.

K. Rachelle Smith-Genthôs is a doctoral candidate in Experimental-Social Psychology at Texas Tech University. Her research and work focus on automatic processes in interpersonal interaction such as non-conscious behavioral mimicry, facial expressions, and the intrapersonal and interpersonal effects of rapport.