Assessing Medical Students' Moral Judgment over the Course of a Four-Year Professionalism and Humanism Curriculum

Sandra Riegle, Ph.D.
Ann W. Frye, Ph.D.
Jason Glenn, Ph.D.
Kirk L. Smith, M.D., Ph.D.

Dr. Riegle is assistant professor, Department of Education, Morehead College, Morehead, KY.

Dr. Frye is the Director of Educational Development in the Office of Educational Development, University of Texas Medical Branch, Galveston, TX.

Dr. Glenn is assistant professor, Institute for the Medical Humanities, University of Texas Medical Branch, Galveston, TX.

Dr. Smith is associate professor, Department of Internal Medicine, University of Texas Medical Branch, Galveston, TX.

Correspondence should be addressed to Dr. Kirk L. Smith, Department of Internal Medicine, The University of Texas Medical Branch, 2.500 Rebecca Sealy Hospital., Galveston, TX 77555-0168.
Email: klsmith@utmb.edu
Phone: (409)226-9763
Fax: (409)226-9776

Abstract

Teachers tasked with developing moral character in future physicians face an array of pedagogic challenges, among them identifying tools to measure progress in instilling the requisite skill set. One validated instrument for assessing moral judgment is the Defining Issues Test (DIT-2). Based on the work of Lawrence Kohlberg, the test's main index, the P score, indicates the respondent's progress within a framework of increasing moral sophistication. To evaluate the effectiveness of the Professionalism and Humanism curriculum at the University of Texas Medical Branch (UTMB) in improving medical students' professional judgment and moral reasoning, the authors administered the DIT-2 to the medical school class of 2011 (N = 236) during orientation at the beginning of school (T1) and again shortly before graduation (T2). 195 tests were subsequently scored by the Center for the Study of Ethical Development at T1 and 72 tests at T2. Analysis of variance was used to assess for main effects of gender and primary language on P scores. The P score mean for the 1st year medical students was 43.67 on a 0 to 95 scale; for graduating seniors, it was 43.31. At T1 and T2, a statistically significant main effect for gender was observed on P scores ($p < .01$) but not for English as primary language. There was no statistically significant difference between the results at T1 and T2, which contrasts with the normal progression of moral reasoning that is typically observed in young adults who seek higher education. However, the UTMB results are consistent with other studies of medical students which have demonstrated a similar inhibition of moral reasoning over four years of medical education.

Key words: Assessment; curriculum; ethics; medical school; professionalism

Introduction

In 2007, the University of Texas Medical Branch (UTMB) began implementing a four-year Professionalism and Humanism (P-H) curriculum for medical students. The P-H curriculum is based on the Accreditation Council for Graduate Medical Education (ACGME) professionalism competency and specifies a skill set comprised of altruism, sound ethical practice, and sensitivity to culture, age, gender and disability (Association of Graduate Medical Education [ACGME] Outcome Project, 2005). The P-H curriculum was conceived as a broad-based and integrative initiative, with every course and clerkship taking up relevant aspects of professionalism. The devastation wrought on UTMB by Hurricane Ike in September 2008 resulted in paring of the scope of the initiative; however, significant elements of the P-H curriculum were introduced into approximately one half of School of Medicine courses and clerkships and a robust extra-curricular program of P-H-oriented small group discussions and community service learning activities was launched and sustained. (Smith et al, 2007, Muller et al, 2010) This paper reports a longitudinal effort to evaluate the P-H curriculum.

Tools for assessing moral development are few in number. Those cited in the medical education literature include Kohlberg's Moral Judgment Test, Hogan's Maturity of Moral Judgment Scale, Gilligan's Sexual Moral Judgment Scale, and Gibbs' Social Reflection Questionnaire (Character Education Partnership [CEP], 2005). While these tools are considered some of the most reliable instruments for measuring moral reasoning and values, the one most often cited in terms of frequency of administration, cumulative results, and suitability to medical education is the Defining Issues Test (DIT) (Bebeau & Thoma, 2003).

The DIT-2 (the most recent version of the test) was created by James Rest, based on Lawrence Kohlberg's typology of moral development (Rest, Narvaez, Bebeau, & Thoma, 1999). It consists of five scenarios, each depicting an ethical dilemma. For each scenario, respondents are provided a series of questions reflecting strategies for resolving the dilemma that they rank in order of sophistication. The rank orderings are tabulated to compute an index of the respondent's moral judgment, typically reported as a P score. P scores may range from 0 to 95. As Elm and Weber (1994) note, Rest's scenarios are rooted in a concept of justice that tests the extent to which individuals are capable of participating in goals external to themselves, or, as Latif (2001) explains, "how one best organizes social cooperation in society by coordinating activities in such a manner so as to maximize human welfare" (p. 120). Such a determination is pertinent to the practice of medicine insofar as the profession's characteristic moral stand, the primacy of patient interests, is usually cast in terms of a contrast between self-interest and duty to others. The DIT's P score indicates the student's readiness to adopt that stand by assessing whether the student's reasoning focuses upon concern for society-at-large, on concern for self, or at some intermediate point (Rest et al., 1999). The P score indicates the extent to which respondents prioritize "postconventional statements" in their responses, reflecting higher-level moral reasoning.

Using the DIT in this context is not without controversy. Of greatest concern is the DIT's relevance to the overall scheme of ethics education. At best, the DIT registers moral discernment and indicates the potential for professional behavior. It remains for the educator to turn that potential into reality. To date, we note that at least 19 universities and colleges in the U.S. and Canada have administered the DIT across disciplines as diverse as medicine, education, and accounting (Patenaude, 2003; Self & Olivarez, 1996; Derryberry, Snyder, Wilson, & Barger,

2006; Lan, McMahon, Rieger, King, & Gowing, 2005; also Chu-Ting Chung, personal communication, 25 September 2007). Schools that have administered the test to medical students include the University of Sherbrooke, New York University, the University of North Dakota, the University of Iowa, the University of Nebraska, Texas A&M, the University of Manitoba, Zagreb University, Tel Aviv University, Ben-Gurion University, the University of Tokyo, Sunkyunkwan University and the University of Copenhagen.

These administrations were turned to multiple purposes. Besides gathering baseline data, the DIT has been used in cohort studies to track changes over time by administering the test at various points in the cohort's progression to graduation (Hren, Marusic & Marusic, 2011; Patenaude et al., 2003; Self & Olivarez, 1996). Studies have also been used to evaluate and assess instructional practices, e.g. the efficacy of small group discussion (Self, Baldwin, & Olivarez, 1998a). Researchers have also suggested using the DIT to screen medical school applicants (Benor, Notzer, Sheehan, & Norman, 1984; Self, Wolinksy, & Baldwin, 2000). At all these institutions, the research aims were similar insofar as they sought insight into the moral capabilities of the respondents.

Methods

In August 2007, we administered the DIT-2 to members of the incoming UTMB School of Medicine (SOM) class of 2011. The instrument was administered during orientation week under the auspices of the UTMB Office of Educational Development, following an IRB-approved protocol. Respondents were informed that participation was voluntary, that anonymity would be maintained, and that test results would not become part of any student's personal record or grade. Students were then asked to provide written confirmation of their consent for use of their scores in this study. Of 236 incoming students, 233 completed the DIT-2. Thirty

tests were subsequently culled either because the student failed to sign the consent form, did not obtain the required witness signature, or otherwise failed to provide data necessary for processing the score sheets. The remaining 203 tests were submitted to the Center for the Study of Ethical Development at the University of Minnesota (which oversees the administration and scoring of the DIT and the archiving of results). Eight of the score sheets were purged at the Center due to internal inconsistencies in the student's responses, leaving 195 valid tests (195/233 or 83.7%) for scoring.

For this initial administration of the DIT-2, we constructed three hypotheses:

(1) The mean P score of the UTMB SOM class of 2011 would be comparable to that reported for students at other medical schools in the U.S. and Canada.

(2) The mean P scores for women students would be higher than those recorded for men, consistent with results from other administrations of the DIT-2.

(3) The mean P scores for students who indicated English was their primary language would be higher than those who did not. This hypothesis was constructed to analyze for cultural bias in the test, language serving as a rough surrogate for culture.

The results of the DIT-2 administration at T1 with respect to Hypotheses 1-3 are presented in Tables 1-3.

In the spring of 2011, we administered the DIT-2 to members of the class of 2011 nearing graduation. For this administration, we invited students to use an on-line version, reminding them of the anonymous, voluntary nature of the test and confirming the consents obtained at T1. To incentivize participation, $25 gift certificates were distributed to participants. Of 232 graduating medical students, 91 completed the on-line DIT-2; on submission to the Center for

the Study of Ethical Development (relocated to the University of Alabama), 19 were purged, leaving 72 valid tests (72/91 or 79.1%) for scoring.

For this follow-up administration of the DIT-2, we constructed two hypotheses:

(4) The mean P score of the Class of 2011 at T2 would be significantly increased over the score at T1.

(5) The mean P scores for women students would be higher than those recorded for men, and the variance would be consistent with that observed at T1.

The results of the T2 DIT-2 administration with respect to Hypotheses 4 and 5 are presented in Tables 4 and 5.

At both T1 and T2, analysis of variance was used to test for P-score differences between men and women students and between students with English as their primary language and those with another primary language. We also calculated effect size statistics using Cohen's d, a standardized difference statistic, for these two comparisons to allow side-by-side comparison of the differences.

Results

[Table 1]

Hypothesis 1: For the 195 processed scores of incoming UTMB medical students, the P score mean was 43.67 (sd = 14.53) out of a possible 95 points. The Center for the Study of Ethical Development identifies the 40s score range as the norm for college students, a range appropriate for students beginning medical school in the U.S.

[Table 2]

Hypothesis 2: Women (mean P score = 46.6, sd=14.9) scored somewhat higher on average than men (mean P score =40.9, sd=13.6). This difference was statistically significant (F=7.624, df=191, p=.006). The Cohen's d effect size statistic for this difference was .40.

[Table 3]

Hypothesis 3: The mean P scores for those students indicating English as their primary language (mean=44.1, sd=14.4) was slightly higher than the mean P score of those who did not (mean=39.2, sd=14.6). This difference, however, was not statistically significant (F=3.8, df=191, p=.188). The Cohen's d effect size statistic for this difference was .34.

[Table 4]

Hypothesis 4: For the 195 processed scores of incoming UTMB medical students, the P score mean was 43.67 (sd = 14.53) out of a possible 95 points. For the 72 processed scores of graduating UTMB medical students, the P score mean was 43.31 (sd = 14.65) out of a possible 95 points. There was no statistical difference between P score means for the UTMB SOM Class of 2011 at the beginning and end of their exposure to the P-H curriculum.

[Table 5]

Hypothesis 5: Women (mean P score = 46.6, sd=14.9) scored somewhat higher on average than men (mean P score =40.9, sd=13.6). This difference was statistically significant (F=7.624, df=191, p=.006). The Cohen's d effect size statistic for this difference was .40.

Discussion

We anticipated that the P score results for the incoming UTMB SOM class of 2011 would be similar to scores from comparable populations reported by other North American medical schools. We found that the mean P score for the UTMB students was consistent with the scores reported by these schools (see Table 1), most of whose mean P scores fell within the 40s (Manitoba's results were slightly below that range, those of the University of Nebraska slightly above, and those of the University of Iowa markedly higher).

We hypothesized that mean P scores would vary by student gender. Our analyses showed a statistically significant main effect for P scores by gender both at T1 and T2. Interpreting that finding requires a caveat. This observation is not intended to influence the debate about whether and to what extent these male/female distinctions might reflect biology or the social construction of sex roles. Of note, Rest et al. (1999) argue that they "know of no evidence to challenge the conclusion that sex is a trivial variable in accounting for DIT variance" (PAGE NUMBER). Our findings merely confirm those of other studies and suggest the need for further investigation of the relationship between gender and DIT performance, as well as the potential differential effect of medical education on the moral reasoning of male and female students.

Our third hypothesis anticipated a difference in mean P scores between students who reported English as their primary language and the mean score of those who did not. This was a salient consideration, given the UTMB student body's marked diversity. In 2007, UTMB ranked 5th among U.S. medical schools in Hispanic graduates, 30th in African-American graduates, and 38th in Asian-American graduates. We used language as a surrogate variable for culture in an attempt to discover any cultural biases in our DIT-2 scores. The language variable is of limited

use for this purpose, as language only very roughly approximates culture. Cultural distinctions

need not correlate with primary language (cf. native-born American Blacks and Whites). We did

not find a statistically significant difference in mean P scores for the two language groups, so our

concerns that DIT scores for our students might be culturally biased were somewhat mitigated.

However, the effect size statistic (Cohen's d =.34) revealed a borderline moderate effect size for

language-related differences, so we advocate caution in drawing conclusions from the inferential

statistical analysis. Only 17 students of the 195 in the T1 study indicated that English was <u>not</u>

their primary language (and only 5 of the 72 in T2 – a sample too meager to analyze); the

unequal size of the comparison groups in this analysis may have made it difficult to measure

accurately a true non-zero mean-score difference. Additional studies involving larger

comparison populations of non-primary English speakers are needed to clarify the role of culture

in the measurement of moral reasoning and the DIT-2's sensitivity to cultural differences in U.S.

medical students.

The students in the T1 study scored in a range consistent with the performance observed

at other medical schools, a gratifying finding insofar as it indicates 1) the validity of our DIT-2

results, and 2) that the UTMB SOM class of 2011 on admission was, on average, at least as

advanced in its moral reasoning as 1st year students at other medical schools or those with

published DIT results. But what is the nature of that reasoning?

As noted above, the DIT is based on the neo-Kohlbergian framework of Rest, et al.

(1999). That framework specifies three progressive schema of moral reasoning. The first

schema represents reasoning based on "personal interest" and the good or ill done to the

individual and those close to him or her (Rest et al., 1999). At T1, our students achieved a mean

P score in the 40s (43.67) which correlates with the second schema, "conventional" moral

reasoning that concentrates on "maintaining norms" (Rest et al., 1999). Such reasoning is preoccupied with formalized rules, universally applied. Adherence to defined roles reinforces a decision-making hierarchy that defers to authority and duty. Thinkers within this schema prefer instruction in following a moral code as opposed to reflection on the purpose and meaning of the code.

We suspect that these characteristics ring familiar with our fellow teachers. In our own experience with classroom ethics discussions, we have observed that, e.g. many neophyte medical students are eager to know "the rules;" we have seen discussion foreclosed by citation of a relevant legal statute, as if legal precedent silenced debate. Nor is such thinking limited to students. Medical faculty attending such discussions have been known to interrupt with "real world" advice for the students, telling them, "don't worry about ethics, find out what the lawyers say." Rule-mindedness is evident as well in the beginning students' avidity for clear definitions of their roles as students and their future duties as physicians. Many show a near-reverential regard for the authority of science and are reluctant to question the "truths" delivered in their textbooks and journals. Students frequently request more "objective" testing of their knowledge and criticize "subjective" evaluation methods such as OSCEs, essays, assessments of small-group participation and the like, describing them as observer-dependent and unfair. For the novice, authority is vested, if not in instructors, then in a presumed realm of neutral facts where "proofs" dwell and normative standards of behavior bear equally on all who follow the rules. According to that standard, cheaters are condemned not by the Socratic judgment that they bring dishonor on themselves, but because their failure to play by the rules harms the system. Students often appear less upset by the impropriety than by the potential disadvantage to them. These ways of thinking are consistent with the "conventional" schema of moral development.

The third schema is characterized by "postconventional thinking." Postconventional thinking involves a flexible and more reflective approach to ethical norms (Rest et al., 1999). Postconventional thinking recognizes that there are various ways to organize society. Morality depends not only on maintaining a normative structure but more profoundly on the moral purpose behind norms. Ideals regulate behavior. Social action is consensual and based on shared ideals. Postconventional thinkers are aware of the fallacy of supposing that things *as they are* are *as they ought to be*. We find that these elements resonate well with the aspirations of professionalism. As members of a self-regulating profession, physicians have options for organizing the practice of medicine, both within the profession and in relation to society-at-large. "Professionalism" expects choices to be made relative to an ideal, most notably, the primacy of patient interests. In preserving the values of the profession, authoritarian recourse to rules is untenable. Reflection and reasoned debate are prerequisite to ethics and to moral and professional maturity.

We accept the neo-Kohlbergian framework insofar as it squares with our pedagogic charge to promote professionalism. "Maintaining norms" strikes us as a fair representation of the moral schema employed by most of our 1st year medical students, whereas the cognitive abilities involved in postconventional thinking recall the professionalism to which we aspire. From a neo-Kohlbergian point of view, our task then is to move students forward in their moral reasoning ability, from a condition bound by convention to a state of postconventional reflection and idealism consistent with professional identity (P scores≥ 50). Insofar as our task involves a neo-Kohlbergian progression, the DIT-2 is indeed the appropriate metric for measuring changes in the students' moral judgment: whether progress, stasis or regression.

Assessing moral judgment

In order to make maximal use of the DIT metric, we plan to re-administer the instrument to the class of 2011 in year 4 of their medical school education to assess changes in moral reasoning after experiencing most of the 4-year P-H curriculum. We also hope to determine specific factors contributing to change. Postconventional thinking requires opportunity for reasoners to reflect on choices and negotiate moral consensus. The P-H curriculum is designed to take advantage of the small group, problem-based format that prevails at UTMB to encourage "reflective capacity," by weaving ethics discussions, reflective writing and mentoring opportunities into existing courses, clerkships and electives. We are eager to see if the investment pays off: will UTMB medical graduates, at the completion of their professionalism training, show improved moral reasoning along with improved clinical skill and scientific medical knowledge? Likewise, if students fail to progress, we would be interested in determining factors contributing to stasis or regression. Our longitudinal assessment will assist in that determination.

We conclude by declaring our hope that researchers at other schools, employing a variety of methods for ethics instruction, will undertake DIT-2 studies to assess the efficacy of their moral training strategies. In the late 1980s through the 1990s, the DIT enjoyed a vogue among medical educators that, for various reasons, has waned. The increased interest in developing professionalism warrants giving the test a second look as the academic medical community meets the challenge of cultivating – and assessing – the "ineffable".

References

Association of Graduate Medical Education (ACGME).(2005). *Outcome Project: Enhancing residency education through outcomes assessment; general competencies*. Retrieved February 22, 2008, from http://www.acgme.org/outcome/project/proHouse.asp.

Baldwin, D.C., Daugherty, S.R., & Self, D.J. (1991). Changes in moral reasoning during medical school. *Academic Medicine*, *66*(September Supplement, 9), s1-s3.

Bebeau, M.J. & Thoma, S.J. (2003). *Guide for the DIT-2: A guide for using the Defining Issues Test, Version 2 ("DIT-2") and the scoring service of the Center for the Study of Ethical Development*. Minneapolis, MN: University of Minnesota.

Benor, D.E., Notzer, N., Sheehan, T.J., & Norman, G.R. (1984). Moral reasoning as a criterion for admission to medical school. *Medical Education*, *18,* 423-428.

Character Education Partnership (CEP). (2005). Moral reasoning/values. Retrieved February 18, 2008, from http://www.character.org/site/c.gwKUJhNYJrF/b.995179/k.8A2D/Moral_ReasoningValues.htm.

Daniels, M.H., & Baker, G.L. (1979). Assessing the moral development of medical students: An empirical study. *Proceedings of Annual Conference on Research in Medical Education, 18*, 87-92.

Derryberry, W.P., Snyder, H., Wilson, T., & Barger, B. (2006). Moral judgment differences in education and liberal arts majors: Cause for concern? *Journal of College and Character*, *7*(4), 1-10.

Elm, D. & Weber, J. (1994). Measuring moral judgment: The Moral Judgment Interview or the Defining Issues Test? *Journal of Business Ethics, 13*, 341-355.

Fleisher, W.P., Kristjanson, C., Bourgeois-Law G., & Magwood, B. (2003). Pilot study of the Defining Issues Test. *Canadian Medical Association Journal, 169*(11), 1145-1146.

Givner, N. & Hynes, K. (1983). An investigation of change in medical students' ethical thinking. *Medical Education, 17*, 3-7.

Lan, G.P.W., McMahon, S., Rieger, F., King, N., & Gowing, M.P. (2005). Differences by gender in the moral reasoning, personal values, and value types of accounting majors: A study. *Online Conference Proceedings for the Association to Advance Collegiate Schools of Business (AACSB) International*. Retrieved December 1, 2007, from http://www.niagara.edu/ciaer/2005/documents/LANPAPER.NIAGARA.doc.

Latif, D.A. (2001). Moral reasoning: should it serve as a criterion for student resident selection in pharmacy? *American Journal of Pharmaceutical Education, 65* (Summer),119-124.

McGaghie, W.C., Mytko, J.E., Brown, W.N., & Cameron, W.C. (2002). Altruism and compassion in the health professions: A search for clarity and precision. *Medical Teacher, 24,* 374-378.

Rest, J.R., Narvaez, D., Bebeau, M.J., & Thoma S.J. (1999). *Postconventional moral thinking: A neo-Kohlbergian approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Self, D., Baldwin, Jr. D.C., & Olivarez, M. (1998a). The amount of small-group case-study discussion needed to improve moral reasoning skills of medical students. *Academic Medicine, 73*(5)*,* 521-523.

Self, D.J., Baldwin, C. Jr., & Wolinsky, F.D. (1992). Evaluation of teaching medical ethics by an assessment of moral reasoning. *Medical Education, 26*, 178-184.

Self, D. & Olivarez, M. (1996). Retention of moral reasoning skills over the four years of medical education. *Teaching and Learning in Medicine*, *8*(4)*,* 195-199.

Self, D., Olivarez, M., & Baldwin, D.C. (1998b). Clarifying the relationship of medical education and moral development. *Academic Medicine, 73*(5)*,* 517-520.

Self, D., Wolinksky, F.D., & Baldwin, Jr. D.C. (2000). Should moral reasoning serve as a criterion for student and resident selection? *Clinical Orthopaedics and Related Research*, *378*, 115-123.

Table 1: Mean P-score comparison

School	Mean P score	SD	N (1st year medical students)
UTMB (2007)	**43.67**	**14.53**	**195**
University of Manitoba (2003)[a]	39.4	Not available	75
Texas A&M (1998)[b]	47.7	Not available	95
Texas A&M (1996)[c]	44.57	Not available	n=25
	46.25	Not available	n= 97*
Texas A&M (1992)[d]	47.28	Not available	n=39**
Rush Medical College (1991)[e]	41.05	Not available	n=27 (male)
	50.55		n=18 (female)
University of Nebraska, Omaha (1983)[f]	50.75	12.48	51
University of Iowa (1979)[g]	52.59	Not available	60

*respondents who did not submit DIT data over 4 year period

** experimental group in study

a. Fleischer, Kristjanson, Bourgeois-Law, & Magwood, 2003

b. Self, Olivarez, & Baldwin, 1998b

c. Self & Olivarez, 1996

d. Self, Baldwin, & Wolinksy, 1992

e. Baldwin, Daugherty, & Self, 1991

f. Givner & Hynes, 1983

   Daniels & Baker, 1979

Table 2: Descriptive statistics and one-way ANOVA table for P score by gender

comparison*

Descriptive statistics

Post Conventional   (P score)

	N	Mean	Std. Deviation	95% Confidence Interval for Mean	
				Lower Bound	Upper Bound
Male	100	40.9241	13.60046	38.2255	43.6227
Female	93	46.5851	14.88123	43.5204	49.6499
Total	193	43.6519	14.47519	41.5968	45.7071

Post Conventional   (P score)

	df	F	Sig.
Between Groups	1	7.624	.006
Within Groups	191		
Total	192		

*Two respondents did not indicate their gender.

Table 3: Descriptive statistics and one-way ANOVA table for P score by language comparison.*

Descriptive statistics

Post Conventional   (P score)

	N	Mean	Std. Deviation	95% Confidence Interval for Mean	
				Lower Bound	Upper Bound
Yes	176	44.0794	14.43230	41.9323	46.2264
No	17	39.2269	14.60737	31.7165	46.7373
Total	193	43.6519	14.47519	41.5968	45.7071

Post Conventional   (P score)

	df	F	Sig.
Between Groups	1	1.749	.188
Within Groups	191		
Total	192		

*Two respondents did not indicate whether English is their primary language.

Table 4: Comparison of UTMB SOM Class of 2011 P scores at T1 and T2

DIT-2 Administration	Mean PSCORE	Std. Deviation	N
T1	43.67	14.53	195
T2	43.31	14.65	72