Moral judgment in computing undergraduates

Suzy Jagger
Roehampton University, London, UK

Received 29 May 2010
Revised 14 November 2010
Accepted 12 January 2011

Abstract

Purpose – The purpose of this paper is to examine whether, when teaching professional ethics, the educational interventions have any effect on improving students' moral decisions. One method often used to measure change is the well-established defining issues test – an American test based on Kohlberg's stage theory.

Design/methodology/approach – Using this test, two before-and-after studies were carried out on cross-cultural cohorts of first year computing undergraduates which both received the same lectures, debates and moral-decision-making exercises.

Findings – One study showed a significant increase in moral judgment whilst the other showed a decrease (although not significant). Both studies indicated mean scores far below the American averages.

Research limitations/implications – As both studies involved relatively small sample sizes, the results are indicative rather than conclusive. However, they bring to light issues associated with the test, in both American and non-American research, indicating that lower than average mean scores could be due to cross-cultural and situational variations.

Practical implications – The paper questions the premise of stage theory as a method for measurement within a cross-cultural context; and the usefulness of measuring one component of moral development (moral judgment) in isolation.

Originality/value – The paper proposes that tests based on more discipline-specific skills, rather than stage theory, would be of greater use in evaluating student levels of moral development.

Keywords Ethics, Undergraduates, Decision making, Professional ethics

Paper type Research paper

1. Introduction

Quick and easy information access via the internet and networked databases has had a positive social and economic impact by providing new possibilities and limitless flexibility in the way we live and work. This technological capability has brought with it, a host of innovative ways for exploitation of systems and networks making the development of courses in computer ethics for would-be professionals an essential part of the curriculum on undergraduate computing courses. This paper examines the impact of a computing ethics module on the moral judgment (as defined by a psychometric testing method) of two small cohorts of first year computing undergraduates. Using empirical research, the study attempts to determine the extent in which teaching computing ethics helped students develop their moral frameworks to enable them to ethically navigate within the prevailing technological environment.

Empirical research in the area of teaching (or assessing) ethics is problematic due to the difficulties associated with the methodology of investigation. Moral appraisal is typically done on the basis of three main areas; what participants say, do and feel (Nicholson, 1994, p. 581) and in this regard there is generally an absence of accepted criteria for assessing the meaning and truth of moral propositions (Hill, 1995).

Journal of Information, Communication & Ethics in Society
Vol. 9 No. 1, 2011
pp. 20-33

1477-996X
DOI 10.1108/14779961111123205

There are a number of studies which utilise empirical measurement to evaluate moral competency (Daniel *et al.*, 1997; Loe *et al.*, 2000; Robinson *et al.*, 2000) and most require some form of value judgment in their analysis. The defining issues test (DIT) is one such measurement which analyses moral judgment capabilities using a "postconventional" approach to moral-decision making. Staehr and Byrne (2003, p. 233) administered the test to a group of final year computer science undergraduates and concluded, "There is plenty of scope for study in a wide variety of aspects of moral development in the computing and engineering professions".

Smolarski and Whitehead (2000, p. 260) described approaches to introducing students to the study of computer ethical issues and felt the question of personal ethics was an important consideration, "It is not an unreasonable question to ask how much a student's attitudes have changed as a result of including such ethical material in a technical course".

The component of moral judgment is only one aspect of moral development as recognised by Rest (1984) in his four component model (the others being moral sensitivity, moral motivation and moral character. It follows that developing the skill to think critically and choosing postconventional solutions will not necessarily transmit into a change of ethical attitude or action and in this way the test used for this study (the DIT) is designed to provide an indication of a particular skill in postconventional cognitive thinking rather than evidence of attitudinal or behavioural change.

Huff *et al.* (2008, p. 306) comment on the DIT:

> [. . .] this sophisticated measurement focuses on one aspect (giving reasons) of one category (reasonableness) of the skills and knowledge relevant to successful moral action. Better, though, would be skill and knowledge measurement that was specifically targeted to the issues and needs of professional action.

The idea of developing skills through expertise is an area explored by Huff and Rogerson (2005) in their work with moral exemplars and is discussed in the conclusion.

The defining issues test and moral development

Piaget engaged in morality research in the 1930s, but it was Kohlberg's work from the 1960s that shaped the theory of cognitive-developmental moral philosophy. He is best known for his moral judgment interviews and his six stage theory. This theory is the most popular and tested theory of moral reasoning and remains the most cited work in contemporary behavioural science (Trevino, 1992; Endler *et al.*, 1978). Kohlberg's theory associated three levels to moral judgment: pre-conventional, conventional and postconventional and split each of these levels into two categories, giving six stages in total.

Rest (1984), a contemporary of Kohlberg's, developed the DIT as a method for measuring moral judgment that did not require interview. It is a paper and pen test which takes about half an hour to complete and uses generic hypothetical ethical dilemmas in which the participant is asked to rate 12 issues for each story and rank by order of importance. The ease of delivery has made the test highly accessible across the USA and, to a lesser degree, globally. It has been peer reviewed in over 400 publications and has gained credence across the academic community as a robust method for measuring moral judgment and evaluating the effectiveness of educational interventions[1]. The test is based on the premise that depending on a person's level of moral judgment, they will "interpret moral dilemmas differently, define the critical issues of the dilemma differently, and have different intuitions about what is right and fair in a situation" (Rest, 1986b, p. 196).

The test uses three scores or schemas which are analogous to Kohlberg's stages. The personal interest (PI) score denotes stages 2 and 3; the maintaining norms (MN) score represents stage 4 and the postconventional score (P-score) stages 5 and 6. These stages are linked to levels of mental maturity. Children have a moral level of stages 1-3 being more personally orientated with a limited social conscience – hence the PI label. Stage 4 relates to those who have reached a level of social awareness and integration – they are motivated to obey the law and maintain the status quo – the MN schema. Most adults, to varying degrees, reside at this level. The ultimate goal from educational interventions at formal education level would be a progression to stage 5 or 6 – that of the principled or postconventional thinker. Such thinking involves a level of moral reasoning whereby a person can make moral judgments which provide long-term benefit to society. Such persons will often question the law but rarely break it unless it directly contravenes a moral principle. They will have high-level skills in critical analysis and be able to view issues akin to a moral philosopher in which all aspects and perspectives of that issues are considered and weighed against levels of harm. The test evaluates the level of postconventional thinking by calculating a P-score from the ranked data.

The DIT-2 is a new version of the test which includes a number of amendments. As well as the three scores, the DIT-2 includes the N2 score. The N2 compares changes that take place in the P-score against reductions in the PI score and research has found it to outperform the P-score in pre- and post-test experiments. The higher effect sizes achieved suggests the N2 is more sensitive to educational interventions (Rest *et al.*, 1997, p. 501). Research using the DIT has suggested that educational interventions, particularly at university level, can have an impact on student moral judgment (Rest, 1986a, p. 81; Rest *et al.*, 1997, p. 501).

This study is an examination of students' level of moral judgment before and after taking part in a computing ethics module, using the DIT as a method for analysis. It examines two cohorts – one in 2006-2007 and other in 2007-2008, in an attempt to ascertain the level of impact the course may have had on students' propensity to make moral judgments.

2. Research method

The computing ethics classes ran every spring term as a compulsory module for computing undergraduates. Typically these students are of average academic ability having gained at least three Cs at A level and a B in maths GCSE for entry to the course; as well as a number of mature students who gain entry through non-traditional methods. Being a London University, the course has a rich cultural diversity with anywhere between 20 and 30 per cent of students being of Middle Eastern, African or Asian decent, 10 per cent from European backgrounds and the remainder being a mixture of English ethnic mixes. English was the first language for roughly 60 per cent of the cohort.

The same lesson plans were used for both studies and incorporated a number of teaching methods to develop moral reasoning skills. Students received a total of 36 hours of ethics instruction, over a period of 12 weeks. Weekly interactive lectures discussed ethical and legal issues associated with the computing profession. Student-led structured debates (which were linked to assessment) focused on ethical topics designed to challenge pre-conceived ideas (illegal downloading of music, the effects of violence in computer games, etc.) and develop argument construction skills. A number of group

workshops were delivered designed to improve critical analysis skills and develop moral decision-making techniques.

All students who attended the class in week 2 were given the DIT to complete. Ten minutes was spent explaining how the test was structured with students being advised to select "not important" for any issue, they did not understand (as advised by the test designers). The test took approximately 35 minutes to complete. The process was repeated in week 12 to the same cohort. The test papers were sent to the DIT centre for analysis and scores were returned in digital form, which included a breakdown of scores as three schemas: P-score, PI and MN as well as an N2. The N2- and P-score are the scores most focused on in this paper as they are the most validated. However, the PI and MN scores are briefly mentioned due to their impact on the N2 score.

2.1. DIT Study 1 (2006 cohort)

2.1.1. Participants. Participants represented an opportunistic sample of first year computing undergraduates who were present in weeks 2 and 12 of the ethics class. From the tests collected, only those who had completed both tests by being present at both sessions were used. The mean age of the class was 23.8 years. The sample reflects the gender imbalance that exists in the field of IT with more men than women.

Of the tests collected in weeks 2 and 12, a total of 17 students were matched as having completed both tests. A *t*-test confirmed no significant difference between mean scores for the entire sample and the selected 17 paired tests, suggesting the 17 pairs were a suitable representation of the whole sample. The average age of the sample was 22.8, all with English as their first language. Table I shows the breakdown of the number of tests received and the final paired sample.

After checking for blanks and illegibility, a total of 67 questionnaires were sent for data extraction by scanner to the centre. Results were returned indicating the relevant schema scores from which further analyses and correlations were carried out using SPSS.

2.1.2. Reliability testing. The sample was tested for internal consistency (story rather than item consistency) using Cronbach's coefficient alpha and produced an alpha score of 0.70. Although a borderline result, this is partly explained by the small sample size and narrow range of subjects and is thus considered acceptable (Nunnally, 1978).

2.1.3. Results. Table II shows the before and after mean scores showing a small increase in P-score and an increase in the N2 score. Figure 1 provides a graphical

Students enrolled	Male	Female	Unspecified	Total
Week 2 respondents	20	12	6	38
Week 12 respondents	20	9	0	29
Matched pre- and post-tests	13	4	0	17

Table I. Tests received and final sample selection based on pairing of pre/post-tests

Schema	Mean before	SD	Mean after	SD
Maintaining norms	38.06	17.54	34.94	12.37
Personal interest	34.04	11.19	33.29	12.96
Postconventional	24.02	14.97	26.71	14.16
N2	20.02	14.90	26.64	14.70

Table II. Schema before and after scores for Study 1

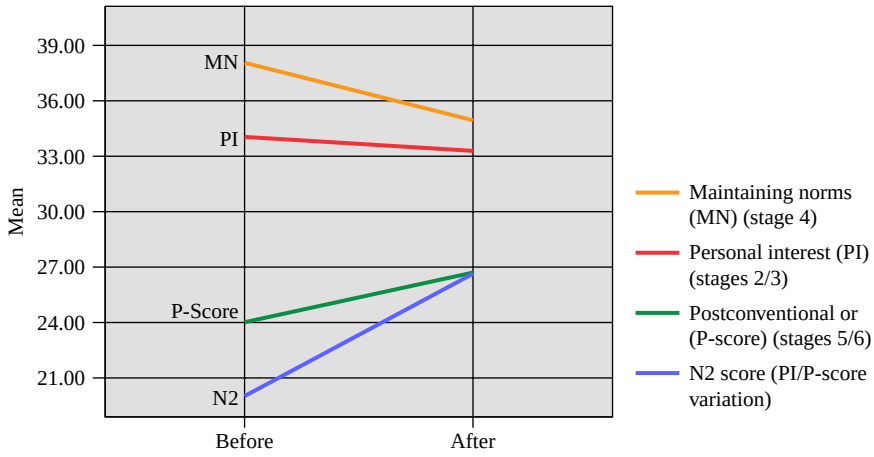

Figure 1.
Schema scores for Study 1 showing upward and downward trends with the N2 upward trend showing a significant increase

representation of the four scores. It can be seen from this graph that pre and post scores suggested favourable results – P-score increasing, PI decreasing and N2 showing a small effect size (0.25). Although the table shows all score results, this paper focuses primarily on the P-score and N2 results as these are the most externally validated. However, all schema scores are shown as they impact on overall findings and are discussed in the final section.

Postconventional scores. The international average for first year university students P-score is 32.62 ± 14.77 (Bebeau and Thoma, 2003, p. 35). The mean for this study was far lower at 24.02 ± 14.97 (Section 3.5).

Pearson's product-moment correlation coefficient confirmed there was a medium, positive correlation between the two variables of age and P-score ($r = 0.502$, $n = 17$, $p = 0.04$) with a shared variance of 25.20 demonstrating that 25 per cent of the variance of respondents' P-scores can be attributed to their age (post-test P-scores were used for this analysis). Past research confirms that there is a relationship between age and P-score and that, within educational interventions, the most powerful treatment effects take place in the adult group (0.61) and least for the secondary school group (0.22) (Rest, 1986a, p. 82).

There was an increase of 2.69 between the before and after results, although a *t*-test confirmed this result was not statistically significant.

N2 score. There was an increase of 6.62 in the N2 before and after results and a paired samples *t*-test was conducted to evaluate the impact of the intervention on these scores which revealed a statistically significant increase between the before ($M = 20.02$, SD = 14.90) and the after ($M = 26.64$, SD = 14.70, $t(17) = -2.235$, $p < 0.0005$). The effect size was calculated using η^2 (Pallant, 2005, p. 212) which produced a small effect size (0.25) (Cohen, 1988, p. 44).

2.2. DIT Study 2 (2007 cohort)

The conditions for this study were the same as for the previous study but took place a year later with a larger sample size and included a control group.

2.2.1. Participants. First year undergraduates on the computing programme formed the experimental group for this study and first year undergraduates on the business

programme (a similar programme with similar entry requirements and cultural diversity) represented the control.

As with Study 1 there was a gender imbalance between the experimental and control group with more women in the control group studying business subjects (Table III). There has been some discussion about the significance of gender in moral development but research has found no clear evidence of this (Snarey *et al.*, 1985; Gibbs and Widaman, 1982; Nisan and Kohlberg, 1982). *T*-tests on mean data showed no significant difference between male and female scores. Also no significant difference was found between those for whom English was not their first language, and no significant difference existed between scores for the entire pool of students who took the first test and the final selected groups. The average age for the two initial samples was 20.3 for the experimental and 20.7 for the control.

The test was administered in week 2 of the spring semester to 32 students in the ethics class (experimental group) and 82 in the business (control group) and again in week 12 to 39 students in the experimental group and 49 in the control group.

Table III shows the gender split between the groups, the number of tests collected for each week and those matched together, and the number of tests purged on the basis of the test's reliability checks. After checking for blanks and illegibility, a total of 202 questionnaires were sent for data extraction by scanner to the centre. Results were returned indicating the relevant schema scores.

Only respondents who had completed both a pre- and post-test were used for this study. This consisted of 27 students in the case of the test group and 36 in the case of the control. *T*-tests done on unused tests within both the experimental and control groups showed no significant difference in the mean between the used and unused tests ensuring that removing these tests did not create bias.

2.2.2. Reliability testing. All groups were tested for internal consistency (story rather than item consistency) using Cronbach's coefficient alpha. Although Nunnally (1978) recommends a minimum level of 0.7, Rest and Narvaez (1997, p. 56) maintain that where a sample focuses on a narrow range of results (in this case a group of peers at the same educational level) the α is likely to be affected. In this study for all groups combined the α was 0.653, just under the desired result. Table IV shows the Cronbach alpha scores for each group.

2.2.3. Results. This section details the pre and post mean schema scores and compares the N2- and P-score against international averages, correlates age to P-score, and graphs upward and download trends of the P and N2 scores between the two groups.

DIT groupings	Week 2			Week 12		
	Male	Female	Total	Male	Female	Total
Experimental	22	10	32	26	13	39
Control	39	43	82	22	27	49
Matched pre/post-experimental completions	18	11	29	18	11	29
Matched pre/post-control completions	20	22	42	20	22	42
Purged experimental			2			1
Purged control			5			3
Pre/post-experimental after removal of purged subjects	18	9	27	18	9	27
Pre/post-control after removal of purged subjects	17	20	36	17	20	36

Table III.
DITs collected for Study 2

A two-way analysis of variance (ANOVA) was performed on all scores but did not show any significance within or between groups ($p > 0.05$).

Postconventional scores. A comparison of the mean for this study and the international average for first year university students (32.62 ± 14.77; Bebeau and Thoma, 2003, p. 35) reveals the average for this group to be lower than the international average by a difference of 7.73. Table V displays the means and standard deviations of P-scores for both groups.

Unlike Study 1, the means for both the experimental and control reduced. Using a two-way ANOVA, it was established that those who received ethics instruction did not increase their P-score any more than did those who received no instruction ($F(1,61)$ for interaction = 0.011, $p = 0.92$). In fact, P-scores for the combined sample decreased slightly from pre to post measurement though not enough to reach significance ($F(1,61)$ for pre-post main effect = 0.497, $p = 0.48$). Figure 2 shows the P-score downward trend.

The relationship between high P-scores and age was investigated in both groups. Pearson's product-moment correlation coefficient confirmed there was a medium, positive correlation between the two variables ($r = 0.353$, $n = 0.63$, $p = 0.005$) with a shared variance of 12.46 demonstrating that 12 per cent of the variance of respondents' P-scores can be attributed to their age.

Table IV.
Cronbach's alpha for Study 2

Group	Cronbach's alpha
E1	0.71
E2	0.70
C1	0.66
C2	0.72

Table V.
Pre/post P-score means for experimental and control groups

Group	*n*	Pre P-score (mean ± SD)	Post P-score (mean ± SD)
Experimental	27	26.01 ± 14.88	24.83 ± 13.99
Control	36	24.89 ± 13.95	24.01 ± 15.53

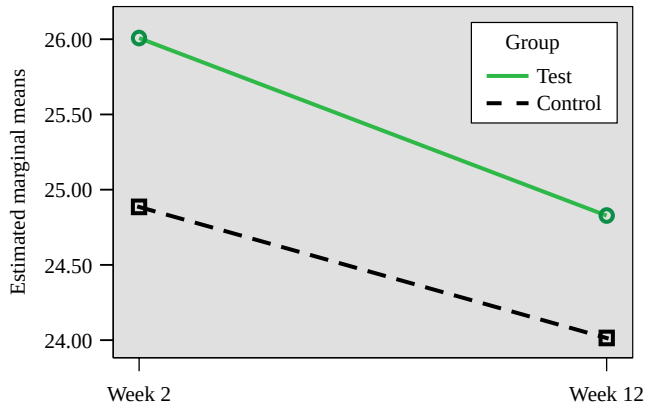

Figure 2.
Pre/post-score plot showing small decrease in Study 2 P-score by both experimental and control group over the semester although this decrease was not significant

The N2 score. The international average for the N2 score for this year group is 31.24 (Bebeau and Thoma, 2003, p. 35) which is much higher than for this sample. This score revealed an upward trend in the experimental group (change of 1.42), but this was not significant (Table VI and Figure 3).

There was a slight downward trend in the control (change of 0.41) although, a two-way ANOVA revealed this was not significant ($F(1,61)$ for interaction = 1.25, $p = 0.27$). The upward and downward trend between the pre- and post-combined samples were not significant ($F(1,61)$ for pre-post main effect = 0.80, $p = 0.38$).

3. Discussion

3.1. Study 1 – pre- and post-test comparison

Study 1 resulted in a small increase in the P-Scores with a correspondingly small decrease in the PI score which resulted in a statistically significant increase in the N2 score overall. The N2 score was developed specifically to identify more sensitive effects as a result of educational interventions by measuring the variation between the PI and P-scores and in the case of this study indicated that although P-scores do not show a significant rise, the combination of reduction in the PI combined with an increase in P-scores was significant.

3.2. Study 2 – pre- and post-test comparison

Contrary to Study 1, Study 2 actually showed the opposite trend – a decrease in P-scores and an increase in PI scores. Although the changes in both experimental and control scores were not statistically significant, there was a large increase in PI scores at the expense of MN scores (quite extreme in the case of the control group). There is very little research on MN and PI scores in the DIT with most studies focusing on the well-known P-score – the theory being that by the time students reach tertiary

Group	*n*	Pre N2-score (mean ± SD)	Post N2-score (mean ± SD)
Experimental	27	21.93 ± 14.87	23.35 ± 15.20
Control	37	21.63 ± 13.65	21.22 ± 16.29

Table VI.
Pre/post-N2 means for experimental and control groups

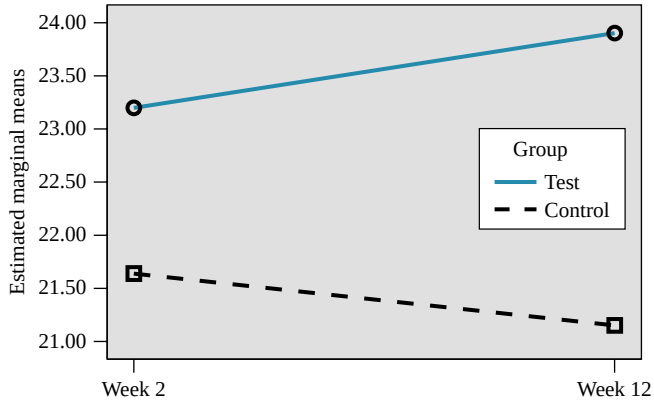

Figure 3.
Pre/post-score plot showing Study 2 increase in N2 score for experimental group and decrease for control group over the semester although this was not significant

education, the results will primarily range between MN and postconventional choices rather than PI. This study calls into question that premise. The low mean scores suggest that, if the test is to be used for this purpose, more attention and discussion should be given to the PI and MN scores in themselves and the interplay between them.

3.3. Comparison of studies 1 and 2

Cohort characteristics and educational content were very similar between the studies 1 and 2 experimental groups, and results were all within the same boundaries with no significant differences between scores between the groups. However, Study 1 suggested an improvement in moral judgment had taken place (even if small) whereas for Study 2, there is no statistical significance or even a suggestion from trends.

Being a quasi-experimental design, Study 1 did not have the benefit of a control group to determine whether the N2 increase was significant within the context of another group and the sample size was far smaller. However, the mean scores for the larger group were not statistically different. Also existence of Study 2 does provide some validation for Study 1 by showing mean scores to be comparable. For this reason it is unlikely the differences between the upward and downward trends of the two studies can be put down to study design, it is more likely differences in results can be attributed to other factors.

3.4. Comparison with past studies

A review of 55 DIT studies by Rest (1986a, p. 85) provides evidence that educational interventions which use dilemma discussions enhance moral judgment skills in students with the effect size being "modest but definite". These studies were a combination of dissertations (30), journal publications (18) and unpublished manuscripts (9). However, only nine of the studies employed a full experimental design with only 19 using analysis of covariance or two-way ANOVA (as used in Study 2) with the others adopting an approach similar to Study 1 (*t*-test without control). Thus, the claim of a modest but definite increase needs to be understood within this context. The effect sizes for these 55 studies were calculated from the P-score (for which neither Study 1 nor Study 2 showed a significant increase).

The studies cite a number of factors as relevant to educational interventions. Three of these factors: age; English as a first language; and the teaching of Kohlberg's theory to participants; are discussed in the context of this study.

Age and ethical development. The studies cite the impact of age as a significant factor on P-score effect size verifying that adults of 24 years and older respond better than younger groups to educational interventions (Rest, 1986a, p. 85). In the case of Study 1 the average age of respondents was 22.9, and in the case of Study 2, 20.3. Coefficient results from both studies show that the factor of age does have a bearing on P-score even within these small samples. Thus, it is possible that age may explain the significant increase in N2 scores for Study 1.

English as a first language. The sample for cohort 1 all had English as their first language whereas for cohort 2, 63 per cent had English as their first language. However, separate *t*-tests performed on both groups did not detect a significant difference in the P-scores between the two groups and there was no positive correlation between P-score and English as a first language. Past research suggests that the lack of English as a first language can have an impact on the validity of the test (Bebeau and Thoma, 2003, p. 10).

This may have been the cause of lack of overall significant results for the second study due to cultural interpretation rather than language difficulty.

The teaching of Kohlberg's theory. The teaching of Kohlberg's theory as part of the intervention was also cited as having a significant effect (Rest, 1986a, p. 81). Although some consider teaching Kohlberg's theory as a useful tool by giving students an indication of what defines postconventional thinking, there is also the distinct possibility that teaching the theory would create research bias by not necessarily reflecting true attitudes and for this reason (and also because it was felt irrelevant to the discipline) in these studies Kohlberg's theory was not taught as part of instruction. This may indeed have been a reason for no quantifiable effect being achieved.

3.5. Mean scores against international averages

Both studies had low mean P-scores compared with international averages (Study 1 a difference of 8.60, Study 2 a difference of 6.61 in pre-test results). There are a number of non-American studies which show very low mean P-scores and three are discussed here: Clarkeburn's (2000) work amongst 195 Bioscience students at the University of Glasgow; Wimalasiri's (2001, p. 620) study involving 106 Australian university students and Ahmed and Gielen's (2002, p. 264) review of studies undertaken across the Arab world.

Clarkeburn (2000) reasoned that the low mean could have been a result of the changing nature of university cohorts due to greater demands and pressures imposed on students and lack of ethical instruction. She concludes in her thesis that the university environment is no longer conducive for moral development to take place.

In an Australian study, Wimalasiri (2001, p. 628) encountered low P-score means using the DIT and commented that all but one dilemma in the DIT related exclusively to issues of rights and legal justice – a "narrow domain" which did not incorporate a larger moral field. It was felt that this affected the assessment's capability to measure higher levels of reasoning that may fit a broader definition. The study concluded that this factor in the DIT assessment method may explain the low scores rather than that the Australian subjects were morally inferior.

Ahmed and Gielen (2002) identified a number of methodological problems preventing conclusions being drawn about the usefulness of the DIT in Arab cultural settings. The studies produced high failures in the DIT consistency tests – particularly among Kuwaiti and Sudanese students with many students finding the DIT's moral arguments "strange" and difficult to understand.

In their comparison of the DIT against the multidimensional ethics scale (Reidenbach and Robin, 1990), Shawver and Senetti (2009, p. 675) determine that the DIT was not effective in explaining the moral reasoning of first year accounting undergraduates and comment that "sources other than cognition influence student ethical sensitivities". Similar to results from this study, they ascertain that DIT scores tend to be lower for accounting undergraduates than in other disciplines and that results tend to be "without sufficient pre and post-test variations to support statistically significant findings" (Shawver and Senetti, 2009, p. 674). They cite possible reasons may be to do with methodological issues associated with the test itself but also that the test cannot be adapted to different disciplines and situations. They conclude that alternative measures are needed.

Both this study and the studies mentioned above suggest that the DIT has cultural and situational limitations. Although Kohlberg and Rest believed ethical theory should

be robust enough to have universal application, it seems reasonable to conclude that to measure to any accuracy may require a more local and discipline-specific approach.

Bebeau and Thoma (2003, p. 35) acknowledge that even within the national spread in the USA in which thousands of studies have been used to create the averages; "Caution should be used in interpreting these scores as data were not selected to counterbalance the variability that appears to exist among college and professional school programmes".

3.6. Limitations

The relatively small size of both cohorts was a limitation with regard to this study and renders the results indicative rather than conclusive. The size of cohort meant, it was not possible to do further analysis to determine other factors which may have contributed to results such as cultural, gender or intellectual/academic variations. Being practitioner based, the study was also subject to a certain amount of bias as the practitioner was also the researcher. However, the incorporation of a psychometric testing instrument within a quantitative study will have controlled this effect to some degree.

4. Conclusion

A lower age mean and lower percentage of English as a first language for the second study have been identified as possible contributory factors to explain the difference in change scores between the two studies. The lack of effect size compared to past research may be due to the fact that, unlike a number of past studies, Kohlberg's theory was not taught. However, independent research also points to issues with the DIT, and stage theory in general as a method for evaluating moral judgment.

As experienced by Wimilarisi (2001) and Ahmed and Gielen (2002) (Section 3.5) thinking postconventionally is not necessarily an indication of moral judgment. Cultural differences may make the universalist choices required for a high P-score on the DIT not appropriate for some audiences and in some situations.

Kavathatzopoulos contends that tests which evaluate using any kind of moral philosophical framework can mislead an assessment and considers a more valid approach is to remove any moral philosophical principles from the analytical process. In his measurement model, he focuses primarily on the cognitive process involved in moral decision making as a method for analysis. However, he acknowledges that "cognitively higher ethical reasoning does not necessarily lead to better morality because there is no moral principle in the model to define what is good and what is bad" (Kavathatzopoulos and Rigas, 1994, p. 58).

Analysis of individual P-scores for the Study 2 cohort was undertaken in separate case study analyses (Jagger, 2011) and indicated large pre- and post-score variations which are not reflected in means analysis (although the high standard deviation gives an indication). Whilst some students achieved high positive P-score improvements, others followed an equally high negative trend. This large fluctuation, combined with low mean scores overall, call into question stage theory as a method of measurement. The invariant sequential nature of the theory does not allow for regression and this stringency of the theory has caused issue in some before and after intervention studies in which both the skipping of, and a reduction in, stages has occurred. Kohlberg himself experienced this phenomenon (Kohlberg and Kramer, 1969) although he does not conclude that this finding is a contradiction of the cognitive-developmental

theory, nor a fault with the scoring system seeing it as a temporary, functional phase "in the service of the ego" (Kohlberg and Kramer, 1969, p. 116). However, the existence of this "temporary, functional phase" calls into question the accuracy of any means-based testing measurement which cannot incorporate these phases within results.

A test for the professions

Moral judgment is only one component of moral development and one which requires a high level of cognitive skill. The DIT does not measure other components as identified by Rest in his four component model, such as moral sensitivity, motivation and character – all of which have a bearing on progression and ethical development. Practitioners should be guiding students in how to morally judge a situation but also motivating them to behave ethically having highlighted the social imperatives and professional benefits.

Huff and Rogerson (2005) explore a more professionally orientated alternative to Kohlberg's theory which focuses on the identification of specific skill sets relevant to the computing profession. These skill sets are comparative to Rest's four components and identify "reasonableness" as similar in nature to moral judgment. This is defined within a professional context as, "constructing data-based and reasoned arguments; engaging in reasoned dialogue; gathering relevant evidence; listening to others; giving reasons; changing plans/positions based on reason" (Huff *et al.*, 2008, p. 301). This component echoes less universalist sentiments than Kohlberg's stages, being more aligned to specific values recognised within the profession.

The development of skill sets through practice as a method for teaching professional ethics is one promoted by Narvaez and Lapsley (2005). Huff *et al.* (2008, p. 306) cite this approach:

> The more often this skill is exercised, the more proficient one will be at recognising the opportunity for its expression and the more efficient one will be at discerning options.

In conclusion, the DIT was used for this study due to its stature in the academic community as a validated, robust measurement of moral judgment, derived from a strong and well-respected theoretical base. However, results from these two studies, and other recent studies, have called into question its efficacy as a measurement tool within certain situations and cultures. Future research which supports the development of teaching and assessment measures which are flexible enough to be incorporated for specific professions (perhaps incorporating skill sets such as those identified in the Huff and Rogerson model) would be worthwhile and timely. This would help to ensure a tool that is more culturally sensitive, as it draws from values and requirements associated with the profession rather than purely generic universalist principles. Research should involve an analysis of older cohorts, such as final year undergraduates, participants on masters' programmes and employees within private and public organisations.

Note

1. DIT Instruments, Services and Materials, Center for the Study of Ethical Development, www.centerforthestudyofethicaldevelopment.net/DIT%20instruments%20service%20materials.htm (accessed 13 October 2010).

References

Ahmed, R.A. and Gielen, U.P. (2002), "A critical review of studies on moral judgment development using the defining issues test in Arab countries", *The Arab Journal of Humanities* (*Kuwait*), Vol. 77, pp. 261-81.

Bebeau, M. and Thoma, S. (2003), *Guide for DIT-2*, Center for the Study of Ethical Development, University of Minnesota Press, Minneapolis, MN.

Clarkeburn, H. (2000), *How to Teach Science Ethics*, Glasgow University, Glasgow.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, Erlbaum, Hillsdale, NJ.

Daniel, L.G., Elliott-Howard, F.E. and Dufrene, D.D. (1997), "The ethical issues rating scale: an instrument for measuring ethical orientation of college students toward various business practices", *Educational and Psychological Measurement*, Vol. 57, pp. 515-26.

Endler, N.S., Rushton, J.P. and Roediger, H.L. (1978), "Productivity and scholarly impact (citations) of British, Canadian, and US departments of psychology (1975)", *American Psychologist*, Vol. 33 No. 12, pp. 1064-82.

Gibbs, J.C. and Widaman, K.F. (1982), *Social Intelligence: Measuring the Development of Sociomoral Reflection*, Prentice-Hall, Englewood Cliffs, NJ.

Huff, C., Barnard, L. and Frey, W. (2008), "Good computing: a pedagogically focused model of virtue in the practice of computing (part 2)", *Journal of Information, Communication and Ethics in Society*, Vol. 6 No. 3, pp. 284-316.

Huff, C. and Rogerson, S. (2005), "Craft and reform in moral exemplars in computing", paper presented at Ethicomp 2005, Looking Back to the Future, Linkoping.

Jagger, S. (2011), *Studies on the Moral Development of Computing Undergraduates* (in press).

Kavathatzopoulos, I. and Rigas, G. (1994), "Training professional managers in decision-making about real life business ethics problems: the acquisition of the autonomous problem-solving skill", *Journal of Business Ethics*, Vol. 13, pp. 379-86.

Kohlberg, L. and Kramer, R. (1969), "Continuities and discontinuities in childhood and adult moral development", *Human Development*, Vol. 12, pp. 93-120.

Loe, T.W., Ferell, L. and Mansfield, P. (2000), "A review of empirical studies assessing ethical decision-making in business", *Journal of Business Ethics*, Vol. 25, pp. 185-204.

Narvaez, D. and Lapsley, D. (2005), "The psychological foundations of everyday morality and moral expertise", in Lapsley, D.K. and Power, F.C.E. (Eds), *Character Psychology and Character Education*, University of Notre Dame, Notre Dame, IN, pp. 140-65.

Nicholson, N. (1994), "Ethics in organizations: a framework for theory and research", *Journal of Business Ethics*, Vol. 13 No. 8, pp. 581-96.

Nisan, M. and Kohlberg, L. (1982), "Universality and cross-cultural variation in moral development: a longitudinal and cross-sectional study in Turkey", *Child Development*, Vol. 53, pp. 865-76.

Nunnally, J.O. (1978), *Psychometric Theory*, McGraw-Hill, New York, NY.

Pallant, J. (2005), *SPSS Survival Manual*, 2nd ed., Open University Press, Maidenhead.

Reidenbach, R. and Robin, D. (1990), "Toward the development of a multidimensional scale for improving evaluations of business ethics", *Journal of Business Ethics*, Vol. 9, pp. 639-53.

Rest, J. (1984), "The major components of morality", in Kurtines, W.M. and Gewirtz, J.L. (Eds), *Morality, Moral Behavior, and Moral Development*, Wiley, New York, NY.

Rest, J. (1986a), *Moral Development Advances in Research and Theory*, Praeger, London.

Rest, J. (1986b), *The Psychology of Morality*, Praeger, London.

Rest, J. and Narvaez, D. (1997), *Ideas for Research with the DIT*, Center for the Study of Ethical Development, University of Minnesota Press, Minneapolis, MN.

Rest, J., Thoma, S.J., Narvaez, D. and Bebeau, M.J. (1997), "Alchemy and beyond: indexing the defining issues test", *Journal of Educational Psychology*, Vol. 89 No. 3, pp. 498-507.

Robinson, R., Lewicki, R.J. and Donahue, E.M. (2000), "Extending and testing a five factor model of ethical and unethical bargaining tactics: introducing the SINS scale", *Journal of Organisational Behaviour*, Vol. 21, pp. 649-64.

Shawver, T. and Sennetti, J. (2009), "Measuring ethical sensitivity and evaluation", *Journal of Business Ethics*, Vol. 88, pp. 663-78.

Smolarski, D.C. and Whitehead, T. (2000), "Ethics in the classroom: a reflection on integrating ethical discussions in an introductory course in computer programming", *Science and Engineering Ethics*, Vol. 6 No. 2, pp. 255-63.

Snarey, J.R., Reimer, J. and Kohlberg, L. (1985), "The development of social-moral reasoning among kibbutz adolescents: a longitudinal cross-cultural study", *Developmental Psychology*, Vol. 20, pp. 13-17.

Staehr, L.J. and Byrne, G.J. (2003), "Using the defining issues test for evaluating computer ethics teaching", *Institute of Electrical and Electronics Engineers* (*IEEE*) *Transactions on Education*, Vol. 46 No. 2, pp. 229-34.

Trevino, L.K. (1992), "Moral reasoning and business ethics: implications for research, education, and management", *Journal of Business Ethics*, Vol. 11 No. 5, pp. 445-59.

Wimalasiri, J.S. (2001), "Moral reasoning capacity of management students and practitioners: an empirical study in Australia", *Journal of Managerial Psychology*, Vol. 16 No. 8, pp. 614-34.

Corresponding author
Suzy Jagger can be contacted at: s.jagger@roehampton.ac.uk