

An empirical investigation of student behaviour when non-originality detection is made available before submission

Fintan Culwin
London South Bank
University
Borough Road
London SE1 0AA

fintan@lsbu.ac.uk

Jon Warwick
London South Bank
University
Borough Road
London SE1 0AA

warwick@lsbu.ac.uk

Mike Child
London South Bank
University
Borough Road
London SE1 0AA

childm@lsbu.ac.uk

Abstract

Although there is conflicting advice as to the advisability of allowing students access to non-originality detection systems in advance of coursework submissions there appears to be no empirical evidence to support either position. This paper reports on an analysis of the pre-submissions and final reports of the final year dissertations of about 100 undergraduate BIT and Computing students. The interpretation of the results would suggest that there is no measurable impact upon student behaviour.

Keywords

Academic misconduct, academic integrity, plagiarism, non-originality detection.

Introduction

Non-originality detection systems such as JISC/Turnitin are now routinely used by educational institutions to assist with assuring the academic integrity of student submissions. However, there is no consensus on the advisability of making such systems available to students to check their own work prior to submission. Although many seem to advise that such systems should be made available (Carroll 2002) there are others who advise differently (Savage 2004). In the authors' own institution there is no common practice with some departments not making the facility available whilst others do. What seems to be common however is that there is no published empirical evidence supporting one position or the other. The study reported in this paper attempts to provide evidence in order to inform debates about the advisability of such practice.

The cohort for this study was the undergraduate students on the BIT and Computing programmes at London South Bank University (LSBU) who were submitting final year projects in 2007. The corpus consisted of the JISC /Turnitin non-originality reports of the drafts and final submissions of the final year project. The successive cohorts of students, and corpora of their submissions, have been subject to previous studies (Culwin 2006, 2008); hence the expected behaviour of the cohort was known to the investigators. Furthermore the team responsible for the final year projects had been awarded an internal teaching excellence prize, indicating both the supportive nature and stability of the environment.

The 15% cut off is the operational limit of noise in JISC/Turnitin non-originality reports as decided by the project team over a number of years. That is reports with less than 15% reported non-originality are unlikely to be compromised and are given less attention compared with reports with more than 15%. A visual tool allows a rapid determination to be made regarding the sub-15% non-originality. That is 15% distributed as fragments throughout the report can be safely ignored, whilst 15% showing as two or three concentrations would merit manual investigation.

For the 2007/2008 session the students were allowed access to the Turnitin system for one week at the end of March 2007 and encouraged to submit the drafts of their projects for non-originality investigation (referred to in this paper as the pre-submission). Students were given the rule of thumb that reports with less than 15% non original content (noc) were unlikely to attract attention for an academic misconduct investigation and reports with more than 15% non original content were more likely to attract such attention. The system was then made unavailable to the students and the final reports were required to be submitted in the middle of May 2007. In previous years a small number of students had either paid agencies for (non-Turnitin) non-originality reports or had managed to make use of the Turnitin system in other departments or institutions.

This resulted in two sets of Non-Originality Reports (NOR) one containing about 100 pre-submissions and the other about 130 final reports. The non-

originality reports identify content which can be shown to be identical to already published or submitted material. There is no guarantee that all non-original material has been identified and the identified material comprises of legitimate and non-legitimate reused material, as well as some ‘false hits’. The Culwin-Lancaster (Lancaster & Culwin 2001) four stage model emphasises that human judgment is required to identify non-legitimate reuse. In keeping with previous studies (Culwin 2006, 2008) the amount of non-original material is taken as a marker for the amount of non-legitimately reused material. That is, the extent to which any individual document is compromised is unknown but is operationally assumed to be consistently related to the amount of non-original material it contains. The legitimacy of this assumption will be explored in the body of the paper and the conclusions.

The study reported in this paper consisted of two separate investigations. The first investigation explored the demographic aspects of the corpus attempting to identify difference between: reports with and without pre-submissions, the gross differences between the corresponding pre-submissions and final reports and if there were any differences in behaviour related to the programme of study. The second investigation was more analytical and used a number of purpose built and adapted tools to describe in more detail the processes whereby the pre-submission was changed into the final report.

The Demographic Investigation

The first part of the investigation will address the question “What are the characteristics of the reports that do and do not have a pre-submission?”. Table 1 presents the gross characteristics of the corpus.

	Report with pre-submission	Report without pre-submission
N	98	28
Mean length (words)	15207	12849
Median noc (%)	8.14	8.32
Median mark (%)	58.00	53.00

Table 1. Gross characteristics of the corpus

There were a total of 126 projects of which 98 had pre-submissions and 28 did not. This would suggest that if the facility is made available then most, but not all, would make use of it. Reports with a pre-submission were, on average, longer than those without and contained less non-original content; however both of these differences were statistically insignificant. Reports with pre-submissions were graded higher than those without and this difference was statistically significant at the $p < .05$ level¹.

¹ Statistical tests were conducted using the non-parametric Mann-Whitney significance test unless otherwise noted.

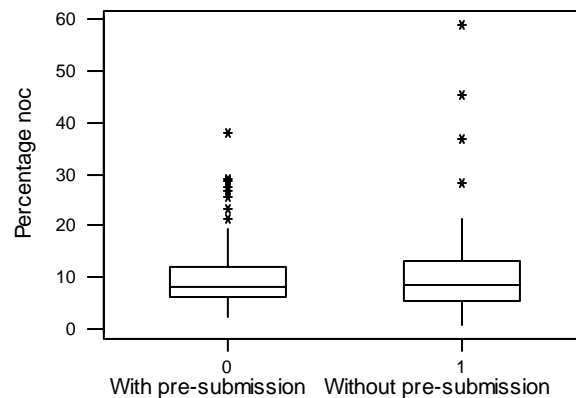


Fig. 1. Differences in non-original content between reports with and without pre-submissions

Fig. 1 illustrates the differences between the amount of non-original content in reports with and without a pre-submission. The information is shown as a box plot where 50% of the values are contained within the boxes. Although the medians of both groups are very similar, the non pre-submission group is more variable with some dramatic outliers. The variability of the data is shown in the height of the boxes which defines the inter-quartile range (IQR). The occurrence of outliers in this and subsequent data sets makes the median a better measure of average than the arithmetic mean. The corresponding box plots for the length of the reports and the mark obtained showed a similar pattern, with noticeably greater variability in the no-pre submission data.

The conclusion from this part of the study might be that the students who did not make use of the facility submitted projects that were graded lower than those who did make use of the facility. Or, put the other way around; weaker students were less likely to make use of the facility. The increased non-original content in the weaker group agrees with the conclusion of a previous study by the authors (Culwin 2006). The reasons why some students did not make use of the facility are unknown but might be assumed to be related to lowered engagement and organisation; both of which have been related to increased academic misconduct (Caruana *et al.* 2000). This possibility is supported by this investigation, where Fig. 2 shows the scatterplot of the extent of non-original content against mark awarded. There is a weak (although insignificant) relationship with larger amounts of non-original content gaining fewer marks. Lower marks are associated with lower engagement and organisation as well as innate ability.

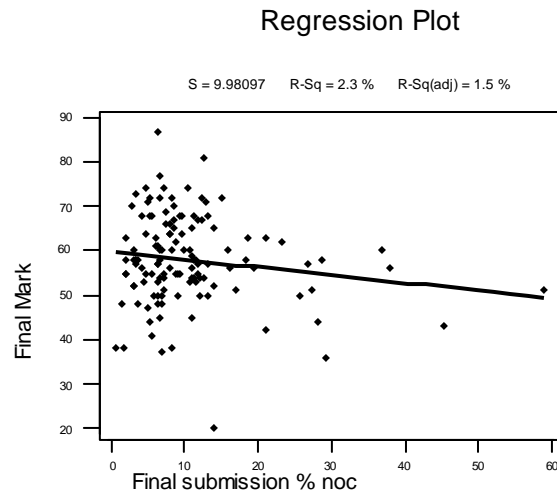


Fig. 2. Scatterplot of amount of non-original content and mark obtained

The second part of this investigation will address the question “What are the superficial relationships between the pre and post submissions?”. Table 2 presents the results of the analysis of the 98 reports which had a pre-submission.

	Pre-submission	Final report
Mean length (words)	8571	15207
Median length (words)	7103	13968
Median noc (%)	10.91	8.14

Table 2. Gross characteristics of the pre-submissions and final reports.

Unsurprisingly the differences between the mean and median lengths were statistically significant at the $p < 0.01$ level. The mean and median non-original content levels were higher in the pre-submissions than the corresponding final reports. These differences were also statistically significant at the $p < 0.01$ and $p < 0.05$ levels respectively.

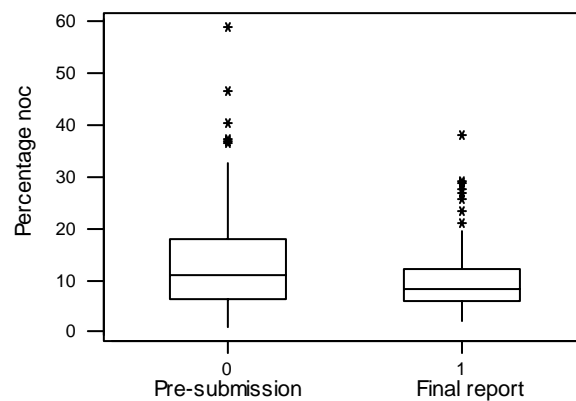


Fig. 3. Differences in non-original content between pre-submissions and final reports.

Fig. 3 shows the box plot for the non-originality data. The final reports show less non-original material overall, are less variable and contain fewer extreme outliers. This data is also shown in the scattergram in Fig. 4.

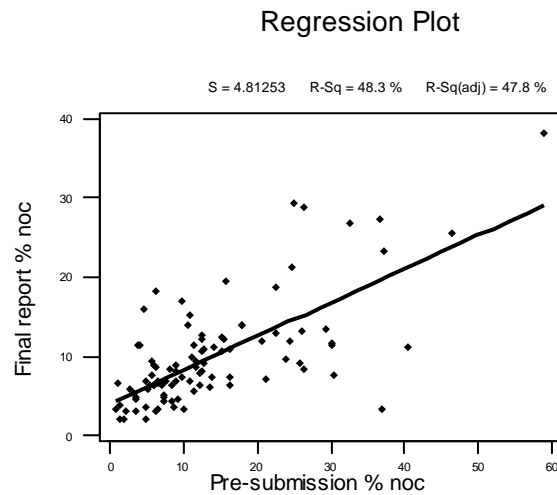


Fig. 4. Scattergram of pre-submissions and final report non-original content

The majority of the points fit close to the regression line with relatively few outliers, indicating that the proportional change in non-original content was fairly constant over the whole range of behaviours. Of the 98 projects analysed 33 (34%) showed an increase in non-original content with the remaining 65 (66%) showing a decline. The box-plot of the pre-submission non-original content of these two groups is shown in Fig. 5.

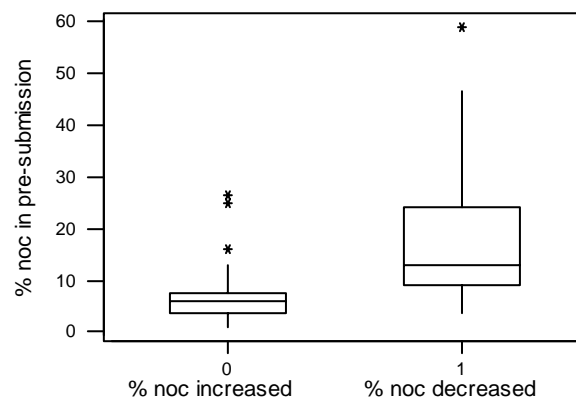


Fig. 5. Differences between projects where non-original content increased and decreased between pre-submission and final report.

The two groups are clearly quite distinct. The group who reduced the non-original content of their reports had a significantly higher ($p < 0.01$) median level of non-originality in their pre-submissions. The median of this group also corresponds closely to the 15% rule of thumb non-original content level given to the students.

The conclusions from this part of the investigation might be that when the facility is made available the amount of non-original content drops and that this effect is particularly pronounced for submissions that have relatively large amounts of non-original content. However the projects also increased in size between the pre-submission and final report and it might be that the most compromised material was already present in the pre-submission and was 'diluted' by less compromised material in the final report. This is a possibility which will be returned to in the second investigation in this report.

The final part of this investigation will address the question "Are there any differences between programmes of study?". The students are studying in either the more technical Computing area, or the less technical Business Information Technology (BIT) area. A previous study by one of the authors had concluded that BIT students were more prone to academic misconduct (Culwin 2006).

	BIT	Computing
N	90	36
Median length (words)	14441	12219
Median noc (%)	8.63	7.27
Median mark (%)	57	58

Table 3. Gross characteristics of the BIT and Computing final reports

Table 3 shows the characteristics of the BIT and Computing final reports. The differences in length were shown to be significant at the $p < .05$ level, but the other two factors were statistically not significant. A more detailed analysis failed to find any significant interactions between 'programme and presence or absence of pre-submission' and 'programme and increase or decrease in non-original content'. Hence the conclusion from this part of the investigation is that, apart from the length of the final report, there are no differences between the two programmes of study.

The Analytic Investigation

In this investigation an attempt was made to divide the reports into 'clean' and 'dirty'² text and then to investigate the relationships between the clean and dirty parts of the pre-submission and final report. Fig. 6 illustrates the basis of the first, preparatory part of the investigation. The left of the diagram illustrates the pre-submission and final report non-originality reports and divides the contents of each of the files into clean and dirty parts. The right hand side of the diagram illustrates the two non-originality reports being split into four plain text files containing the pre-submission clean text (**pc**), the pre-submission dirty text (**pd**), the final report clean text (**fc**), the final report dirty text (**fd**).

² The phrases 'clean' and 'dirty' are being used as a shorthand to designate text that has not been marked as non-original and text that has been marked as non-original in the reports. No further connotation of academic misconduct is implied.

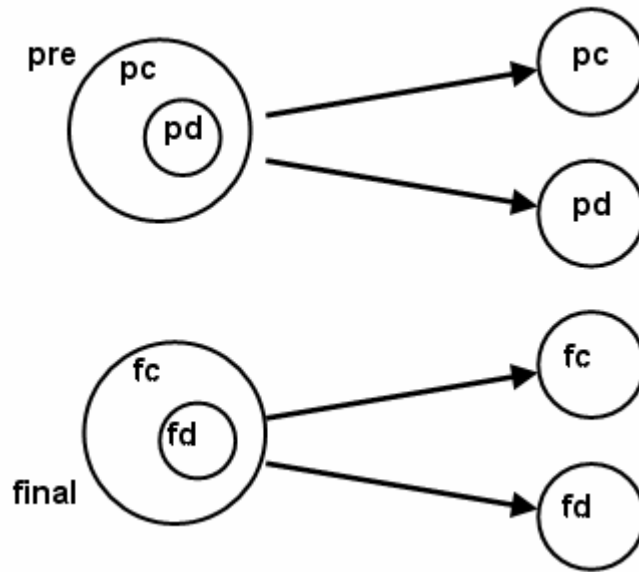


Fig. 6 Conceptual splitting of the pre and final submissions

An unpublished tool called JiscView which makes 'maps' from JISC/Turnitin non-originality reports was adapted to split the reports as shown and four text files were obtained for all 98 reports in the corpus which had a pre-submission and a final report. Fig. 7 illustrates the next stage of the investigation where two intersections between various files can be considered.

The first intersection is between the pre-submission dirty file and the final report dirty file. Any content that appears in both files is material which was marked as dirty in the pre-submission and also appears as dirty in the final report (dirt stay). Likewise material in the pre-submission which did not appear in the final report is dirt which has disappeared (dirt gone) and the last area is new dirt which appears only in the final report.

The second intersection is between the pre-submission clean file and the final report clean file. The three areas would be material which exists in both files (clean stay), material which appears in the pre-submission only (clean gone) and new clean material in the final report (new clean). By examining the overall relationships between the six areas some indication of the patterns of editing which transformed the pre-submissions into the final reports can be gleaned.

Another existing unpublished tool Vertical ALignment Tool (VALT) was adapted to process two files and produce three output files: the material that is common to both files (the intersection) and the remainder of the two input files. Operationally the adapted tool was configured to recognise a minimum length of match of four words (including stop words). This makes the tool operate a little differently from the JISC/Turnitin tool which can flag even single words as matches under some circumstances.

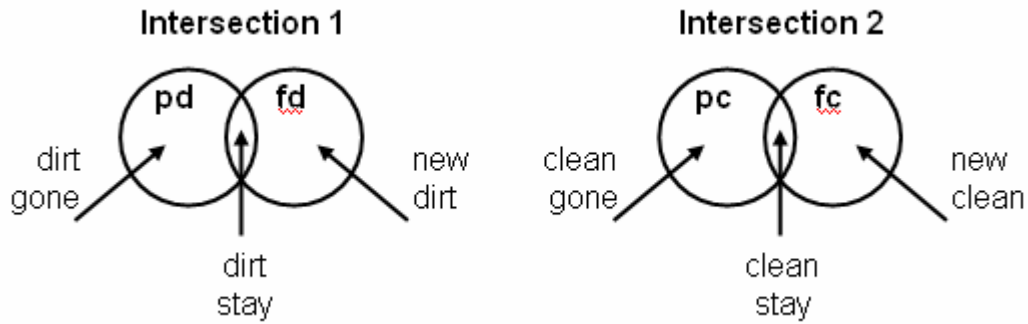


Fig. 7 Clean & dirty, pre-submission and final report intersections

Table 4 shows the number of pre-submissions which were above or below the 15% cut-off level and the number of corresponding final reports which were above the cut-off. This shows that about one third (32/98) of the pre-submissions were above the cut off and about one third of these (10/32) corresponded to final reports which were above the cut off. In contrast only about one twentieth (4/66) of the pre-submissions that were below the cut off resulted in post submissions that were above the cut off. This might seem to suggest that the inclusion of non-original material is a pervasive characteristic.

	final <= 15%	final > 15%	
pre-sub <= 15%	62	4	66
pre-sub > 15%	22	10	32
	84	14	98

Table 4. Pre-submission & final report, above & below 15%

Fig. 8 presents the scattergraph of the amount of non-original content in the pre-submission and the percentage amount of dirt removed as measured in intersection 1. Although the pattern looks a little chaotic there is a significant correlation ($p < .001$) indicating the greater the amount of non-originality the greater the proportion of dirt removed. This would suggest that students throughout the range of behaviours are responding to the information in the non-originality report.

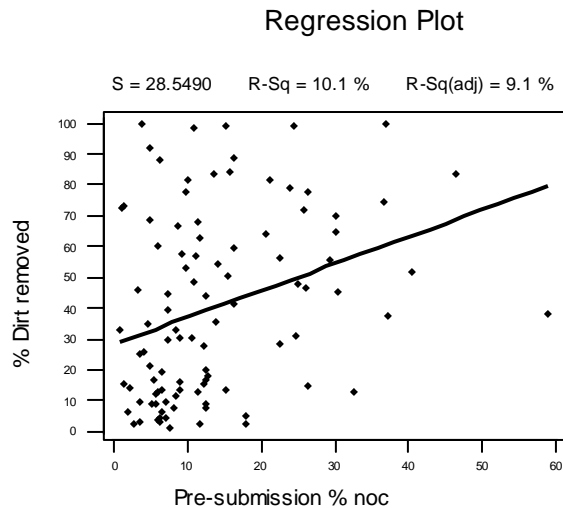


Fig. 8. Scattergraph of %noc against % dirt removed.

Fig. 9 (a) presents the box plots of the dirt gone and clean gone percentages from intersections 1 and 2, and Fig. 9 (b) the box plots of new dirt and new clean. These plots indicate that students remove a smaller percentage of clean material from the pre-submissions and the variation in the percentage of dirty material removed is much greater. However, the median amounts of clean and dirty material added are approximately equal, but again the variability of the amount of dirty material added is greater.

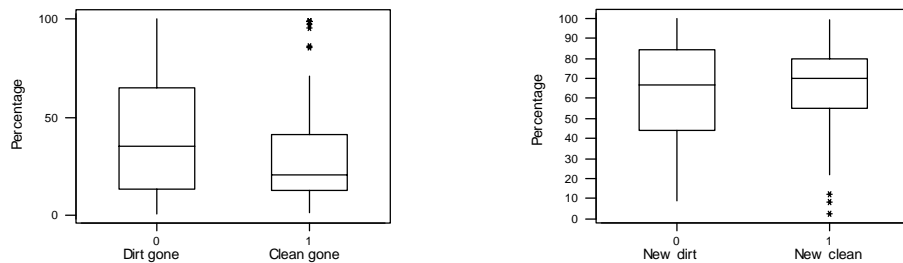


Fig. 9. (a) (left) dirt gone and clean gone box plots, (b) (right) new dirt and new clean.

The overall suggestion from this is that although students are removing the dirty material from the pre-submissions they are then adding more dirty material to produce the final report. This suggestion is reinforced in Fig. 10 which shows a scatter graph of the percentage of dirt in the final submission against the amount of new dirt as a percentage of the overall dirt. The vertical line delineates the 15% cut off with the 14 reports indicated in Table 4 visible to the right of the line. The horizontal line indicates the median percentage of new dirt added. Fig. 9 (b) illustrates the large variation in the percentage of new dirt added to students' final submissions with the median percent added being 67.13. Fig. 10 gives a little more detail showing that for those 14 students who made a final submission containing more than 15% dirty material, 10 had added 80% or more of this dirt since the pre-submission was made. The remaining 84 students showed no discernable pattern in the percentage of new dirt added.

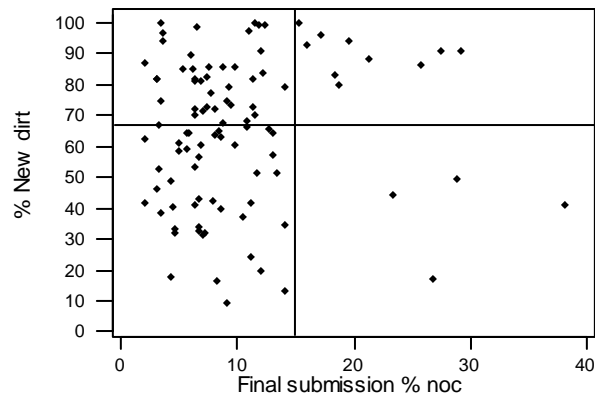


Fig. 10. Scattergraph of final report dirt against dirt added.

Conclusions

Not surprisingly when access to a non-originality detection system is made available to students most, but not all, will make use of it. Those students who do not make use of the system tend to submit shorter projects, with more non-original content and which gain fewer marks. Where a pre-submission and a corresponding final report are available the pre-submission contains significantly more non-original material and this effect is proportionally greater for larger amounts of non originality. In this study no differences were observed between a more technical and less technical programme of study. The analytical investigation suggested that the effect of the pre-submission report was to 'clean up' that part of the dissertation, but new non-original material was introduced as the final version was prepared. Hence the suggestion that the dirt in the pre-submission is diluted by relatively clean material is not supported. Were this to be the case it would be indicative of a change of behaviour after exposure to the initial report.

Accordingly, the indication as to the advisability of allowing students access to a non-originality detection system prior to submission is that; although it might lead to a small reduction in the amount of non-original content in the final report, it does not seem to change behaviour per-se. This finding is in accord with the general best practice guidelines that detection systems are needed to locate those students making extensive illicit reuse of material, but of themselves they are not an answer to the problem of plagiarism (Carroll 2002). However this study might suggest that there is a need for students to be better educated regarding how to make use of non-originality reports.

The measures that were made were of non-original content, both appropriately and illicitly used. This was taken as being a marker of illicit use and the validity of that assumption can be verified from the observation that the greater the amount of non-original content, the greater the proportional amount of material marked as non-original removed from the draft. If this material was being used appropriately then there would be no need to remove

it and the amount of 'clean' and 'dirty' material removed might be approximately equal. That it is being removed and that more 'dirty' material is being removed would seem to indicate it was being inappropriately used.

There are further investigations to be made upon this corpus. The dirty material that has been removed may have simply been dumped. Alternatively it may have been disguised to make it less visible to the non-originality detection system. Of the non-original material that remains it may have been legitimised by being correctly attributed or cited. Preliminary work on addressing these questions suggests that all three behaviours can be detected from automated textual analysis of the submissions.

Academic Integrity Statement

This research was unfunded and this is the first publication reporting the investigations.

References

- Carroll J. (2002). *A Handbook for Deterring Plagiarism in Higher Education*. Oxford Brookes University, ISBN 1-873567-56-0.
- Savage S. (2004). Staff and Student Responses to a Trial of Turnitin Plagiarism Detection Software. *Proceedings of the Australian Universities Quality Forum*.
- Culwin F. (2006). An Active Introduction to Academic Misconduct & the Measured Demographics of Misconduct. *Assessment & Evaluation in Higher Education*. **31**(2), 167-182.
- Culwin F. (2008). A Longitudinal Study of Non-Original Content in Final-Year Computing Undergraduate Projects. *IEEE Transactions on Education*. To appear.
- Lancaster T. & Culwin F. (2001). Plagiarism Issues for Higher Education. *Vine*, **31** (2), 36-41.
- Caruana A., Ramaseshan B. & Ewing M. T. (2000). The effect of anomie on academic dishonesty among university students. *International Journal of Educational Management*. **14** (1), 23 – 30.