

STATISTICAL MEASURES OF INTEGRITY IN ONLINE TESTING: EMPIRICAL STUDY

Tom Wielicki

California State University, Fresno , 5245 North Backer Street, M/S PB7, Fresno, CA 93740-8001, USA

ABSTRACT

This paper reports on longitudinal study regarding integrity of testing in an online format as used by e-learning platforms. Specifically, this study explains whether online testing, which implies an open book format is compromising integrity of assessment by encouraging cheating among students. Statistical experiment designed for this study focused on combining such variables as numerical scores on tests and quizzes with type of feedback received by students during the test and question randomization. Results obtained proved that cheating during well designed online tests is more of a myth than reality.

KEYWORDS

Online testing, integrity of testing, e-learning platform

1. INTRODUCTION

Online delivery of education became a permanent part of educational landscape in spite of many challenges including integrity of online testing. It has been found to be extremely cost effective in delivery of internal corporate training (Zhang, D., Nunamaker, J., 2003) in spite of some evident barriers exhibited especially among SMEs (Anderson, R., Wielicki, T., 2010). The same cannot be said about education – especially higher education, where objectives of instructional activities are broader and more complex than objectives of typical training. Also, universities seem to have more problems with incorporating this new technology into an overall strategy and business processes since – ironically – they are more resistant to change (Jones, N., O'Shea, J 2, 2004).

This may be a reason for apparent differences between number of online credit courses and degree programs offered by lower tier unaccredited institutions and those fully accredited. Accredited degree programs seem to be much more cautious in adopting e-learning format out of concern about quality of education and requirements of accrediting institutions as well as questions about integrity of distance learning. Big part of this skepticism is attributed to legitimate questions about reliability of online testing and assessment, especially at the undergraduate level. Specifically, issue of security or lack of it in a web based testing has been preoccupying researches like Adams and Armstrong (1998) leading to numerous software solutions like their Eval program used for testing at undergraduate level.

However, some studies suggest a negative overall trends in increasing academic dishonesty among new generation of students along with increasing societal permissiveness of our society (Kitahara, R.T., Westfall, F., 2007). Therefore, we should be careful not to attribute problems with integrity of testing solely to an online format of instructional delivery.

Hodgins (2002) in his vision paper developed for the American Society for Training and Development (ASTD) emphasized “Assessment and Certification” as one of the main areas where impact of technology on e-learning has to be closely monitored and controlled. Similarly, Dobbs (2002) in his definition of the state of online learning is concentrating on four fundamental obstacles to high quality of e-learning. Number one problem identified by him is a flawed perception that “reading is learning”. He is suggesting that more interaction should be built into the e-learning as well as effective assessment mechanism.

Assessment seems to be an important part of study in the area of designing and evaluating online learning environment like the one proposed by Hoffman and Ritchie (2001). However, its impact on the quality of educational experience is hardly ever measured and assessed in empirical settings.

A series of quantitative studies based on a solid sample of web based students performance has recently been completed shedding more light on the issues of viability of e-learning (see Wei-Fan Chen, 2005).

At the same time some authors warned against Digital Doctrine that greatly overestimates impact of technology on economy and education (see – Albrecht and Gunn, 2000). Some anticipate that dot-com bust could be repeated with disappointments in the field of e-learning, due to irreplaceability of some important components of face to face learning process. This study attempts to continue a trend of verifying myths created around e-learning with statistically sound samples of data.

2. METHODOLOGY AND HYPOTHESIS

A sample of 230 students took an upper division undergraduate MIS course, which was delivered fully online using Blackboard LMS – a comprehensive e-learning environment. At the same time another 186 students took the same course with the same instructor and using the same text book but in a web enhanced mode. Web enhanced mode is defined here as a paperless class with all materials, handouts and communication delivered in a digitized form (using Blackboard content), with all tests administered online but with students still participating in a traditional lecture in classroom settings.

Couples of hypothesis were formulated addressing different dimensions of quality of assessment process:

- Online open book delivery format of quizzes and tests is conducive to cheating and abuse, therefore test scores will be impacted by the type of assessment feedback
- Online open book delivery format of quizzes and tests is conducive to cheating and abuse, therefore test scores will be impacted by the level of questions randomization used in the assessment
- Combined impact of type of feedback and question randomization will cause significant difference in the mean scores of online tests and quizzes due to cheating

3. EXPERIMENT DESIGN

A sample of 416 students took 12 quizzes and 2 tests during one semester upper division MIS course. This means that total number of graded assignments (quizzes and tests) used in this study is equal to 5824. It has been insured that the level of difficulty was uniform for all students by using the same pool of questions, the same textbook and the same time frame for the assignments. About a half of the sample were web based students (online course), which had almost no face to face contact with the instructor and each other. The other half of the sample included students that participated twice a week in a regular lecture, knew each other and benefited from instructor's face to face consultation hours.

3.1 Variables and Treatments

Blackboard environment provides numerous settings for designing of an online test. Every design could be more or less conducive to cheating, depending on such parameters as:

- time allocated to every question,
- enforcement of sequential way of answering questions (one at a time) versus scrolling page,
- type of provided feedback (just the score, identification of questions missed and the score, identification of question missed and correct answer)
- questions randomization from a larger pool versus the same set of questions presented to every student

We will define every combination of these parameters as a statistical treatment. For purpose of this study only two of those parameters were used to create treatments in statistical analysis of scores:

- type of provided feedback (3 levels) and
- randomization (2 choices)

Those treatments represented arrangements under which cheating during an open book online quiz or test could be either “very easy” or “very difficult” or anywhere in between. A variable that was measured for every treatment was an average score (class mean) on a given test or quiz with specific format (combination of parameters). It was assumed that - should students abuse an online format of testing – the mean of scores should consistently drop as we move from “easy to cheat” treatments to “difficult to cheat” treatment. In other words – if there was any abuse of online testing among students, it was expected that difference between the mean scores will be statistically significant as we compare different combined setups shown in Table 1 below.

Table 1. Combined setups for delivery of online assignments

SCA-NR	show correct answers (SCR); the same set of questions –NR (not randomized)
DRNA_NR	show missed questions but no correct answer (DRNA); the same set of questions - NR
SCO_NR	show only total score; the same set of questions – NR (not randomized)
SCA-R	show correct answers (SCA); randomized questions - R
DRNA-R	show missed questions but not correct answer; randomized questions - R
SCO-R	show only total score; randomized questions - R

It is reasonable to assume that above formats (assessment setups) represent an increasing degree of difficulty in cheating; therefore treatments from the first row to the last may be viewed as a scale of increasing “degree of difficulty in cheating.”

Separate statistical tests were conducted on quiz scores and tests scores due to the difference in settings of the assessment process in both cases. All remaining setup parameters of the assessment process were the same for all collected data: all questions were multiple choice questions, there was always 60 seconds time allocated to every question, and there was always a possibility of answering questions in any order (a scrolling mode enabling student to answer questions in any sequence).

3.2 Statistical Tests

Numerous statistical tests have been conducted to verify some of the hypothesis listed above. Primary focus of this analysis was on the issue of searching for statistically significant difference in the mean scores on online assignments administered under different settings, which were more or less conducive to cheating and abuse by the students.

The first test was conducted using One-Way ANOVA F-test for verification of significant difference among the mean scores on assignments administered with different level of feedback (treatments). Null hypothesis H_0 about equal means on scores obtain in assignments delivered with different level of feedback could not be rejected even at $\alpha = .05$ with value of $F=1.77$ and $p\text{-value} = .1759$. Post hoc Tuckey analysis of p -values for pairwise t-tests confirmed this result.

Similarly, One-Way ANOVA F-test was used for verification of significant difference in the mean scores obtained on online assignments administered with different form of randomization (treatments). Surprisingly, mean scores on assignments with and without randomized questions shown even more uniformity. Null hypothesis H_0 about equal means on scores obtain in assignments delivered with and without randomized questions could not be rejected even at $\alpha = .05$ with value of statistics $F=0.60$ and $p\text{-value} = .4406$.

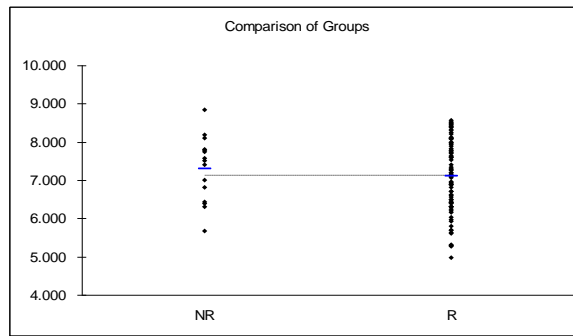


Figure 1. Distribution of means of scores between two types of questions randomization

Lack of impact of questions randomization on the mean score is clearly visible on the Figure 1 above. Post hoc Tuckey analysis of p-values for pairwise t-tests confirmed this result.

The next test utilized Randomized Block Design experiment with blocks identified as two different levels of randomization (R and NR) and treatments as three levels of feedback. Its intention was to remove any variance between investigated means that could be possibly caused by the fact that some assignments used randomized questions and some did not. Again, null hypothesis H_0 about equal means on scores obtain in assignments delivered with combined settings of randomization and feedback could not be rejected even at $\alpha = .05$. Value of $F=0.20$ for treatments (level of feedback) and $F=0.33$ for blocks (randomization) with p-values equal respectively 0.83 and 0.62 would clearly indicate statistically solid uniformity of means. Combined and clearly inconsistent impact of feedback and randomization on the means of scores is shown below in Figure 2.

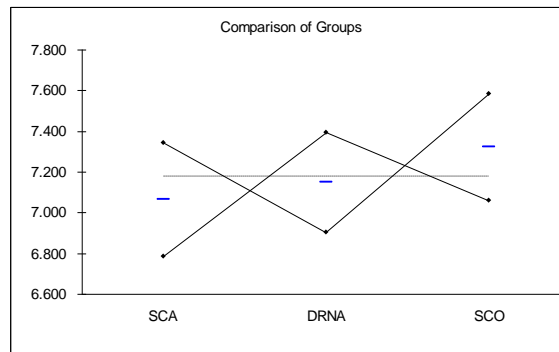


Figure 2. Comparison of impact of feedback and randomization on mean scores

4. SUMMARY OF RESULTS

Preliminary results seem to contradict couple of myths to which academic community often prescribes:

- in general, delivery of quizzes and tests in an online/ open book format does not seem to be conducive to cheating as it does not lead to variations in scores obtained by students under different assessment setups,
- it appears that making answers to questions available to students right after completion of an assessment (treatments SCA) does not have statistically significant impact on average score regardless whether questions were randomized or not,
- randomization of questions when delivering an online quiz or test does not cause statistically significant difference in the means of scores

5. CONCLUSIONS

An overall conclusion should perhaps be formulated in the following way: an average student taking an online class is less mischievous and interested in cheating as he/she is overworked, disconnected and ill organized to be an effective cheater in digital world. Cheating and abusing online testing environment through copying questions, sharing, taking screenshots etc. can be easily made very time consuming and difficult for students by a skillful instructor. Randomization of the questions seems to have a minimal effect on mean scores, whereas revealing answers upon completion of the assignment does not increase possibility of cheating.

REFERENCES

- Adams, J.C., Armstrong, A. 1998. Web-based testing: A study in insecurity. *World Wide Web*, Vol. 1, No. 4, pp 193 – 208
- Albrecht, K., Gunn, R. 2000. Digital Backlash. *Training and Development*. Vol.11, No.3, 2000, pp. 24-34. ASTD Publications
- Anderson, R., Wielicki, T. 2010. Barriers to Application of E-Learning in Training Activities of SMEs. *International Journal on E-Learning*. Vol. 9, No.2, pp.159-167
- Chia-I Chang, F. 2003. Quantitative Analysis of Distance Learning Courseware. *Multimedia Tools and Applications*, Vol. 20, No. 1., pp: 51 - 65
- Dobbs, K. 2000. The State of Online Learning – What the Online World Needs Now: Quality. *Training and Development*, Vol. 9, No.2, pp.78-90
- Hodgins, W. 2002. Learnativity: Into the Future, in *E-learning Handbook*, McGraw Hill, New York, pp.38-64
- Hoffman, B., Ritchie, D. 2001. An instructional design-based approach for developing online learning environment, in Kahn,B. (ed.), *Web-Based Training, Educational Technology Publications*, Englewood Cliffs, NJ, pp.123-156
- Jones, N., O'Shea, J. 2004. Challenging hierarchies: The impact of e-learning. *Higher Education*. Vol. 48, No.3. pp. 379 – 395
- Kitahara, R.T., Westfall, F. 2007. Promoting Academic Integrity in Online Distance Learning Courses. *MERLOT Journal of Online Learning and Teaching* . Vol. 3, pp.34-45
- Stella, A., Gnanam, A. 2004. Quality assurance in distance education: The challenges to be addressed. *Higher Education*, Vol. 47, No.2, pp: 143 – 160
- Wei-Fan Chen. 2005. Effect of web-browsing interfaces in web-based instruction: a quantitative study. *IEEE Transactions on Education*. Vol. 48, No.4; pp. 652- 657
- Zhang, D., Nunamaker, J. 2003. Powering E-Learning In the New Millennium: An Overview of E-Learning and Enabling Technology. *Information Systems Frontiers*. Vol. 2, No. 5, pp: 207 – 218