INVESTIGATING GENDER BIAS IN FORMATIVE AND SUMMATIVE CAA

Karen Fill and Dr. Sally Brailsford

Investigating Gender Bias in Formative and Summative CAA

Karen Fill, Centre for Learning and Teaching
Dr. Sally Brailsford, School of Management
University of Southampton
Highfield, Southampton. SO17 1BJ
k.e.fill@soton.ac.uk

Abstract

Over two academic years, some four hundred first year students of Management at the University of Southampton had the opportunity to self-test using computer-based quizzes before taking a summative, online assessment. Over 97% of the students practised at least once; the average number of attempted quizzes was nine. This paper presents analysis of the accumulated data, with particular attention to possible gender differences in the number of practice quiz attempts, best formative scores and final summative results. The authors conclude that these computer-assisted assessments and the question bank were gender neutral.

Introduction

It has been asserted that the use of computers to support learning in higher education is increasingly gender neutral (Ory et al, 1997, Gunn et al 2002). However, in the specific area of computer-assisted assessment (CAA), the debate and the research seem as yet inconclusive. On one hand it has been suggested that female students may do worse than males because of computer anxiety (Brosnan, 1999; Todman, 2000) and, on the other, that young male students may perform badly because they tend to practise less (McSporran and Young, 2001; Gunn et al, 2002). Some researchers have found that males generally do better in objective tests, including those based on multiple choice questions (MCQ) which are often the mainstay of CAA (Birenbaum, and Feldman,1998; Hopkins, 2003).

This paper describes and discusses the use of formative and summative CAA by first year undergraduates taking an introductory unit in Management. In the analysis of usage and results from two academic years, particular attention is paid to gender differences in the number of formative tests taken, best formative scores and best summative scores.

Background

The 'Introduction to Management' unit is team taught to approximately two hundred students each year. Traditionally, assessment has been by two written essays but, in academic year 2001/2002 a thirty minute computer-assisted assessment replaced the first essay. Formative tests, of the same style and duration, were made available for the students to practise for six weeks before the exam. Use of these tests was optional but all attempts and scores were logged automatically in a database.

Based on the core text, (Bartol and Martin, 1998), a test bank of four hundred questions was developed in WebCT. The questions were predominantly multiple choice, requiring the selection of one correct answer from a list of five, with a few True/False. Three hundred of the questions were available for the students to use formatively, one hundred were reserved for the summative assessment.

About one week of tutor time was involved in development activities. A research assistant spent two weeks setting up the questions and tests in WebCT. There was a minor adjustment to question allocation in the second year of use to ensure that all students received four of the questions that most students had found very difficult in the previous year.

Administration

Students registered on the unit were allocated a WebCT password. They could use the formative tests, known as 'quizzes' as little or as often as they liked. A quiz consisted of twenty-five questions randomly drawn from the database. The students had thirty minutes to answer the questions and submit the test for grading. All grades were recorded for submitted tests, regardless of the time actually taken. Feedback to students was immediate, indicating their grade and the correct answer for each question, as well as more detailed feedback if they had given an incorrect answer.

The computer-assisted summative exam was taken at the end of the first lecturer's six week teaching block. The exam took place in strictly invigilated sessions in university computer rooms. In 2001/02 invigilators had some concerns about security issues during the exam and one session was somewhat disrupted by a network failure. In 2002/03, there were no technical problems but more invigilators were used, to ensure students only accessed the WebCT test and no other websites.

For the exam, each student had a different, random, selection of twenty-five questions to answer in thirty minutes. Students with special needs were allowed an extra ten minutes, in accordance with university regulations. Each question was worth four marks and all were deemed to be of equal difficulty. The computer generated grades were made available to students on-line the same evening. No students failed the exam in 2001/02; eight failed in 2002/03 scoring less than the 40% pass mark. The average grade was 67% in 2001/02 and 63% in 2002/03.

The summative marking load was significantly reduced. In response to a question posed by staff supporting the project, the lecturer wrote: "The time saving in terms of marking was breath-taking. Marking 220 essays normally takes about two person-weeks, but for the computer-assisted assessment marking took about five minutes."

Student feedback

In 2001/02, students were asked to complete an anonymous evaluation form one week after the summative test. This was in addition to the standard unit evaluation form given out at the end of the unit (week 11). Only 56 forms, representing 28% of the cohort, were returned for analysis. Of these, 43 (76%) found the practice tests useful for revising and 54 (96%) felt that CAA was better than writing an essay. Some students voiced concerns about whether all questions were of equal difficulty but this was not raised formally with staff. All respondents found the software easy to use. In 2002/03 the student response (in the end-of-unit evaluation) was slightly less favourable although it was generally positive. Four students expressed the view that giving people different questions was “unfair”.

Research Method

The WebCT database provided access and score data for all students who used the formative and summative CAA. Essay marks and gender were provided by the course administrators. All data was anonymised for analysis. Normal distribution of scores was observed, permitting use of statistical correlations and analyses of variance.

In addition to analysing the overall usage of formative quizzes, results from them and from the summative assessment, comparison was made with the results from essay-only assessment, and particular attention was paid to possible differences due to student gender.

Results

Comparison with the previous non-computerised assessment

In academic year 2000/01, students were assessed by means of two essays. The overall results are shown in Table 1 against those for the two years in which computer-assisted assessment (CAA) replaced one essay.

	2000/01 216 students		2001/02 199 students		2002/03 229 students	
	Essay 1	Essay 2	CAA	Essay	CAA	Essay
Minimum grade (%)	35	25	40	39	24	29
Maximum grade (%)	85	78	96	80	88	80
Mean grade (%)	61	55	67	60	63	54
Standard deviation	10	9	12	7	12	10
Grade correlation	0.32		0.21		0.24	

Table 1. year on year results

Impact of the formative tests (quizzes)

Over the two years, a total of 428 students took the summative CAA (exam). Table 2 summarises formative and summative results for each of the two years grouped by how many times the students used the quizzes.

Group	Year	% of cohort	Mean best quiz score (%)	Mean CAA exam grade (%)
Students who did not take the quiz	2001/02	3	-	61
	2002/03	2	-	66
Students who took the quiz 1-10 times	2001/02	67	72	65
	2002/03	73	69	62
Students who took the quiz 11-20 times	2001/02	19	88	70
	2002/03	21	85	68
Students who took the quiz 21+ times	2001/02	11	97	72
	2002/03	4	87	65

Table 2. summary of quiz and exam scores

In year one (2001/02), only six (3%) of the students who took the exam, had not attempted the quizzes at all. The mean exam grade of this small group was 61%, six percentage points ***lower*** than that of the students who had tried the quizzes at least once. The 193 students who used the quizzes did so an average of ten times.

In year two (2002/03), only five (2%) of the students who took the exam, had not attempted the quizzes at all. The mean exam grade of this small group was 66%, three percentage points ***higher*** than the mean for students who had tried the quizzes at least once. The 224 students who used the quizzes did so an average of eight times.

The correlation between the number of quiz attempts and exam grades for all students was 0.23 in year one and 0.22 in year two. Correlations between best quiz scores and exam grades was 0.34 and 0.48 in years one and two respectively. Analysis of results by the groupings shown in Table 2 revealed

that the already weak correlation between number of quiz attempts and exam grades diminished with more practice: from 0.26 for 1-10 attempts, to 0.08 for 11-20, to -0.03 for 21+ (all students over both years).

Highest and lowest exam scores

Analysis of results by question, revealed no statistical differences between the two years, that is the mean scores for every question were indistinguishable. Over the two years, eighteen students recorded exam scores between 36 and 40, thirteen students recorded exam scores between 88 and 96. Those recording the lowest exam scores only used the quizzes between one and eight times (mean = 4); those recording the highest exam scores practised between two and twenty-two times (mean = 11). The average best quiz score recorded by the low exam grade group was 55; that of the high grade group was 84.

Results by Gender

Over the two years, results are available for 189 female and 237 male students. The overall results for each year are shown in Table 3.

		No. of students	Quiz		Exam
			Mean no. of tries	Mean Best Score (%)	Mean Grade (%)
2001/02	Female	80	10	78	67
	Male	117	9	78	66
2002/03	Female	109	8	73	62
	Male	120	7	74	64

Table 3. overall results by gender

Two-tailed t-tests revealed no statistically significant differences between male and female students with respect to number of quiz attempts (p=0.34), best quiz scores (p=0.59) or exam scores (p=0.49) in either year.

Over both years, 2% of males and 3% of females did not use the formative quizzes at all; 73% of males and 68% of females submitted between one and ten quiz attempts; 19% of males and 22% of females submitted between eleven and twenty attempts; 5% of males and 4% of females submitted between twenty-one and thirty attempts; 1% of males and 4% of females submitted over 31 attempts.

The **highest exam grade** for a female student was 96%; this student had submitted twenty-two quizzes, with a best score also of 96. The highest male exam grade was also 96%; this student had submitted eleven quizzes, with a best score of 84. The **lowest exam grade** for a female student was 28%; the two students scoring this had both submitted two quizzes, with a best score of 48. The lowest male exam grade was 24%; this student had submitted five quizzes, with a best score of 60.

For **all female students** the correlation between the number of quiz attempts and final exam grades was 0.20; the correlation between best quiz scores and exam grades was 0.39. For **all male students** the correlation between the number of quiz attempts and exam grades was 0.27; the correlation between best quiz scores and exam grades was 0.48.

Analysis of results by the groupings in Table 2 revealed the widest deviation from mean exam grade for the fourteen female students with more than 21 quiz attempts than for any other group (mean = 68, sd = 14.8). This group also showed a weak negative correlation (-0.25) between number of formative attempts and final grade.

Although not the focus of this research, it is interesting to note that over the two years there was no significant difference (p=0.42) in the marks for male and female students on the essay assignment later in the unit. However, year by year analysis did reveal a significant difference (p=0.015) in the 2002/03 cohort, with female students achieving a mean of 56.2 and male students 52.8 on the essay assignment.

Discussion

Comparison with the previous year's non-computerised assessment

There was only a very weak correlation between results on the two summative assessments for all students in all three years (Table 1). This may be because they are testing different components of the unit. However, it is noticeable that the maximum grade given for any essay was 85%, whereas, with the objective marking implicit in CAA, a student could theoretically score 100% on the computer-based exam. In 2000/01 only three students (1% of the cohort) achieved 80% or more on the first essay, but in 2001/02, thirty seven students (19% of the cohort) scored 80% or more on the CAA which replaced it. In 2002/03, twenty seven (12% of the cohort) scored 80% or more on the CAA. In the lecturer's view, this is partly due to the inherent reluctance of markers to give very high marks for essay-type questions, but also suggests the need for "tougher" marking schemes for the CAA, possibly including negative marks for certain wrong answers. However, it is unlikely that this would be introduced because a significant aim of the exam, the first degree-level assessment for these students, is to boost their confidence. Indeed, 12% of students achieving 80% is a better profile than 1%, and may suggest that the original essay marking scheme was flawed.

Impact of the formative tests (quizzes)

Sly (1999) reported significantly higher summative results for first year Economics students who opted to do one formative test than those who did not practise and suggested that 'practice tests should be offered to all students' (p. 343). Overall results for our first year Management students appear to endorse the positive effect of practice but, as is clear from Figure 1, this cannot be taken for granted.

In 2002/03, students who practised between one and ten times recorded a poorer mean exam result than those who did not practise at all and lower than the mean for all students in both years.

Generally, the correlation between amount of practice and exam results was very weak. It could be that the distribution of practice tests is bimodal, one peak of weak, anxious students and another peak of hard-working, clever students. Some students would be familiar with the material if they had done A-level Business Studies and so might do very well without any practice at all. Indeed one student who submitted only two quizzes, answering just one question (correctly) each time, scored 92% in the exam. Students may have used the quizzes purely to familiarise themselves with the computer interface or style of question, rather than to test their learning prior to the exam.

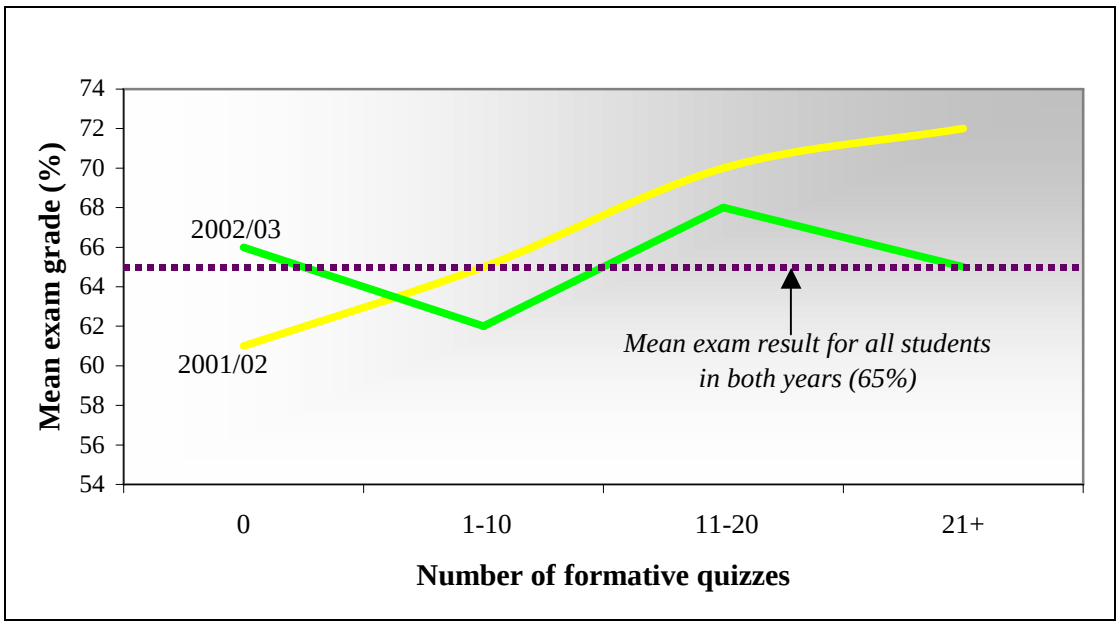

Figure 1. impact of practice on exam grade

Reliable data on language competence was not available for all students, but it is possible that those for whom English is not their first language may have practised more than native speakers, yet still done relatively poorly in the exam. Future research should attempt to explore these and other pre-existing parameters, such as 'A' level or other previous study, language and IT skills, or special needs, more thoroughly.

Gender

There were no statistically significant differences between male and female students with respect to the number of times they practised, nor their actual formative or summative results. This appears to endorse the assertions of Ory et al (1997) and Gunn et al (2002) who contend that the use of CAA in higher education does not disadvantage one or other gender group. There was no evidence of higher levels of anxiety amongst the female students, and male students did not practise less nor achieve better grades.

All students who completed module evaluation forms found the CAA software easy to use and agreed with the statement “computer exams are a good idea, overall.” This appears to support the contention of Struyven et al (2002) that

> Within conventional assessment practices, namely multiple choice and essay typed examinations, students perceive the multiple choice format as more favourable than the constructed response/ essay items, especially students' perceptions on the perceived difficulty, lower anxiety and complexity, and higher success expectancy give preference to this examination format.

Summary of main findings

Full data has been analysed for a total of 426 students over the two years: 189 female and 237 male. Only 3% of students did not use the formative quizzes at all. The average number of quiz attempts was 9. The average summative grade was 65%.

There was only a very weak correlation between the number of quiz attempts and final exam grades for all students. There was a slightly stronger correlation between best quiz scores and exam grades.

There were no statistically significant differences between male and female students with respect to number of quiz attempts, best quiz scores or exam scores in either year.

Students found the software easy to use and the summative marking load was significantly reduced.

Conclusions

Over two academic years, a gratifyingly high proportion of first year Management students took advantage of the facility to self-test as often as they liked before sitting a summative computer assisted assessment. There were no statistically significant differences attributable to gender in use of the formative tests, results for them or for the summative CAA. We conclude that these computer-based tests, and the question bank they draw on, are gender neutral and their use does not disadvantage either male or female students.

Perhaps surprisingly, no striking correlations were found between formative use (number of practice attempts or scores) and summative grades. Further research might enable us to establish the factors that influence the number of times a student attempts a practice test. One such factor may be language competence, another may be prior exposure to the subject matter. Detailed investigation into how the formative quizzes are used might lead to specific recommendations to improve the impact of practice on final result.

Postscript

It had been intended to continue and refine this research in academic year 2003/4. Between times, the University migrated to a different virtual learning environment and CAA engine. This in itself was not a barrier to comparison, as the question sets were also transferred, and the same approach was taken to teaching and learning on the unit. However, there was a catastrophic failure at the summative exam stage and, consequently, a full set of data was not available. Interested readers will find details of how the failure was resolved and proposed strategies for risk management in large scale CAA in Harwood (2005).

Acknowledgements

The authors would like to thank Adam Warren, Centre for Learning and Teaching, University of Southampton, for his unstinting help with WebCT, firstly during the development and use of the quizzes and exam, and secondly with access to the database for analysis purposes.

References

Bartol, K. M. and Martin, D. C. (1998) *Management*, McGraw-Hill, London.

Birenbaum, M., and Feldman, R. A. (1998) 'Relationships between learning patterns and attitudes towards two assessment formats', *Educational Research,* Vol 40 No 1, pp. 90-97.

Brosnan, M. (1999) 'Computer Anxiety in Students' in Brown, S., Bull, J and Race, P. (eds.) *Computer-Assisted Assessment in Higher Education*, Kogan Page, London.

Gunn, C., French, S., McLeod, H., McSporran, M. and Conole, G. (2002) 'Gender issues in computer-supported learning', *Association for Learning Technology Journal*, Vol 10 No 1, pp. 32-44.

Harwood, I.A. (2005) 'When summative computer-aided assessments go wrong: disaster recovery after a major failure', *British Journal of Educational Technology*, Vol. 36, Special issue on Thwarted Innovation in e-Learning, July 2005.

Hopkins, S. (2003) 'Assessment modes in first year macroeconomics: gender differences in performance', *Economic Society of Australia,* [online], http://www.findarticles.com/p/articles/mi_m0PAO/is_2_22/ai_106142468/print (last accessed May 2005).

McSporran, M. and Young, S. (2001) 'Does gender matter in online learning?', *Association for Learning Technology Journal*, Vol 9 No 2, pp. 3-15.

Ory, J C, Bullock, C, and Burnaska, K. (1997) 'Gender similarity in the use of and attitudes about ALN in a university setting', *Journal of Asynchronous Learning Networks,* Vol 1 No 1, [online], http://www.aln.org/publications/jaln/v1n1/v1n1_ory.asp (last accessed May 2005).

Sly, L. (1999), 'Practice tests as Formative Assessment Improve Student performance on Computer-managed Learning Assessments', *Assessment & Evaluation in Higher Education*, Vol 24 No 3, pp. 339-343.

Struyven, K., Dochy, F. and Janssens, S. (2002) 'Students' perceptions about assessment in higher education: a review', [online] http://www.leeds.ac.uk/educol/documents/00002255.htm (last accessed May 2005).

Todman, J. (2000). 'Gender differences in computer anxiety among university entrant since 1992', *Computers & Education*, Vol 34 No 1, pp. 27-35.