DOES REGULAR ONLINE TESTING ENHANCE STUDENT LEARNING? EVIDENCE FROM A LARGE FIRST-YEAR QUANTITATIVE METHODS COURSE

Judith Watson

School of Economics

University of New South Wales

J.Watson@unsw.edu.au

Simon D Angus

School of Economics

University of New South Wales

(currently Department of Economics, Monash University)

simon.angus@buseco.monash.edu.au

ABSTRACT

In recent years, online tools have increasingly contributed to both course delivery and assessment across various domains with mixed quantitative results. We examine the introduction of a regular online assessment tool ('e-quiz') delivered by the WebCT Vista content management system to a large first-year quantitative methods course at UNSW, Sydney (Australia). Unique aspects of this study include the quantitative (mathematical) nature of the online learning, the larger than usual sample size (n > 1500 *in toto*), the inclusion of several difficult-to-obtain explanatory variables, and the high diversity in sub-population attributes. We principally measure the effectiveness of the e-quiz tool from the perspective of student performance in an end of semester closed-book examination. We also consider student opinions and attitudes to the introduction of the online tool. Data taken on attendance at voluntary peer-assisted learning classes, used as a proxy for student effort, along with pre-enrolment mathematical aptitude data are used to identify the specific contribution of the e-quiz tool to student learning. We find support for the hypothesis that regular usage of online learning tools (not grade) significantly and positively contributes to student performance. Furthermore, we find that the hypothesis that this improvement is due mainly to improving student feedback is supported by an analysis of student surveys. Finally, in an extension, we consider the use of such low-cost, online quizzes as a possible identifier for 'at-risk' students.

INTRODUCTION

> *Assessment is the most powerful lever teachers have to influence the way students respond to courses and behave as learners.* (Gibbs, 1999, p. 41)

Educational practitioners live in an exciting age, where information – their basic material of trade – is now available to both themselves and students alike in astonishing quantities at the click of a button. In recent times, the information technology (IT) revolution has provided new tools for all stages of the learning process, from instruction, to formative and summative assessment alike. Mathematical instruction and assessment has received a smaller amount of attention in the literature (reviewed below) due most probably to difficulties in implementing an online system that can handle mathematical formulae, receive the syntax of numerical answers and 'mark' such answers in a way that recognises rounding and alternate representations (e.g. decimal vs. percentage).

We report here on the implementation of a regular, online assessment tool for students studying a broad first-year tertiary mathematics course. Several important features distinguish the present results from other studies of e-learning. First, we have a very large data-set to work with, with a combined observation count in excess of 1600, approximately an order of magnitude greater than equivalent studies. Second, for robustness we consider two distinct sub-populations, both studying exactly the same course, with the same online learning treatment, but constituted of vastly different characteristics such as mathematical ability and cultural background. Third, and importantly, since this aspect is most often lacking in the literature, we are able to control for a range of important explanatory factors such as prior mathematical ability and in-course aptitude. Taken together, this study aims to provide a sober assessment of what can be achieved with the new online learning technologies whilst also discussing just how costly, and/or radical such assessments are to implement.

The rest of this paper is organised as follows. First we establish that online tools have a receptive student audience by reviewing the literature on student attitudes towards information communication technology. Second, we review related work in online learning. Following this introductory section, we introduce the background, methodology and results of the present work before finishing by discussing these results and wider issues that one should consider in implementing an equivalent online assessment.

Digital Natives? Do students want to go online?

At least in the Australian and similar contexts, students entering the higher education system are clearly digitally-aware, and most probably digitally-active. A broad-ranging study by Kennedy et al. (2008) of 1973 students across all faculties at the University of Melbourne found that students were heavily engaged in the new information technology age. A clear majority of students (73%) had unrestricted broadband internet access, whilst practically all students indicated access to a variety of digital hardware such as mobile phones (96%), desktop computers (90%), digital cameras (76%), MP3 players (69%) and laptop computers (63%). Indeed, the authors note that, 'students were overwhelmingly positive about the use of ICT [Information Communication Technology] to support their studies.' (p.3) They found that key activities students wanted to use to pursue their studies included some predictable

activities such as, 'using a computer for general study' (94%), 'searching for information' (93%), or 'general course administration' (84%), and other less expected uses such as 'communicating via SMS' (84%) and 'instant messaging' (75%). Even the common 'Learning Management System' received 81% support in the survey.

These data match with those reported in the US context (Young, 2004) where a survey of 4,374 freshmen and seniors across 13 colleges (of various disciplines) found 74.4% of students wanted either 'entirely', 'extensive', or 'moderate' use of IT in teaching, with only 2.9% saying they preferred 'no IT at all'. Furthermore, 76.1% of students recorded having a 'positive' or 'very positive' experience with course-management systems, as opposed to only 6.9% of students recording a 'negative' or 'very negative' experience. Notably, these positive attitudes arose in courses using more than just course-outline dissemination tools online, with 70% of students who had used an online course management system also saying it featured online quizzes.

A finer detailed look at online tool usage is given by Hoskins (2005) who considered students taking a biological psychology unit. They found that those students who were more frequent users of online features such as bulletin boards either had a higher 'achievement orientation' or were male. However, in a larger study of web usage, Hargittai and Shafer (2006) reported no significant differences between men and women in their online abilities, although they do find women have a lower *self-assessment* of their abilities.

So one might conclude that students are ready and willing to engage with online learning, and in those courses where they have already been exposed to such tools, the experience has been overwhelmingly positive. On the surface, this should provide excellent reasons to pursue an online course facilitation agenda. However, one final statistic from Young's report is telling – where students were asked to list the 'greatest benefit' of classroom technology. To this question, only 12.7% responded that 'improving learning' was the greatest benefit, instead, 48% cited 'convenience' as the greatest benefit. Clearly, at least in the minds of the students, the technology-driven classroom may be cost-effective, efficient, and convenient, but not necessarily actually helpful in terms of student learning. For these reasons, we might urge a closer inspection of online tools to assess accurately their benefit to students. As will be reported below, we believe that it is possible to create the right online tools such that students not only enhance their mastery of the subject, but also prefer the experience.

Online assessment, does it help?

A number of studies have attempted to investigate this question. Of course, within the simple category 'online assessment' there is a broad church of approaches. Whilst not the purpose of this paper, the reader is referred to a review article by Allen (2003) which discusses a range of quiz-making software, mathematical typography converters, and comments on course management tools. As might be expected, there are strengths and weaknesses to each. Inherent in most tools is the common trade-off between flexibility and ease of use. Software optimized for mathematical notation and assessment (e.g. AIMS, covered below) has excellent typography and automated feedback for mathematical contexts, but requires learning at least two different (albeit

simple) software languages to get going. Others, such as WebCT (the focus of the present study) and Blackboard, rely on a very simple point-and-click interface, but are difficult to apply to the more notation-intensive mathematical environment.

In general, Allen notes that 'to keep the students on task, many more quizzes should be given each semester [than a single exam]' (p.274). Here, Allen is alluding to the importance of *formative assessment*, which the small, online 'quiz' environment is well suited. This kind of assessment emphasises feedback given to the student, as opposed to *summative* assessment which emphasises the attainment of a grade. Indeed, Yorke (2001) argues for the rediscovery of formative assessment, especially in the 'first year of a higher education program' (p.115).

In reality, online quizzes are usually implemented so as to represent a combination of formative and summative assessment. A small number of marks are often attached to the assessment to increase student engagement, but the assessment is structured so as to provide methodological feedback to the student. In this domain, the results in the field have been positive. For example, in a study of educational psychology students using the INQSIT software for online, multiple-choice quizzes, Cassady et al. (2001) find end of semester examination performance improves for students labelled as either 'moderate' or 'heavy' users. Importantly, they find that the regularity of the assessment do not adversely affect student anxiety, emotionality or study behaviours, and in fact, there was an advantage in the domain of 'perceived threat' imposed by the impending final examination. Again, students responded positively to this kind of testing, with only six out of 64 students disagreeing with the statement, 'I found the online quizzes to be helpful in preparation for the exam' (p.7). Similarly, Zappe et al. (2002) report on the use of 'TigerNet', an online performance, feedback and student tracking system for high school students. They find that motivated students sought more feedback through the system, and thus performed better in their assessments.

These findings are echoed by others using the more powerful ALEKS mathematical feedback system. This system is designed specifically to use formative assessment to move students from introductory to mastery skills across instructor-specified domains. It 'intelligently' adapts the question difficulty in a topic area to the particular student's abilities as demonstrated in previous ALEKS interaction sessions. The studies of Stillson (2003) and Hagerty et al. (2005) add significant weight to the benefits of formative online quizzes used as a parallel component of a traditional instruction course. In the first case, Stillson studies a basic algebra course taught across three sections with the same instructor and finds that final examination grades are highly correlated with ALEKS achievement scores. Furthermore, it is found that the higher ALEKS scores are highly correlated with time spent on the ALEKS system. Similarly, Hagerty et al. consider an introductory algebra class and compare pre- and post- semester summative assessment. The study finds that students using ALEKS outperformed others (across 4 sections, n = 119) by 8% on average (significant at $p < 0.001$ level). However, although these results are very encouraging, a note of caution is needed it seems, with respect to the ALEKS system, with both authors finding that students' attitudes towards the system were not uniformly positive. In the case of Stillson, a very high drop out rate (approx. 50% more than usual) was recorded for the semester that ALEKS was introduced with some students citing difficulty in learning/adapting to the ALEKS system as reason for their

decision. Likewise, in a survey of student attitudes that asked whether students felt ALEKS should be continued to be used in the course, Hagerty et al. found that only 24 of 53 students agreed or strongly agreed, whereas 17 students were neutral and 12 disagreed or strongly disagreed with its further use.

Other online quiz systems that are reported, but not with quantitative analysis include FCAT (Florida Comprehensive Assessment Test) Explorer (Martindale et al., 2005), Blackboard (Groen, 2006), simple HTML systems such as found in Sanchis (2001), or the Alice Interactive Mathematics (AIM) integrated system of Sangwin (2003, 2004).
In the domain of *summative* assessment, Engelbrecht and Harding (2004) find that online assessment does not differ significantly from that of paper-based assessment. This result is of particular importance, since it was conducted in the context of a calculus course, and using the WebCT online course management system, suggesting that despite the concerns of Allen (2003) mentioned above, the online assessment feature of WebCT can be used to effectively assess students in an analogous way to that of the traditional paper tests.

To sum up this brief survey, it would appear that online assessment is a reasonable proxy for paper assessment, and that where it is used formatively, it can significantly enhance student performance. However, few of these studies tackle analysis of the online system with a rigorous methodology that controls for previous aptitude, subject specific ability, and demographic features. Furthermore, whilst some authors such as Cassady et al. (2001) find strong support from students for such formative assessment methods in helping their examination performance, this seems to be system specific. The difficulties cited by students of the ALEKS technology in Stillson (2003) and Hagerty et al. (2005) need to be borne in mind.

In this paper, we analyse the introduction of a joint formative and summative online assessment procedure, assessing outcomes of both student perceptions and student performance. We begin by describing the methodological set-up and course design followed by a presentation of results, finishing with a discussion of these and some important considerations in going 'online'.

BACKGROUND

Quantitative Methods A (QMA) is a first year core applied mathematics course in the Australian School of Business at the University of New South Wales, until recently know as the Faculty of Commerce and Economics. QMA introduces students to topics such as financial maths, linear algebra (matrices), linear programming and optimization, and calculus with up to several variables. The course had changed little over many years and was in need of review. Results from CATEI, the university's Course and Teaching Evaluation and Improvement Process, in 2004 showed that students were not particularly satisfied with the course and were critical of the amount of feedback they received on their progress. A review process was begun in 2005 and led to a significant re-design of the course by the authors. The changes were implemented in 2006 in Session 2 where there is a smaller cohort than Session 1. Changes were made in number of areas including content, lecture presentation, web

design and tutorial materials. In this paper we focus on a key aspect of change which was designed to promote changes in student engagement: the assessment.

A comparison of the assessment under the old and new structures is shown in Table **1** below. The QMA assessment is now intended to encourage regular engagement by students throughout the session and to give them ongoing feedback. Assessment tasks are scheduled approximately every two weeks and consist of four e-quizzes, a group assignment submitted in two parts and a mid-term multiple choice exam as well as a final written exam.

Table 1: QMA Assessment Structure

Old structure	% of total	New structure	
2 x 15 minute written tutorial quizzes, weeks 6 and 12	10%	4 x 1 hour e-quizzes, weeks 4,6,10 and 14	8%
Mid-term multiple choice exam, week 8	20%	Mid-term multiple choice exam, week 8	20%
Computer labs - attendance only	5%	Group computing assignment, Part A week 7, Part B week 12	12%
Final Written Exam	65%	Final Written Exam	60%

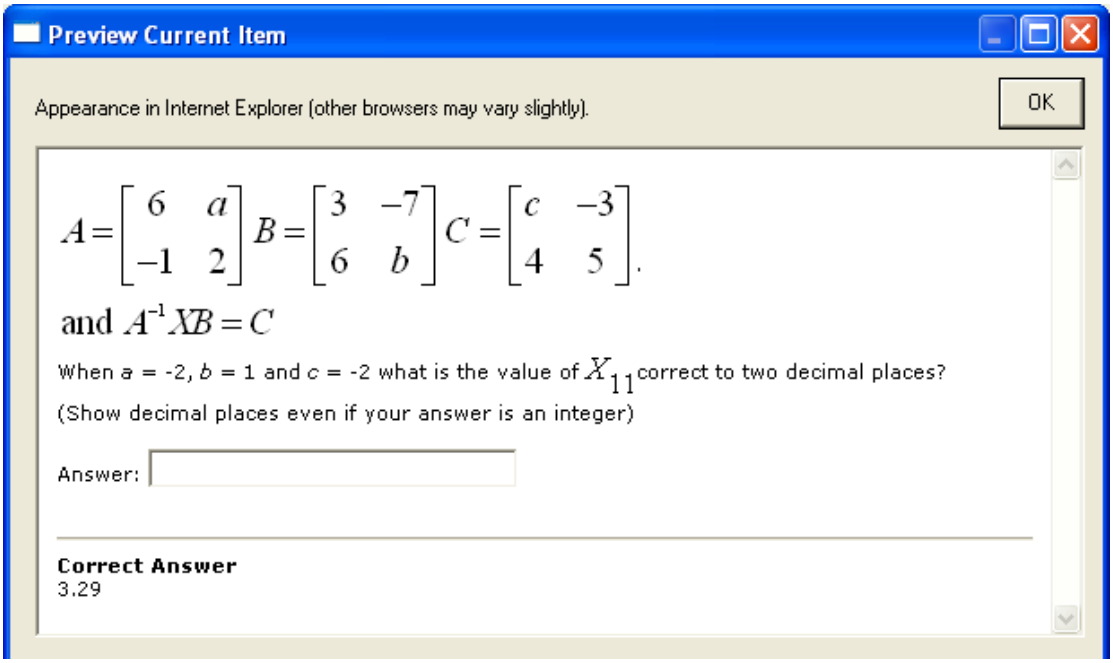
The e-quizzes are designed to be formative as well as summative tools. They each consist of eight to ten questions delivered on WebCT Vista. Students are allowed two one-hour attempts over a one-week period so they are encouraged to revise and learn from their mistakes. Because the e-quiz is primarily a tool to encourage participation in the learning process, but also because of the possibility of cheating on online assessment, only two marks are allocated per quiz. Students can undertake the online quiz on campus or at home and as the quizzes are set to finish at midnight on Sunday it is not surprising that many attempts are made on Sunday nights.

With sizeable cohorts of students it would not be possible to create a large enough bank of multiple choice questions so multiple choice questions with feedback are used only for the practice quiz which is made available on WebCT before each e-quiz begins.

The e-quizzes are written with the Respondus program using calculated questions. Such questions include a number of variables for which up to 80 sets of random numbers within a specified range can be automatically generated. Thus in Session 2 2006 each student was given the same set of questions but with different numerical values and was asked to calculate and enter an answer for each question. Some modifications were made the following session to include a larger question bank which provided greater randomisation of questions. Only the correct answer is given as feedback so students must re-work the question to discover where errors have been made. A second e-quiz attempt, which usually has different numbers in the questions, can be made after a break of at least one hour.

Respondus has a number of inbuilt functions such as logarithms and exponentials which enable certain types of calculated questions to be written easily. It does lack functions for other topics such as derivatives and matrix algebra, but questions still can be constructed if some lateral thinking is employed. An example of a completed matrix question is shown in Figure **1** below.

Figure 1: Completed Respondus Question



The screen showing the question's construction and answer formula using variables can be seen in Figure **2** and a sample of some of the answer sets is in Figure **3**. The question tests the ability to solve the matrix equation for X and requires calculation of a 2 x 2 inverse and multiplication techniques. Because Respondus does not recognise matrix formatting or have matrix calculations available, the question asks for the value of just one element in the matrix product.

Figure 2: Respondus Interface Showing Setup Using Variables

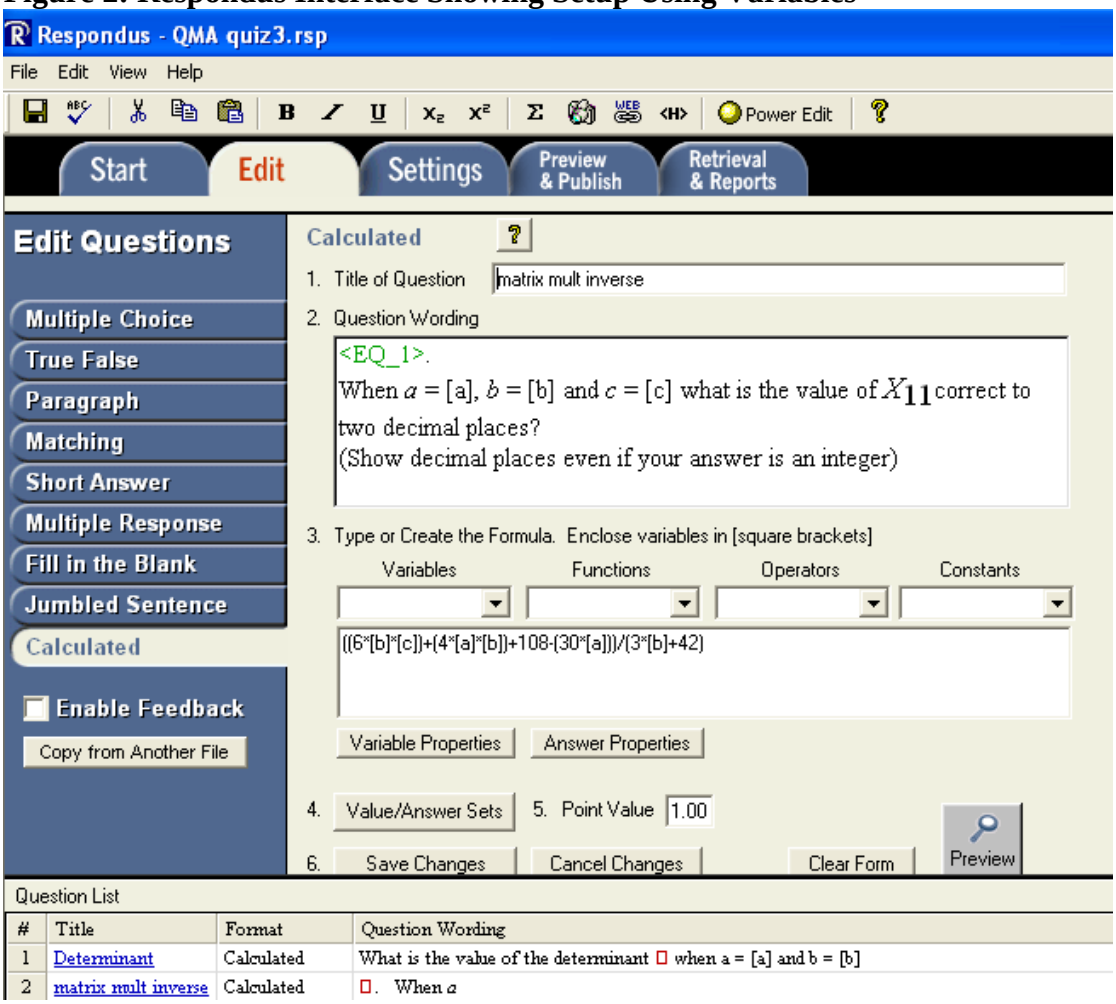
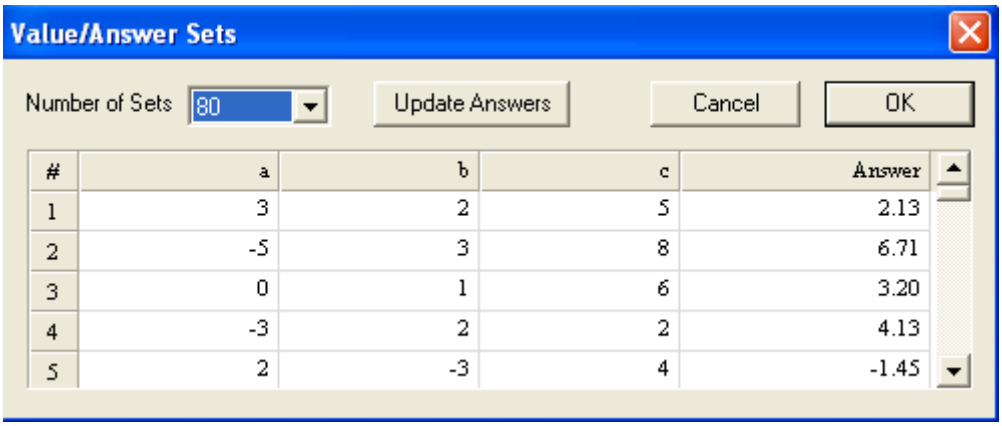


Figure 3: Examples of Answer Sets



#	a	b	c	Answer
1	3	2	5	2.13
2	-5	3	8	6.71
3	0	1	6	3.20
4	-3	2	2	4.13
5	2	-3	4	-1.45

The quizzes have run with very few problems. There were initially some issues relating to the number of significant figures required in answers as these are not clearly documented in Respondus. In 2006 we found that as the session progressed some students were adopting a strategic approach of logging on to the e-quiz briefly

and downloading the questions and answers before making a proper attempt later. In 2007 further questions were included so that there was only a small probability that a student would receive exactly the same question set (with different numbers) on the second attempt. This had the desired effect on gaming behaviour.

AIMS OF THE QUANTITATIVE ANALYSIS

This study aims to assess the extent to which the online quizzes have resulted in improved learning by requiring more consistent work. It also looks at student perceptions of the feedback received. The related issue of the overall satisfaction with the course is also examined. Finally it seeks to analyse whether the e-quizzes might be useful as an early diagnostic to target students in difficulty.

Analysis is in two parts. The first looks at course evaluations and this is followed by econometric analysis of a large body of data.

EVIDENCE FROM COURSE EVALUATIONS

CATEI student evaluation results for comparable sessions before and after the changes were implemented have been examined. Chi square tests of independence were carried out for questions relating to feedback and overall satisfaction with the course. As the distribution of students varies considerably between sessions, two separate comparisons were performed.

Table **2** below shows that the proportion of students who agree with the question is markedly higher post change. The hypothesis that response is independent of year can be rejected at the 1% level for the Session 1 comparison. Session 1 has a large cohort of approximately 1400 students, most of whom have not studied QMA previously. The result is not as strong for the Session 2 data but this can be explained by the fact that a large proportion of Session 2 2006 students had failed and were repeating the course so their responses would reflect their previous session's (pre-change) experience to some extent.

Table 2: Pre- and Post-Change Evaluation of Feedback

Question: I was given helpful feedback on how I was going in the course							
Session	Strongly Agree	Agree	Disagree	Strongly Disagree	n	Chi-square	p-value
1, 2006 (pre)	33	216	125	25	399		
1, 2007 (post)	124	252	94	25	495	**50.1724**	**7.34E-11**
2, 2005 (pre)	4	51	38	9	102		
2, 2006 (post)	17	92	35	8	152	**10.55132**	**0.014417**

Table 3: Pre- and Post-Change Evaluation of Satisfaction

Question: Overall, I was satisfied with the quality of this course							
Session	Strongly Agree	Agree	Disagree	Strongly Disagree	n	Chi-square	p-value
1, 2006 (pre)	42	254	75	29	400		
1, 2007 (post)	150	315	30	5	500	**93.56039**	**3.76E-20**
2, 2005 (pre)	3	61	26	9	99		
2, 2006 (post)	22	95	27	10	154	**10.45954**	**0.015038**

As Table **3** shows, the results for the question on satisfaction show a similar pattern to those for the feedback question except that the chi square value for the Session 1 test is even higher.

The final result from CATEI presented is a response to the question asked in Session 1, 2007 only: "The online e-quizzes were a useful tool to help me study consistently throughout the course". Students clearly thought that the e-quizzes had been useful as 51% strongly agreed and a further 41% agreed.

REGRESSION DATA

Data were collected for QMA students enrolled in Session 2 2006 (Sample 1) and Session 1 2007 (Sample 2). As explained earlier these cohorts have quite distinct characteristics with Sample 1 containing a large proportion of repeat students and new international students. The majority of Sample 2 students are in their first session of university after leaving high school.

Records of all those who did not attempt the final examination or who were granted a supplementary exam were removed. This resulted in a Sample 1 size of 397 observations and 1273 observations in Sample 2.

A Peer Assisted Support Scheme (PASS) operates in QMA. Students are able to attend peer led study groups on a drop-in basis. Previous study has linked higher QMA results with increased PASS attendance (Watson, 2000) so a PASS attendance variable was included as a proxy for effort. For the PASSB dummy variable, attending PASS more than twice per session was scored as 1 and fewer or no attendances were coded as zero. Missing data for some weeks in 2007 precluded having a further variable for high PASS attendance.

Demographic data and information on the level of mathematics taken at high school were obtained from university records. Prior mathematics background was considered

likely to be an important factor in explaining performance in this course. Although students have studied under a variety of education systems, a large number who had taken HSC mathematics in New South Wales could be directly ranked according to the levels they had studied. The middle level i.e. Mathematics plus Extension 1 was considered the baseline category. General Mathematics or Mathematics alone was categorised as Low, and the Extension 1 plus Extension 2 combination was High. It was not possible to obtain prior mathematical data for those who had studied overseas and no attempt was made to include the disparate levels of interstate students.

REGRESSION RESULTS

Do regular online e-quizzes enhance student performance?

Following on from the above, we now test the hypothesis that the online e-quiz regime effectively contributes to student learning. In particular, in this exercise we are not concerned with predicting a student's final mark per se, rather, we wish to ask whether a student who has been exposed to the e-quiz testing (regardless of performance) will perform better than a colleague who has not had the same exposure. In this way, we seek to provide insight as to whether (as we hypothesise) the act of doing the e-quizzes is a significant contributor to a student's learning, perhaps via a greater sense of feedback and/or motivation for consistent study.

Furthermore, since we are concerned to distinguish the e-quiz signal from other possible predictors of final examination performance, we include in our first model explanatory variables for prior student mathematical ability (encoded as dummies *HIGHM*, or *LOWM* where the student's secondary school maths included higher, or only lower subjects respectively); their in-session mid-term examination mark (encoded *MT*, in %); and whether or not the student attended the additional in-session PASS classes (explained above) (encoded as the dummy variable *PASSB*). We also include a gender dummy (encoded as *GEN*, and taking 1 if female) to account for any gender-based variation in online usage. Finally, we encode our variable of interest as the dummy *QUIZB*, which takes the value 1 if the student attempted each of the four online quizzes, and 0 if the student missed one or more of these quizzes. The dependent variable used is the student's final examination mark (in %), encoded as *FE*. The model is thus as follows:

$$FE_i = \beta_0 + \beta_1 QUIZB_i + \beta_2 MT_i + \beta_3 PASSB_i + \beta_4 LOWM_i + \beta_5 HIGHM_i + \beta_6 GEN_i + \varepsilon_i \qquad \textbf{(1)}$$

where ε_i is the error term. However, since the LHS variable is bounded on the interval [0,1], we instead perform the Logistic function transformation,

$$L_i = \ln[FE_i / (1 - Fe_i)] \ , \qquad \textbf{(2)}$$

such that the error terms will be unbounded[1].

[1] For numerical reasons, we actually use the standard $L_i = \ln[(FE_i + 0.5) / (1 - Fe_i + 0.5)]$ formulation.

Summary statistics of all continuous and dummy variables are given in Tables **4** and **5** (at the end of the paper) respectively. As these data indicate, the two samples differ markedly in some important features. As would be expected, the first sample, representing the off-session course delivery, is a much smaller sample by approximately a factor of 1/3. Furthermore, due to the nature of the off-session course timing, a significant number of students in this sample are either repeating the subject, or are fresh students from the mid-year intake. Of the latter, these students are predominantly international (non-local) in origin. These factors are clearly evident in the average values of the two prior-mathematics ability dummies, *HIGHM* and *LOWM*, where Sample 2 has approximately a uniform distribution between the low, medium and high maths attainment, whilst Sample 1 shows an approximately direct transfer from the high regime to the low regime of around 14%. The international dummy also represents these demographic changes with Sample 2 being made up of approximately 20% more international students than the ratio found in Sample 1. Further, since many international students are female, Sample 1 has an elevated female make-up in comparison with Sample 2. Taken together, these data indicate that by demographics, the two samples are extremely diverse and thus form a robust basis for hypothesis testing. The summary statistics are discussed further below.

Due to the fact that prior maths attainment data could only be reliably obtained for those who had completed high school in NSW, model **(1)** was first estimated just on this subpopulation of each sample (the left-hand column in each table). As can be seen in Tables **6** and **7** below, the quiz-exposure dummy was found to be positive and significant at $p < 0.01$ in both samples. Indeed, one can quickly calculate that for the average student in Sample 1 or 2 (mean *FE* mark of 0.55 and 0.60 respectively), by attempting all four quizzes, the model predicts an increase in their final examination mark of approximately 11% or 10% respectively relative to a colleague who attempted three or less quizzes. Other coefficients in this model move in an understandable and consistent direction across the samples, with higher attainment in either the mid-term or in high school mathematics contributing significantly to a higher expected final examination mark. On the other hand, having lower maths attainment was found to be a significant negative factor, as might be expected. The two other explanatory variables which relate to the additional PASS study group scheme, and a student's gender produced mixed results. PASS appears to be weakly positive in Sample 1, or insignificantly negative in Sample 2. Similarly, females appear to fare better than males in both samples, but only in a significant way in Sample 2.

As a further robustness check, we run an adapted model on both samples,

$$FE_{\mathrm{i}} = \beta_0 + \beta_1 QUIZB_{\mathrm{i}} + \beta_2 MT_{\mathrm{i}} + \beta_3 PASSB_{\mathrm{i}} + \beta_4 GEN_{\mathrm{i}} + \beta_5 INTL_{\mathrm{i}} + \varepsilon_{\mathrm{i}} \qquad \textbf{(3)}$$

which is essentially as per model **(1)** but without the prior maths attainment dummies and includes instead, a dummy *INTL* which takes the value 1 if the student was a non-local (international) student. As explained above, model **(3)** can be estimated on the full population of each sample since it does not include the restrictive maths dummies. Again, **(3)** was estimated with the Logit transform as per **(2)** and results are presented in the second (right-hand side) column of Tables **6** and **7** respectively. In

line with the previous estimation, the estimated coefficient for the *QUIZB* dummy again shows a very similar sign, magnitude and level of significance, despite the change in specification and associated increase in number of observations. Furthermore, the other explanatory variables have coefficients very much in line with the previous model, with the mid-term result and PASS coefficients telling a consistent story. The new dummy for non-local students, whilst not affecting the *QUIZB* coefficient of interest in any significant way, does enter significantly and relatively strongly in the Sample 1 population, whereas the Sample 2 population does not yield such a strong effect. One can explain this difference by the demographics of the two populations. Recall that the Sample 1 population is constructed from the off-session teaching of the subject, and hence, local students who take the subject in this session are highly likely to be re-sitting the subject, whereas non-local students comprise the bulk of the new intake. With reference to the small and marginally significant coefficient found in the Sample 2 population, one can conclude that non-local students do not appear to be particularly more likely to have a higher final grade than local students, all else being equal.

To sum up, given the consistency of the sign, size, and significance of the quiz exposure dummy across the two diverse samples under both model specifications, we may conclude that the null hypothesis that full exposure to the e-quizzes does not contribute positively to student performance can be rejected.

Can online e-quizzes serve as a low-cost warning signal?

Given the efficacy of the online e-quizzes in promoting student learning as found above, and due to the low set-up cost requirements for the quiz instrument, we now extend the analysis to ask whether the e-quizzes can be used as a low-cost signal to identify struggling students in the early part of a teaching semester. To test this hypothesis, we construct a purely discrete model that attempts to explain variation in the probability that a student will pass the final exam or not. To this end, we incorporate various independent variables from either pre- or early- stages of the course, that might explain the propensity for a student to pass the final exam. This model is constructed as follows,

$$FEPASS_i = \beta_0 + \beta_1 QABPASS_i + \beta_2 MTPASS_i + \beta_3 LOWM_i + \beta_4 HIGHM_i + \beta_5 GEN_i + \varepsilon_i \qquad \textbf{(4)}$$

where *FEPASS* and *MTPASS* took the value 1 if the student obtained 50% or higher for the final exam and mid-term respectively, and the new variable *QABPASS* took the value 1 if the student obtained 50% or higher for the average of their first two e-quizzes[2]. Additionally, the two prior maths attainment dummies are again included, together with the gender dummy. In this way, the model seeks to use any available early- or pre- semester predictor of a student's final performance such that the true marginal effect of early success in the e-quizzes can be determined.

[2] Where a student attempted a particular quiz twice, the higher of the two grades was taken to calculate the student's average grade for the two early quizzes. This is consistent with the proclaimed policy on the e-quizzes, that the higher grade obtained in each quiz would be used for marking purposes.

The model was estimated as a true Logit regression[3] with coefficients and marginal effects as reported in Table **8** (at the end of the paper). Again, one finds that coefficient signs, magnitudes, and significances are extremely consistent across the two diverse sample groups. In particular, as one would expect, passing the mid-term exam, or having high prior maths attainment both contribute positively and significantly to the probability of passing the final examination. However, low maths attainment contributes significantly but negatively, as is understandable. It is found that the coefficient of the variable of interest, *QAPASS*, enters positively and significantly in both estimations. In particular, a study of the marginal effects[4] indicates that in both samples, early e-quiz performance has one of the highest predictive values for final examination success, on par with the mid-term and prior low maths attainment effects. On the basis of the evidence across the two sample groups, one may conclude that early e-quiz performance could be employed as both a useful and significant signal in determining students who may be at risk of failing an end of semester comprehensive examination.

DISCUSSION

This paper has presented results on student preferences and student outcomes based on data taken from a large first-year, multi-section, mathematical course. A number of ongoing questions in the literature have been addressed with this data-set. Firstly, we have found that the act of *taking part* in the regular online testing, regardless of the mark obtained by the student for these tests, is a positive indicator of final examination performance. We have suggested that the key effects driving this result are first, that the four online quizzes (spread over 14 weeks) increase the propensity for students to work *consistently* throughout the course, and second, that formative student feedback is increased by the combined 'practice' and 'quiz' testing procedure. It would appear that via the online CATEI questionnaire and the Chi-square analysis we can answer in the affirmative for both of these effects. Students are happy to say that the e-quizzes helped them to study 'consistently' through the course whilst cross-year comparisons suggest that students find feedback increased significantly after the introduction of the regular quizzes. It is true that other course structures were changed alongside the e-quiz treatment, however, the four new practice and assessment items (eight in total) added by the e-quiz component must certainly shoulder much of the ‘blame’ for this improvement.

Second, this study has shown that the online testing technique is a valid option for identifying 'struggling' students early in the semester. This opens the way for an intriguing analysis of different intervention styles that could be implemented on this basis. For example, one could easily send a generic email to all students who register a fail on average for their first two attempts, inviting them in particular to take advantage of additional learning support related to the course material. One could

---

[3] The estimation was achieved with SHAZAM's 'LOGIT' procedure which uses the maximum likelihood estimation technique.

[4] For the Logit estimation, the marginal effects give the change in probability of a success in the dependant variable due to changing the binary variable from 0 to 1, given the 'modal' observation characteristics (i.e. the most common student characteristics).

think of a randomised trial that could be run on this basis to assess the effectiveness of this identification/encouragement treatment.

Third, it should be noted that this study gives an indication that the addition of online testing, in this case by the administration of four additional assessments, need not be feared from the perspective of student satisfaction with the course as a whole. This is a particularly pleasing result, since it indicates that good assessment design can defensibly be sold to students on the basis of better performance, whilst at the same time knowing that they will be happier for it.
So one might summarise the results presented in this paper by saying that online mathematical testing appears to be both helpful to student learning, and generally well received. We shall finish by discussing some of the ongoing issues with this kind of assessment.

As has been mentioned in the introduction, the online testing environment is potentially more open to abuse than other forms of assessment. In particular, the authors were concerned that students may take advantage of the 'two attempts' procedure by simply printing-off the quiz in their first 'attempt' (without actually answering any questions), so as to prepare for their second, and hopefully successful attempt. Whilst one might be tempted to think that *any* mechanism which causes a student to do more subject specific study is a good one, we find that in fact, this is not necessarily the case. Although not reported here, a brief study of time-use data for each student produced by WebCT Vista suggested a strong correlation between low (i.e. zero) first-attempt scores followed by high second attempt scores and spending approximately 1 minute on the first attempt, and the full 60 minutes on the second attempt. Furthermore, it was found that students tended to undertake this kind of 'strategic' activity later in the semester.

To investigate this effect, we coded any students with two attempts in e-quiz 3 (C) or e-quiz 4 (D) who obtained an e-quiz mark of 0/10 in the first attempt, and 5/10 or more in their second attempt with a 1 in the variable *STRATC* and *STRATD* respectively. Only students with at least one attempt in the particular quiz were considered. As can be seen in Table 5, 7% and 9% of students were identified in such a way in Sample 1 compared with only 1% and 2% in Sample 2 (significantly less on both counts at the 1% level). The decline in this pattern can be explained by the added randomisation mechanism in the Sample 2 semester such that students couldn't be assured of seeing the same questions (or numbers) in each attempt. Whilst this should offer encouragement to those who would worry about such strategic play in online assessment, is such strategic behaviour actually a bad thing? A model similar to **(1)** was subsequently investigated (not reported in full here) with the addition of the two strategic dummies. Interestingly, both STRATC and STRATD strategic variable coefficients across the two samples had negative signs. However, both samples gave rise to insignificant effects at the 10% level for the fourth e-quiz strategy dummy and *p-values* for the third e-quiz strategy value were close to the 10% level of significance. So we may tentatively conclude that interacting with the online assessment procedure in this non-standard way is not beneficial to student performance, despite appearances. However, further work would need to be done to

clarify this point. In any case, it would seem that this behaviour can be easily eradicated by the larger question bank approach of the second sample.

Finally, whilst concluding that from a teaching and learning perspective, the online testing procedure as implemented through WebCT Vista, is an effective one for student outcomes, it is not without practical problems. For instance, mathematical notation is in the main handled poorly by the system, with inventive usage of pronumerals required to render linear algebra and calculus questions effectively. Whilst apparently not a big problem, this process does increase the cognitive load on a student, who must appropriately fit the pronumeral data to the mathematical formula rendered as an image. Similarly, attention must be paid to both the practice, and announcement, of tolerance settings in assessing student answers. It took some time for the present authors to become accustomed to the 'absolute' and 'percentage' approaches afforded by the Respondus software to mark a student's answer. For financial maths, we found that a tolerance of 5 units of the least significant figure (e.g. if the correct response was 1.234, the system should accept answers between 1.229 and 1.239) was appropriate for most small to medium sized financial calculations where a dollar amount resulted. For percentage or decimal calculations, a tolerance of 2 units of the least significant figure sufficed. An associated problem is that students must not enter anything other than numeric notation, i.e. no '$' or '%' signs. Again, with adequate prior-notice, this problem can be handled.

Notwithstanding these minor quibbles, this paper has found significant support for the use of online testing in the mathematical classroom. Moreover, our experience presented here, suggests that not only do students perform better with the aid of regular online testing, but they seem to also prefer the experience, an outcome that is rarely achieved when it comes to adding to the number of assessments a student must face!

REFERENCES

Allen, G.D. (2003) “A Survey of Online Mathematics Course Basics”, *The College Mathematics Journal,* **34**(4), pp. 270-279

Cassady, J.C. Budenz-Anders, J., Pavlechko, G. and Mock W. (2001) “The Effects of Internet-Based Formative and Summative Assessment on Test Anxiety, Perceptions of Threat, and Achievement”, Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA, April 10-14.

Elzinga, K. G. (2001) “Fifteen Theses on Classroom Teaching”, *Southern Economic Journal*, **68**(2), pp. 249-257.

Engelbrecht, J. and Harding, A. (2004) “Combing Online and Paper Assessment in a Web-based Course in Undergraduate Mathematics”, *Journal of Computers in Mathematics and Science Teaching*, **23**(3), pp. 217-231.

Gibbs,G. (1999) “Using Assessment Strategically to Change the Way Students Learn” Brown, S. and Glasner, A. eds., *Assessment Matters in Higher Education*, Buckingham, S.R.H.E. and Open University Press

Groen, L. (2006) “Enhancing learning and measuring learning outcomes mathematics using online assessment”, UniServe Science Assessment Symposium Proceedings

Hagerty, G. and Smith, S. (2005) “Using the web-based interactive software ALEKS to enhance college algebra”, *Mathematics and Computer Education*, **39**(3), pp. 183-194.

Hargittai, E, Shafer, S, (2006) “Differences in Actual and Perceived Online Skills: The Role of Gender” *Social Science Quarterly,* **87**(2), pp. 432-448

Hoskins, S. L., and van Hoof, J. C. (2005) “Motivation and ability: which students use online learning and what influence does it have on their achievements?”, **36**(2), pp. 177-192.

Kennedy, G., Judd, T., Churchward, A., Gray, K., and Krause, K. (2008) “First year students experiences with technology: Are they really digital natives?” *Australasian Journal of Educational Technology* **24**(1), pp.108-122

Martindale, T., Pearson, C., Curda, L. K. and Pilcher, J. (2005) “Effects of an Online Instructional Application on Reading And Mathematics Standardized Test Scores”, *Journal of. Research on Technology in Education.*, **37**(4), pp.349-360

Sanchis, G. R. (2001) “Using web forms for online assessment”, *Mathematics and Computer Education*, **35**(2), 105-113.

Sangwin, C.J. (2004), “Encouraging higher level mathematical learning using computer aided assessment”. In J. Wang and B. Xu, editors, *Trends and Challenges in Mathematics education*, chapter 20, pages 255–265. East China Normal University Press, 2004.

Sangwin, C.J. (2003) “New opportunities for encouraging higher level mathematical learning by creative use of emerging computer aided assessment”, *International Journal of Mathematical Education in Science and Technology*, **34**(6) pp. 813–829.

Stillson, H. and Alsup, J. (2003) “Smart ALEKS ... or Not? Teaching basic algebra using an online interactive learning system”, *Mathematics and Computer Education*, ***37***(3), pp. 329-340.

Watson, J. (2000) “A Peer Assistance Support Scheme (PASS) for First Year Core Subjects” paper presented at the 4th Pacific Rim First Year in Higher Education Conference, Brisbane, Qld, 5-7 July.

Yorke, M. (2001) “Formative Assessment and its Relevance to Retention”, *Higher Education Research and Development*, **20**(2), pp. 115-126.

Young, J. R. (2004) “Students say technology has little impact on teaching”, *The Chronicle of Higher Education*, August 13.

Zappe, S. M., Sonak, B. C., Hunter, M. W. and Suen, H. K. (2002) “The Effects of a Web-Based information Feedback System on Academic Achievement Motivation and Performance of Junior High School Students”, paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, April 1-5.

TABLES

Table 4: Continuous Variable Summary Statistics

	Sample 1		Sample 2	
	FE	**MT**	**FE**	**MT**
n	397	395	1240	1264
mean	0.55	0.47	0.60	0.71
std	0.19	0.17	0.20	0.16
min	0.00	0.10	0.01	0.15
max	0.98	0.90	1.00	1.00

Table 5: Dummy Variable Summary Statistics

	Sample 1		Sample 2	
	n [a]	$\overline{x}$	**n** [a]	$\overline{x}$
FEPASS	397	0.64	1273	0.69
GEN	397	0.55	1273	0.47
HIGHM	198	0.17	990	0.31
INTL	397	0.35	1273	0.17
LOWM	198	0.45	990	0.31
MTPASS	397	0.45	1273	0.91
PASSB	397	0.14	1273	0.15
QABPASS	368	0.90	1188	0.89
QUIZB	397	0.80	1273	0.83
STRATC	397	0.07	1189	0.01
STRATD	397	0.09	1106	0.02

Notes to table

[a] Sample sizes (*n*) for each variable represent count of non-missing data in each sample. Actual sample size reported in regression results tables below indicates total number of complete observations for the specific regression model being tested.

Table 6: Final Examination (%), Sample 1[a]

	Local Obs	All Obs[b]
QUIZB	**0.48*** **	**0.50*** **
	(4.06)	(4.99, 4.71)
MT	**1.91*** **	**2.63*** **
	(6.27)	(11.1, 10.21)
PASSB	**0.25***	**0.18***
	(1.83)	(1.58, 1.7)
LOWM	**−0.34*** **	**-**
	(-3.3)	
HIGHM	**0.52*** **	**-**
	(3.84)	
GEN	**0.13**	**0.05**
	(1.43)	(0.62, 0.61)
INTL	**-**	**0.35*** **
		(4.13, 3.86)
CONSTANT	**−1.18*** **	**−1.56*** **
	(-6.6)	(-11.4, -10.05)
n	196	395
Adjusted *R-squared*	0.37	0.34
rho	-0.013	-0.025
B-P-R[c] *p-value*	0.10	0.03

Notes to table

[a] Significance levels indicate *p-value* < 0.10 (*****), < 0.05 (******), and < 0.01 (*******) respectively; *t-stats* for each coefficient given in parenthesis, where two regressions were run (see second note), significance levels were drawn from the second (corrected) regression.

[b] In cases where the B-P-R test suggested the presence of heteroskedasicity, a second regression was run, using the heteroskedasticity-consistent covariance matrix approach, with the resultant *t-stats* given in the parenthesise alongside the first estimation for comparison.

[c] B-P-R indicates *p-value* obtained by the Breusch-Pagan-Godfrey test for heteroskedasticity, evaluated on the (first) uncorrected procedure for each model.

Table 7: Final Examination (%), Sample 2[a]

	LOCAL OBS[b]	ALL OBS[b]
QUIZB	**0.44*** **	**0.57*** **
	(6.94, 6.69)	(9.16, 8.49)
MT	**2.76*** **	**3.53*** **
	(17.58, 17.99)	(24.81, 22.97)
PASSB	**-0.03**	**-0.16*** **
	(-0.44, -0.49)	(-2.66, -3.02)
LOWM	**−0.40*** **	**-**
	(-7.09, -7.82)	
HIGHM	**0.43*** **	**-**
	(7.91, 7.55)	
GEN	**0.13*** **	**0.11*** **
	(2.86, 2.84)	(2.58, 2.52)
INTL	**-**	**0.08***
		(1.40, 1.46)
CONSTANT	**−1.91*** **	**−2.58*** **
	(-15.02, -15.85)	(-24.26, -22.45)
n	968	1239
Adjusted *R-squared*	0.51	0.42
rho	0.028	0.045[d]
B-P-R[c] *p-value*	0.00	0.00

Notes to table

[a] Significance levels indicate *p-value* < 0.10 (*****), < 0.05 (******), and < 0.01 (*******) respectively; *t-stats* for each coefficient given in parenthesis, where two regressions were run (see second note), significance levels were drawn from the second (corrected) regression.

[b] In cases where the B-P-R test suggested the presence of heteroskedasicity, a second regression was run, using the heteroskedasticity-consistent covariance matrix approach, with the resultant *t-stats* given in the parenthesise alongside the first estimation for comparison.

[c] B-P-R indicates *p-value* obtained by the Breusch-Pagan-Godfrey test for heteroskedasticity, evaluated on the (first) uncorrected procedure for each model.

[d] Since the DW test indicated the presence of autocorrelation, a regression was run with the standard Cochrane-Orcutt procedure which successfully reduced *rho* to the value of -0.003. However, this procedure had no significant affect on estimated coefficients or *p-values*.

Table 8: Probability of Passing the Final Examination

	SAMPLE 1		SAMPLE 2	
	Coefficient	**Marginal Effect**	**Coefficient**	**Marginal Effect**
QABPASS	**1.23****	**0.30**	**1.38*****	**0.28**
	(2.23)		(5.24)	
MTPASS	**1.13*****	**0.19**	**1.52*****	**0.31**
	(3.10)		(4.95)	
LOWM	**-0.79****	**−0.19**	**-1.71*****	**−0.36**
	(-2.11)		(-8.88)	
HIGHM	**0.77***	**0.14**	**0.74*****	**0.08**
	(1.37)		(2.97)	
GEN	**0.65****	**0.16**	**0.35****	**0.04**
	(1.93)		(2)	
CONSTANT	**-1.13****		**-1.31*****	
	(-1.88)		(-3.28)	
n	182		927	
Iterations to	4		4	
Correct predictions (%)	119 (65)		728 (79)	
Likelihood ratio test				
test stat (*dof*)	32.50 (5)		266.75 (5)	
p-value	0.000		0.000	