British Journal of Educational Technology *Vol 40 No 2 2009* 255–272
doi:10.1111/j.1467-8535.2008.00916.x

Does regular online testing enhance student learning in the numerical sciences? Robust evidence from a large data set

Simon D. Angus and Judith Watson

Simon D. Angus is a lecturer in the Department of Economics, Monash University, Melbourne, Australia; Judith Watson is a lecturer in the School of Economics, University of New South Wales, Sydney, Australia. Address for correspondence: Dr Simon D. Angus, Department of Economics, Monash University, Clayton, 3206 VIC, Australia. Email: simon.angus@buseco.monash.edu.au

Abstract
While a number of studies have been conducted on the impact of online assessment and teaching methods on student learning, the field does not seem settled around the promised benefits of such approaches. It is argued that the reason for this state of affairs is that few studies have been able to control for a number of confounding factors in student performance. We report on the introduction of a regular (every 3 weeks) low-mark online assessment tool in a large, first-year business mathematics course at the University of New South Wales, a major Australian university. Using a retrospective regression methodology together with a very large and rich data set, we test the proposition that exposure to the online assessment instrument enhances student learning. Significantly, we are able to control for prior student aptitude, in-course mastery, gender and even effort via a voluntary class attendance proxy. Furthermore, the study incorporates two large, and statistically diverse cohorts as well as manipulations in the model tested to robustly examine the outcomes. Our central result is that higher exposure to the online instrument robustly leads to higher student learning, all else being equal. Various implications for online assessment design, implementation and targeting are also discussed.

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)

In recent years, online assessment tools have become increasingly used. Instructors are attracted by savings in both marking time and administrative costs of mark compilation, while for students, online quizzes give instantaneous and detailed feedback and greatly enhanced flexibility around the time and place of sitting the test. However, the

educational benefits of such quizzes have been the subject of considerable debate in the literature. 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 online assessment. First, we have a very large data set, with a combined observation count in excess of 1500, approximately an order of magnitude greater than equivalent studies. Second, for robustness, we consider two distinct subpopulations, both studying exactly the same course, with the same online learning treatment, but with characteristics, such as mathematical ability, that are very different. Third, and importantly, because this aspect is most often lacking in the literature, we are able to control for a range of important confounding effects such as prior mathematical ability, in-course aptitude and to a certain extent, student effort. Finally, rather than studying the correlation between student achievement in the online mode of assessment with traditional forms of assessment (as is often the approach), we specifically focus on student *exposure* to the online assessment instrument to gauge what, if any, additional benefit is conferred on learning via the instrument.

Our main finding is that exposure to regular (low-mark) online testing significantly improves student learning as measured by a final proctored examination. Importantly, this result is independent of a student's actual *performance* on each online quiz. Additionally, we use our large data set to show that some frequently used study designs can significantly overpredict the effects of learning instruments such as online quizzes.

Related studies

Broadly speaking, the literature regarding online assessment can be broken down into three principal questions:

1. Are online methods an adequate proxy for traditional methods?
2. Are students at least indifferent between online methods versus traditional methods?
3. Do (regular) online formative assessments enhance student-learning outcomes?

For a successful online learning deployment, each of these questions needs to be satisfied. If the assessment is preferred and helps the learner, but is testing a skill orthogonal to the desired skill-set, it is not worthwhile. Likewise, if the online assessment helps the students and has sufficient overlap with course content, but is less preferred by students, the online component will likely fail through student resistance.

The present study seeks to determine the answer to the third question through rigorous investigation and as such, does not focus on the first two questions. However, from the literature, we may reasonably argue that both question one and two can be answered in the affirmative for online tools. For instance, Bonham, Deardorff and Beichner (2003) noted that online assessments (as compared to manually marked paper assessments) resulted in no discernable difference in student performances over a range of summative assessments (see also Engelbrecht & Harding, 2004). Likewise, Smith (2007) found

that online quiz scores showed higher correlation with final examination marks than laboratory or assignment marks. These data are supported by survey responses of students in Kibble's (2007) study, where 90% of students either agreed or strongly agreed that quizzes were an adequate replacement for formal in-class tests.

For question two, the evidence on whether students prefer online assessments over traditional assessments also appears to be relatively clear. Henly's (2003) study of second year multidisciplinary dentistry/biological sciences students who had the opportunity to complement their studies with a voluntary, multichoice and short-answer online assessment found that 92% of students agreed or strongly agreed that the assessment helped their learning. Only 6% disagreed or strongly disagreed. Similar results can be found in a variety of other contexts (Cassady, Budenz-Anders, Pavlechko & Mock, 2001; Dinov, Sanchez & Christou, 2008; Kibble, 2007; O'Dwyer, Carey & Kleiman, 2007). However, these gains can be extinguished if the online tool is not administered carefully. The experiences reported in Cann (2005) and Ricketts and Wilks (2002) show that if either automatic marking or online assessment manipulation features are not well thought through students can respond very negatively.

Turning now to the third question, which is of especial significance to this study, it would appear that the literature does not provide a definitive answer. A number of studies have attempted to investigate online formative or summative assessment. Of course, within the simple category of 'online assessment', the wide variety of approaches makes comparison difficult.

Martindale, Pearson, Curda and Pilcher (2005) conduct an analysis of the locality specific *Florida Comprehensive Assessment Test* (FCAT) *Explorer* program, an online formative testing resource for students in the Florida jurisdiction. The study compares schools in the jurisdiction that used the FCAT Explorer software with similar schools (eg, by district, size, performance) that did not. An ANOVA analysis indicated that schools using the FCAT Explorer tool had statistically higher mathematics outcomes in fifth-grade mathematics, but not in eighth- or tenth-grade mathematics. Unfortunately, although sample sizes were very large ($n = 586; 491; 1379; 1505$), the study did not apparently control for student exposure to the FCAT instrument, and by the authors' own analysis, usage patterns 'varied wildly' between cohorts. Nonetheless, at the aggregate school level, a weak but significant effect remained. The authors argue that elementary teaching staff may have been more likely to recommend FCAT Explorer use, over secondary school staff, given the perceived heightened importance of FCAT results to younger learners. In contrast, Li and Edmonds (2005) present data from a small study ($n = 22; 10; 16$) in the context of an adult education fundamental mathematics course. The instrument of online instruction was a teacher-built website comprising of 'explanations, examples and interactive exercises' for topics in the course. The treatment group had access to the online component for 1 hour per week (15 hours in total) on top of the normal teaching activity exposure that the control group received. The authors report that along with there being no significant differences in initial pretest results between the treatment and control groups, there were no significant differences

in joint posttest results administered to both groups (although there were some gains for the treatment group in specific skill areas). It would appear that the power of this study was severely hampered by the small sample sizes and furthermore, no accommodation was given to the control group for the additional 15 hours of course specific activity undertaken by the treatment group during their laboratory time.

As referred to earlier, Kibble's (2007) study of a large ($n \sim 350$) Medical Physiology course presented very interesting results concerning the effect of varying the incentives given to students to complete online quizzes. Their design varied the incentive from 0%, to 0.5%, 1% and 2% per quiz. The two quizzes per semester used ANGEL (a virtual learning environment) quiz software and consisted of 20–30 multiple choice questions. In addition to the marks available to students for taking each quiz, the means of obtaining the marks was also varied between login only (for the 0.5% treatment), score >30% (one 1% treatment) and the actual best score from the student's two attempts (a 1% and a 2% treatment). Kibble summarises his findings under three headings: first, under the 0% treatment, students who took the online quizzes did significantly better in a final examination than those who did not; second, when the incentives were increased, student participation rose dramatically (from 52% for the first treatment, to over 95% for the 2% treatment); and third (as mentioned earlier), quiz scores (on the first attempt) were significantly correlated with final examination results.

While this was an intriguing study, and certainly sheds light on how one should set incentives for online quizzes, it is impossible to untangle various confounding effects of quiz engagement and final summative scores. For instance, it is quite possible that students who attempted the quiz in the first treatment were self selecting as high achieving students in the first place. This concern is supported by the fact that there were no significant differences found in final examination results between students who attempted both quizzes versus those who attempted just one.

In the same vein, Henly (2003) uses WebCT (*Web Course Tools*, a proprietary virtual learning environment) to administer a voluntary (formative) online assessment in the form of a mixture of multiple-choice, short-answer and extended matching set questions. She compares the usage patterns of the online assessments and overall course marks of the top and bottom 10% of students and finds that the former uses the online tool significantly more often than the latter ($n = 51$). However, again it is hard to distinguish the actual online quiz effect because it appears, predictably, that good students are found to use the online quiz more, rather than eliciting the added benefit of the online quizzes to any student, after controlling for aptitude.

In perhaps the most comparable study found in the literature to the present work McSweeney and Weiss (2003) conduct a control/treatment survey of the '*Math Online*' (MO) tool—a system which produces randomly generated multiple choice questions around specific skills. Students in the study were encouraged to use the online tool at their leisure, although they were required to complete a minimum number of proctored online tests using the tool during the session. The study is significant since the online

tool used appears to be very mature, with different questions generated each day for each skill, and the proctored tests ensuring that student responses are real rather than as a result of cheating. The study included data from 12 different sections from an applied calculus course with approximately 25–35 students in each section. Additionally, each instructor taught two sections concurrently—one that used the MO tool, and one that did not. The authors compared pre- and posttests of all students and found that the sections using the MO tool had a significantly higher average mark than the control group. Other data showed that the large gains made by the MO sections required active encouragement by the instructor in order that they be realised. Furthermore in the second year of the trial, class time for the MO sections was reduced by 7.5% to accommodate the average additional time students were spending out of class using the tool. Even with the reduced teaching time, the effects remained.

These findings are echoed by others using the more powerful ALEKS (*Assessment and Learning in Knowledge Spaces*, an adaptive assessment and learning system) 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 Hagerty and Smith (2005) and Stillson and Alsup (2003) appear to 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. However, because they do not control for student ability, there is no way to discriminate between their stated hypothesis (that use of ALEKS improves student learning) and an alternate hypothesis that good students use homework tools more. On the other hand, 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 four sections, $n = 119$) by 8% on average (significant at $p < 0.001$ level). Additionally, the authors find via regression that students' learning 'growth' as a result of prior mathematical scores was higher for students using ALEKS than those who did not.

Drawing this section to a close, we note that with only one or two exceptions (eg, Hagerty & Smith, 2005; McSweeney & Weiss, 2003) previous studies of regular online assessment tools have been hampered by small sample sizes (eg, Li & Edmonds, 2005), confounding effects such as type selection bias (eg, Henly, 2003; Kibble, 2007; Stillson & Alsup, 2003) or high variance in instrument exposure (eg, Martindale *et al*, 2005). Despite this, the literature seems willing to suggest at the least a neutral outcome with respect to online formative (or low-mark summative) assessment, and in the most detailed analytical work, a better than neutral outcome. We shall return to this literature further in the study, with references to the outcomes from the present work.

Aims of the study

With the various, and at times patchy, results from the literature studied, it was identified that the literature is full of disparate results, experimental designs and quantitative methodologies. Hence, this study aims to provide a robust analysis of the proposition that regular, online testing improves student learning. While it is acknowledged that not all student learning can be adequately captured by quantitatively measuring student performances, we shall assume for practical purposes that such measures are at least well correlated with student learning outcomes.

Significance

Key to the present study are several layers of robustness. First, we test a fully identified model of student performance, which includes explanatory variables for prior student attainment, unit-specific aptitude, other demographic information and a proxy for effort in addition to our variable of interest. The addition of such factors ensures that any significant effect of the online instrument can be attributable to the instrument and not to a confounding effect, and represents to our knowledge the first such study in the literature. Second, the analysis is conducted on two statistically different cohorts of students. Third, both cohorts are large, and indeed, one is very large—a further factor that has seldom been represented in the literature. Hence, the study aims to distinguish itself by representing a solid resting point for future inference concerning the efficacy of online testing, which offers a low proportion of overall marks.

Instruments

The course of inquiry for this study is called *Quantitative Methods A* (QMA), a first-year core applied mathematics course taught by the School of Economics within the Australian School of Business (until recently known as the Faculty of Commerce and Economics) at the University of New South Wales, Sydney, Australia. QMA introduces students to topics such as financial maths, linear algebra (matrices), linear programming and optimisation, and calculus with up to several variables.

In the period under study, the course included four main assessment instruments, three of which are focussed on in the present study:

1. 4 × online quizzes: (worth 2% each) described in detail later;
2. Mid-term examination: (20%) closed-book, auto-marked multiple-choice (20), proctored, administered half-way through teaching semester;
3. Group assignment: (12%) excel spreadsheet assignment, scenario-based, groups of 1–4 students, manually marked (with template), worksheet and report handed in;
4. Final examination: (60%) closed-book, manually marked short-answer problems, proctored, administered after the semester.

Online e-quizzes

As with several approaches in the literature (eg, McSweeney & Weiss, 2003; Smith, 2007; Stillson & Alsup, 2003) the online quiz instrument in this study was, strictly speaking, summative in nature because 2% of overall marks were awarded for each

quiz, however, the design of the quizzes was heavily informed by formative assessment models. A very similar structure to that of Kibble's (2007) 2% model was employed, with the best attempt from two 1-hour attempts over a 1-week testing period used for marking purposes. Unlike Kibble, who apparently did not vary the questions between students or attempts, our quizzes were prepared from a reasonably small set of questions that used variables to generate 80 different numerical variations. Thus a very large set of 'different' questions and answers was created and could be easily modified. The size of our question bank ensured a low probability that two students would receive exactly the same quiz. In the second sample cohort reported further in the text, additional questions were prepared and a randomised subset was used to compile the quizzes for each attempt. This ensured that not only would the numbers change, but the question order and composition would also vary between attempts.

All quiz questions required the entry of a single (calculated) number in a blank form. This kind of quiz can cause very real (and extremely unpopular) negative outcomes if tolerancing and educative measures are not employed (see Cann, 2005). However, we gave clear instructions to students about the format of answers such as the number of decimal places to use and characters to avoid. We set tolerancing within the WebCT Vista software such that in almost all instances, a five-unit error of the correct least significant value would be accepted as correct. Even in the very large cohort of over 1000 students, complaints arising as a result of the mismarking of a student's answer were rare and easily handled by manual inspection of their input. Feedback in the form of the correct final answer was immediately given to students on completion of their whole attempt. Students were encouraged to go over the questions where they made mistakes and so, to attempt to find working that would yield the correct answer. As an additional learning tool, a purely formative multiple-choice quiz was made available online in preparation for the week of the actual summative quiz with questions of a similar difficulty and content. Once this 'practice-quiz' was completed by a student, fully worked solutions were automatically fed back to the screen for further study. Taken together, the quiz approach in this study used marks as incentive but in all other aspects sought to provide a rich formative experience for student learning. The data presented below uses only the summative portion of this approach.

The quizzes were written with the third-party software called *Respondus*[1] before being uploaded into WebCT Vista. *Respondus* has a number of inbuilt functions such as logarithms and exponentials, which enable certain types of calculated questions to be written easily; but for more exotic constructions (eg, matrices) a little lateral thinking had to be employed. An example of a completed matrix question is shown in Figure 1. 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.

Methodology

The study employed a retrospective regression methodology somewhat similar in approach to that of O'Dwyer, Carey & Kleiman (2007) and Brown and Liedholm

[1]Software details available at http://www.respondus.com/

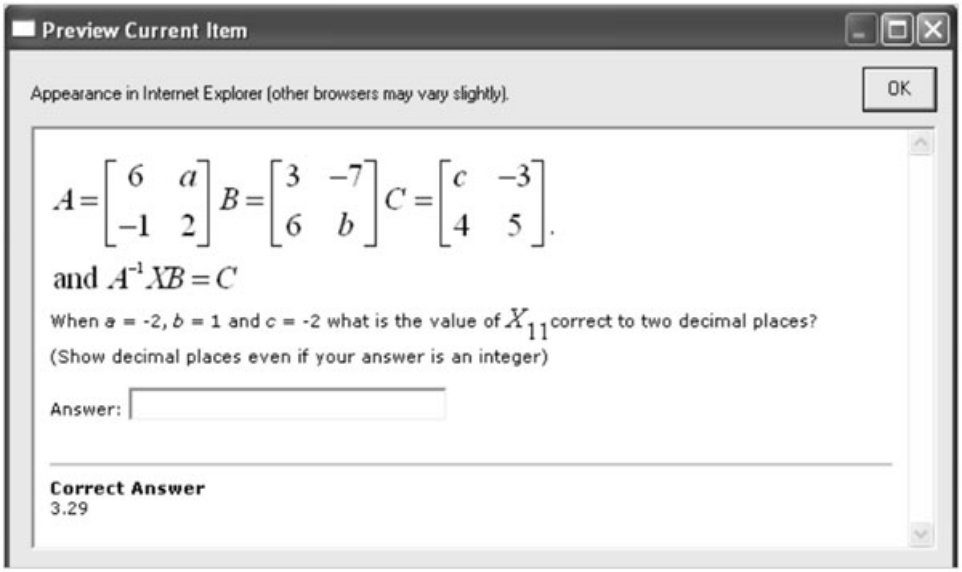

Figure 1: Completed respondus question

(2002). In line with our study aim, we interpreted 'improvements in student learning' as a result of 'regular, online testing' to be signalled by *the attainment of a higher final examination mark for students who completed a higher number of the online assessments than average*. Notice that our study did *not* consider the actual mark obtained by a student in the online instrument (as in eg, Smith, 2007), rather it focussed solely on the extent to which the student was *engaged with* that instrument (as in, eg, Stillson & Alsup, 2003). We believe this is an important distinction to make in testing our proposition since we were not concerned with correlating student performance in (formative) online assessment with summative examination results (as is popular in the literature). Rather, we were investigating the link between engagement with the online instrument and summative final examination performance. Because we wished to examine the connection between the online instrument and student learning more robustly than previous authors, our study utilised a more elaborately specified regression approach with many explanatory variables besides the variable representing the online quiz instrument. In particular, we focused on three main sources of confounding information: a student's prior mathematical ability, a student's in-unit level of mastery and a student's level of effort towards the unit. To our knowledge, such a comprehensive use of explanatory variables has not previously been applied to the question of online learning.

To be precise, we developed the following model to investigate our central proposition:

$$FE_i = \beta_0 + \beta_1 QUIZB_i + \beta_2 MT_i + \beta_3 PASSB_i + \beta_4 HSCLOWM_i + \beta_5 HSCHIGHM_i + \beta_6 NONHSC_i + \beta_7 GEN_i + \varepsilon_I \quad (1)$$

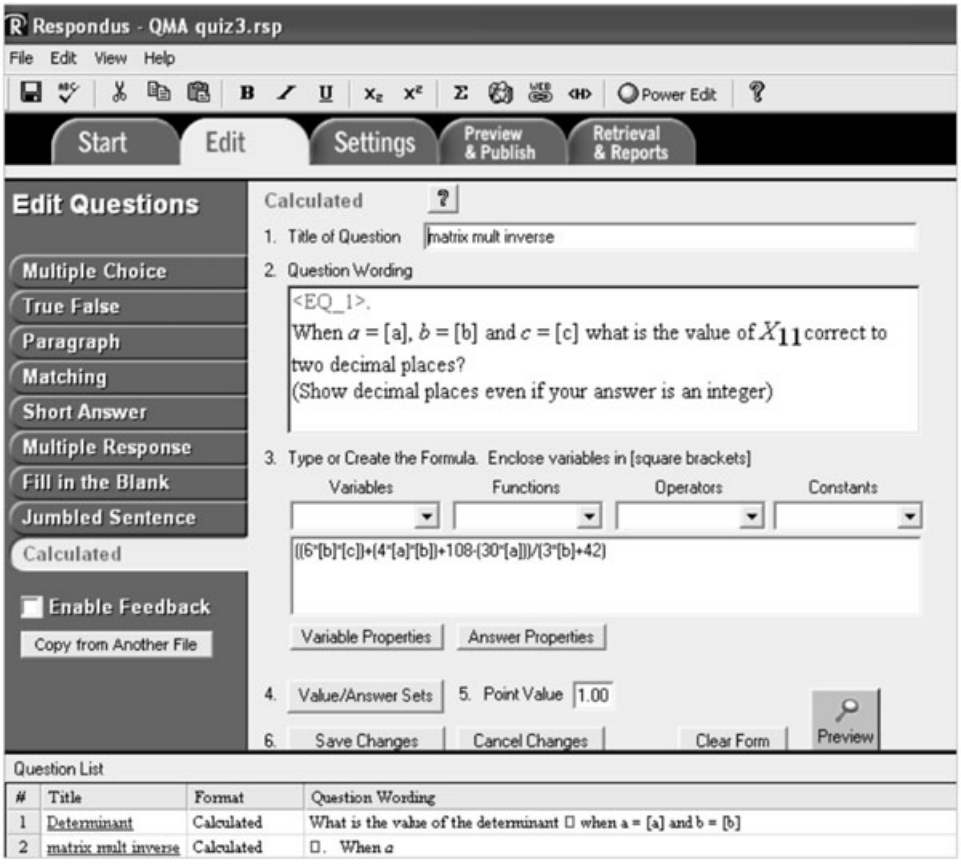

Figure 2: Respondus interface showing setup using variables

where the dependent variable used is the student's *final examination mark* (in %), encoded as *FE*. The variable of interest is the online quiz binary variable *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.[2] Attempting four online quizzes was chosen for the dummy to increase the power of the regression, as this would ensure that the greatest number of observations would obtain 0 for *QUIZB*. Next, for prior mathematical ability we include the three dummy variables *HSCLOWM*, *HSCHIGHM* and *NONHSC* to indicate the student's prior exposure to mathematical courses in their final year of secondary school. The mathematics marks from different high school courses are not on a continuous scale so dummy variables were necessary. Although the stu-

[2]The following analysis was repeated with QUIZB taking 1 if the student recorded a score for 3 or more (of the 4) online quizzes and 0 otherwise. This alternate specification changed neither the sign, the magnitude or significance of any results in an important way.

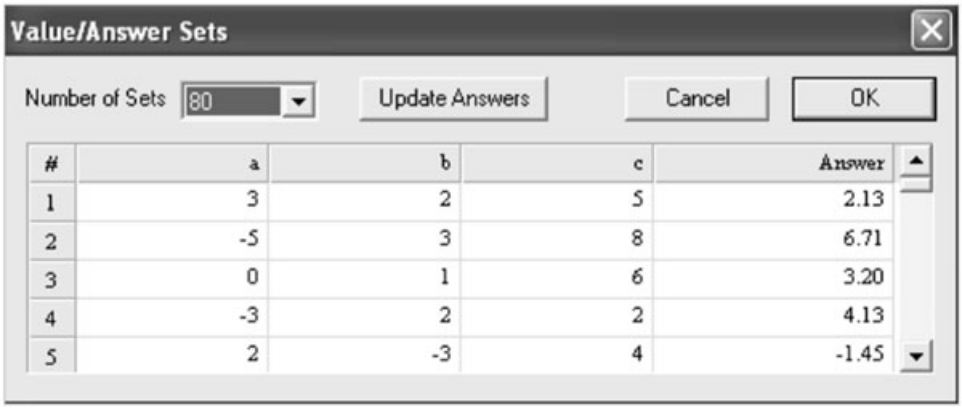

Figure 3: Examples of answer sets

dents studied under a variety of education systems, it was found that a large percentage (over 85%) had taken Higher School Certificate (HSC) mathematics in New South Wales. For these students, middle-level mathematics, ie, *Mathematics* plus *Extension 1* was considered the baseline (omitted) category. The low prior maths attainment dummy *HSCLOWM* took the value 1 if a student took either *General Mathematics* or *Mathematics* alone. The higher than average mathematics attainment dummy *HSCHIGHM* took the value 1 if the student took the HSC *Extension 1* plus *Extension 2* combination. It was not possible to obtain prior mathematical data for those who had not studied the HSC, so no attempt was made to distinguish between the disparate levels of alternate high school mathematics. For these students, the *NONHSC* dummy took the value 1. It is important to note, therefore, that the baseline student case, ie, *HSCHIGHM* = *HSCLOWM* = *NONHSC* = 0 represents a student who sat under the HSC mathematics system in high school at the middle-mathematics level.

As a proxy for a student's in-course level of mastery, we used the student's midterm examination result as a simple percentage score and encoded this as *MT*. While understandably no perfect proxy for student effort was available, we did have data on student attendance at a *Peer Assisted Support Scheme* (PASS), which operates in QMA. Students are able to attend peer-led study groups on a drop-in basis. Importantly, a previous study by Watson had linked increased PASS attendance with higher QMA outcomes (Watson, 2000). We were aware from this study and our own analysis that as PASS is not a remedial scheme attendance was not confined to students from lower mathematical backgrounds. Indeed, attendance was uniformly or slightly bimodally distributed with respect to unit aptitude, with both high performance and weaker students taking advantage of the additional materials and instruction that PASS affords, presumably seeking to further maximize their marks. Hence, the inclusion of PASS as an effort proxy seemed reasonable in the present study. As a result of some missing attendance data during a few weeks of the second sample period, the *PASSB* dummy was con-

strained to take the value of 1 if a student attended PASS more than twice in the given session, although a variable based on more frequent attendance would have been preferred. The PASS study groups were available in eleven weeks of the session. Finally, we also included a gender dummy (encoded as *GEN*, and taking 1 if female) to account for any gender-based variation contribution to final examination marks.

On a final technical note, because the LHS variable is bounded on the real interval [0,1], to conduct our study we estimated the logistic function transformation to the FE variable,

$$L_i = \ln[(\mathrm{FE}_i + 0.5)/(1 - \mathrm{Fe}_i + 0.5)] \tag{2}$$

Hypothesis

For clarity, we restate our aim and associated hypothesis to be tested. As in previous sections, the study sought to investigate the link between online quizzes and student learning by testing the connection between exposure to the online learning instrument and end-of-session examination performance. Thus, in what follows, we shall focus on the sign and significance of the estimated value β_1. If β_1 is found to be positive and significantly different from zero then we may conclude that *ceteris paribus* students who attempted all the online quizzes received a higher examination mark than their colleagues, controlling for their prior mathematical attainment, gender, their in-course aptitude and their level of effort displayed elsewhere in the course. It is important to note that because (like, eg, Bonham *et al*, 2003) we do not compare the online instrument with (say) a control group exposed to a paper version of the same tests, we cannot actually discriminate between the online nature of the instrument, and other factors such as its regularity or effort-inducing incentives. However, to the extent to which the online instrument is a relatively low-cost, efficient and formative component, we may judge that such an instrument is supportable on cost-benefit terms, a consideration we return to below.

Data

Data were collected for QMA students enrolled in Session 2 2006 (hereafter referred to as *Sample 1*) and Session 1 2007 (*Sample 2*) (the same course is presented in each session). Our experience in QMA suggested that Sample 1 contained a large proportion of repeat students and new international students while the majority of Sample 2 students were in their first session of university after leaving high school. The two samples made for a diverse data set to test our hypothesis and were thus a key feature of the study's robustness. 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 as opposed to the significantly larger 1239 observations in Sample 2.

The QMA course has a multicultural enrolment, so it was not surprising to find that the majority of students in both samples were born overseas, predominantly in China,

Hong Kong, Indonesia, South Korea and other Asian countries. In the main session (Sample 2), while only 45% of students were born in Australia, 82% were Australian residents. By contrast Sample 1 had more students from overseas. Only 32% of students were Australian-born and fewer (61%) were Australian residents. Comparing all Australian universities in 2007, we find that a much higher percentage (76%) of internal students were born in Australia, and more of those who were born overseas came from countries such as England and New Zealand, where English was a first language (Department of Education, Employment and Workplace Relations, 2008).

There were more repeating students in Sample 1. The vast majority (91%) of Sample 2 students were in their first session of university after leaving high school whereas only 33% of Sample 1 students were commencing. Age differences between the samples result from many students in Sample 2 either repeating or arriving from overseas at a later age. Twenty five per cent of Sample 1 students were aged 20 years or older compared with 12% of Sample 2 students. There were differences in gender as well, with males representing 53% of Sample 2 students but only 45% of Sample 1.

Summary statistics of all continuous and dummy variables are given in Tables 1 and 2 respectively. As these data indicate, the two samples differed markedly in some other

Table 1: Continuous variable summary statistics

	Sample 1 (n = *397*)		*Sample 2* (n = *1239*)	
	FE	*MT*	*FE*	*MT*
Mean	0.55	0.47	0.60	0.71
SD	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 2: Dummy variable summary statistics

	Sample 1 (n = *397*) $\bar{x}$	*Sample 2* (n = *1239*) $\bar{x}$
GEN	0.55	0.47
HSCHIGHM	0.17	0.31
HSCLOWM	0.45	0.31
NONHSC	0.35	0.17
PASSB	0.14	0.15
QUIZB	0.80	0.83
STRATC	0.07	0.01
STRATD	0.09	0.02

important features. The differing intake routes for each cohort are evident in the prior maths attainment dummies. In contrast to the main (larger) Sample 2 session cohort, which had an equal portion of higher and lower mathematics students by HSC course, Sample 1 had approximately 2.5 times more students whose maths background was classified as low rather than high as a result of their HSC studies. Similarly, where Sample 2 had only a relatively small number (17%) of students without the local HSC mathematics background, this figure was over a third for Sample 1. The proportion of female students in the cohort was also unsurprisingly higher for Sample 1, matching our own observation that females make up a higher proportion of international students arriving midyear.

In terms of the other descriptive statistics, the proportion who qualified for the *PASSB* and *QUIZB* dummies were very similar in the two samples. In each, around 80% of students attempted all four of the online quizzes. The complete breakdown is as follows, for Sample 1, >1% attempted no quizzes, 2% attempted 1 quiz, 5% attempted two quizzes, and 13% attempted three quizzes; and in Sample 2, 1%, 2%, 4% and 11% attempted zero, one, two and three quizzes respectively. The fact that the PASSB proportions were similar for the disparate cohorts is understandable as the study sessions are not remedial but a resource that appears to attract students of all mathematical abilities. Finally, it was not surprising to find that the two cohorts performed in line with expectations, with the Sample 2 final examination and midterm examination results both significantly higher (at $p < 0.01$ level) than those of the Sample 1 group.

Taken together, the demographic and prior maths attainment data, together with the average performance statistics, indicate that the two samples represented markedly different student cohorts and thus provided an excellent joint test bed for the strength of any resultant analysis.

Results

Ordinary least squares regression results of model (1) are presented in Table 3. At first pass, we note that the main coefficients have largely expected signs and magnitudes; that both estimations passed tests for autocorrelation (Durbin-Watson) and model specification (F-test); and that heteroskedasticity (present by Breusch-Pagan test) was handled by rerunning the regression using the heteroskedasticity consistent covariance matrix approach. Hence, we move on to the analysis of the coefficients.

We recall that if the estimated coefficient on *QUIZB* was positive and significantly different from zero then we may claim support for our proposition that regular online testing enhances student performance. Looking at the table, this finding appears to be confirmed. The *QUIZB* coefficient was positive and significant (at the $p < 0.01$ level) for both samples. Although not reported here, as a robustness check, this result was found not to change when the *QUIZB* dummy was relaxed to take the value of 1 if a student recorded a score for three *or* four quizzes (out of four). Furthermore, the value of the elasticity at means for *QUIZB* of around 2.5% indicates a reasonably high return to

Table 3: Final examination (%)

	Sample 1		*Sample 2*	
	Estimated coefficient	*Elasticity at means*[a]	*Estimated coefficient*	*Elasticity at means*
QUIZB	**0.582****	**2.12**	**0.509****	**2.76**
	(5.38)		(7.75)	
MT	**2.491****	**7.56**	**3.018****	**13.67**
	(9.91)		(19.92)	
PASSB	*0.255**	*0.17*	–0.075	–0.07
	(2.49)		(–1.45)	
HSCLOWM	**–0.312**[**,b]**	**–0.43**	**–0.363****	**–0.74**
	(–3.02)		(–6.90)	
HSCHIGHM	**0.520****	**0.71**	**0.406****	**0.83**
	(4.14)		(7.23)	
GEN	0.094	0.20	**0.123****	**0.38**
	(1.18)		(2.94)	
NONHSC	0.144	0.11	0.087	–0.10
	(1.52)		(–1.43)	
Constant	–1.519**		–2.155**	
	(–9.57)		(–17.47)	
n	397		1239	
Adjusted R-squared	0.366		0.486	
Regression F (from zero)	34.33		216.64	

Notes. t-ratios for each coefficient given in parenthesis, where two regressions were run (see second note), significance levels were drawn from the second (corrected) regression. In both cases, the Breusch-Pagan test suggested the presence of heteroskedasicity, in which case, *t-stat* values are reported from a second regression using the heteroskedasticity-consistent covariance matrix approach.
Significance levels indicate $p < 0.05$ (*), and <0.01 (**) respectively;
[a]Elasticity at means gives the expected percentage change in Final Examination scores for a 10% change in the given variable estimated at mean values of all variables (as in Tables 1 and 2).
[b]Coefficient significant at $p < 0.01$ level when *QUIZB* defined for students recording >3 quiz scores, but only significant at $p < 0.05$ level when *QUIZB* defined for students recording >2 quiz scores.

engagement with the online quiz instrument—if students were 10% more likely to sit for all four quizzes, they could expect to receive around 2.5% higher on their final examination, *ceteris paribus*.

We noted other features of the regression results. The most striking initial observation concerns the similarity in sign, magnitude and significance between results for the two samples on the *QUIZB*, *MT*, *HSCLOWM* and *HSCHIGHM* estimated coefficients. It was not surprising that students with higher prior mathematics attainment or good midterm assessment scores had associated higher final examination results. For the same reasons, those students who studied the relatively lower high school mathematics courses received lower final examination results on average. As these results were consistent across the two diverse student samples, we may conclude that these influ-

ences on student performance are robust to changes in overall student attributes. In other words, despite our best attempts, student outcomes in the final examination appeared to be heavily influenced by their initial attributes. Of the other variables, this study is in weak agreement with the earlier reported result (Watson, 2000) on the efficacy of PASS on student learning with the *PASSB* dummy positive and significant ($p < 0.05$) for Sample 1 only. As reported earlier, construction of the variable *PASSB* suffered somewhat because of a few missing weeks in the data set, so the lack of a clear statement on this point could be explained in this manner. Similarly, no clear statement appears possible on the relationship between gender and the final examination.

Discussion

The present study's finding supports the apparent consensus of the literature, and in particular, resonates with the study of McSweeney and Weiss (2003) (MW). As noted above, although the study of MW fundamentally differed in research methodology—they use a treatment/control approach whereas the present study utilised a retrospective regression methodology—there are similarities in design that perhaps contribute to their coherent findings. For instance, MW focus on the *change* in pre- and post- test algebra scores and so effectively control for domain-specific aptitude as was achieved in the present study by the midterm, and HSC maths dummies. One could also compare the results of the present work to that of Kibble's creative multi-incentive study (Kibble, 2007) which found that in the no-marks (and thus, purely formative) treatment, those who undertook at least one of the online quizzes (of the two), gained a statistically higher final exam score. However, as noted earlier, this result is potentially tangled with a selection bias influence.

Following on from what was previously presented, it is possible to advance this discussion by asking, what would be the implication of *not* including all of the other explanatory variables in the regression (as per [1]) and simply correlating the *QUIZB* exposure dummy with at student's final examination score? Doing this yields a significant coefficient for *QUIZB* almost exactly *twice* the magnitude of the actual full regression outcome when the complete model in Equation 1 is run. For completeness, another popular question of the literature concerning whether a student's quiz score is correlated with their final examination score was tested. This approach, as reviewed earlier, is the approach of very many studies that test the *correlation* of online assessment scores with a summative score (eg, final examination). Running the simple regression of FE on mean online quiz score (recall, taking the best of each two attempts, in four quizzes) unsurprisingly yields a positive, strongly significant (t-ratio = 27.19) correlation between the two variables.

The lesson we would urge is that without careful attention to study design, or data analysis, one can easily infer stronger statements than the results suggest. In the case of simple regressions of final examination versus online quiz engagement (or score) (as in Kibble, 2007; Smith, 2007) the danger exists that much of the variation as a result of student-specific aptitude, prior mathematics attainment or gender could be confounded in a quiz coefficient.

It is perhaps timely to acknowledge the limitations of the present study. First and foremost is the fact that, because of the retrospective regression design rather than a treatment/control online/offline assessment process, it was impossible to distinguish the effects of 'regular quizzes' from 'regular *online* quizzes' in this study. That is, our study is entirely consistent with one that simply administers a regular (low-mark) quiz instrument, regardless of whether the administration is virtual or manual. However, as noted in the methodological section of this paper, we would argue that there are at least two main objections to running regular manually marked quizzes, namely (1) the increased course administration costs; and (2) the 'low level' formative component of normal proctored testing. It is for these reasons that an online study of such regular low-mark testing deserves merit—in particular, the formative aspects of this kind of assessment (eg, the opportunity for multiple attempts, immediate feedback, randomised questions and numbers) are arguably only attainable in the online format. Hence, we might moderate the finding of the present study to assert that 'regular testing *undertaken with online methods* enhances student learning'. The shift in interpretation is subtle, but important.

A second limitation of this study was the treatment of missing student data. As noted in our Data section, students who dropped out of the course, or who sat a supplementary examination were not included in the study. A very quick survey of the Sample 2 data set indicates that the number of students for whom this occurred was around 33 or less than 3% (in Sample 2). This is a small, but potentially significant, fraction. Indeed, given the experience of some authors using the more muscular ALEKS package (eg, Stillson & Alsup, 2003) where students who dropped the course appeared heavily influenced by the imposition of the new online testing format, the dropout cases should not be discounted out of hand. While we are not aware of any of these students who dropped the course as a result of the online testing instrument, it is our hope to more thoroughly investigate such student experiences in further work.

With reference to the other two questions posed by the literature, we can concur with the findings of the several authors mentioned earlier (eg, Cassady *et al*, 2001; Dinov *et al*, 2008; Henly, 2003; O'Dwyer *et al*, 2007). We found that via the end-of-session course questionnaire undertaken by 495 students (Sample 2), 51% of respondents *agreed* and a further 41% *strongly agreed* with the statement 'The online e-quizzes were a useful tool to help me study consistently throughout the course'. While this exact question has not apparently been posed in the literature, variants on the same theme have been and the present results are found to be in agreement. Moreover, with respect to the difficulties in student satisfaction that Cann (2005) and Ricketts and Wilks (2002) encountered, the positive response for the quiz instrument used in the present study further supports the observation that going online can be administered in an appealing way so long as student–quiz interactions are simple and reasonable.

Finally, we might ask, is there anything particular about the students who did not record at least one attempt for all available online quizzes? In particular, having established that low exposure to the online quizzes had a significant and negative effect on

final marks, *ceteris paribus*, was there an identifiable group of students who were most likely to fall into this category *ex-ante*? To answer this question, a Logit estimation of the following model was estimated:

$$QUIZB_i = \beta_0 + \beta_1 HSCLOWM_i + \beta_2 HSCHIGHM_i + \beta_3 NOHSC_i + \beta_4 GEN_i + \varepsilon_i \qquad (3)$$

The model supposes that a student's likelihood of attempting all available quizzes will be a combination of their pre-existing mathematical aptitude (and local or international status) or their gender. For (the smaller) Sample 1, the model would be accepted at the $p < 0.05$ level only because the Likelihood Ratio Test statistic was 12.56 (4 d.o.f., $p = 0.0136$). If one proceeds on this basis, then only the coefficient on *GEN* was positive and significant (t-ratio = 3.02) with a marginal effect of around 12%. Whereas, for Sample 2, owing to the much larger sample, the Likelihood Ratio Test statistic was over 46, which is acceptable at the $p < 0.01$ level. In this sample, the coefficient on *HSCLOWM* was significant (t-ratio = −4.78) indicating that local students who studied the lower *General Mathematics* or *Mathematics* course only during their final year of high school were more likely not to complete all of the quizzes even though they completed the course. Interestingly, the gender dummy, although positive, was not significant for Sample 2 (t-ratio = 1.24), indicating that gender did not play a big role in determining quiz activity. Taking the two sample results together, there appears no clear signal on predictors for low quiz engagement. The gender finding of Sample 1—that females were more likely to attempt all the quizzes—can be contrasted to the findings of Hoskins and Hooff (2005) who found males more likely to engage in online tool use, although their discussion shows that other authors find agreement with the present result. It seems that further work is required on the role of gender in online engagement. Returning to the larger Sample 2 result, one could potentially conclude that extra effort would be well spent at (say) time of enrolment to encourage students with a lower than average mathematics background to persevere with online tools and perhaps with other aspects of study. However, this finding would require further supporting evidence from the field.

To sum up, the study suggests three main findings: first, that exposure to a regular (low-mark) online quiz instrument has a significant and positive effect on student learning as measured by an end of semester examination; second, that caution is required in the interpretation of simple regression-, or correlation-, based studies, because we show that large, but ultimately questionable, effects can arise without careful study design and/or data analysis; and finally, that tentative evidence was uncovered indicating that low attainment students *ex ante* are less likely to voluntarily attempt all online quiz components, thus suggesting a possible direction for learning intervention.

Acknowledgements
The authors wish to gratefully acknowledge Peter McGuinn for invaluable support in setting up the WebCT Vista system used in this study, and Denzil Fiebig for helpful comments on an earlier version of this manuscript. The authors also thank two anonymous referees for generous comments and suggestions, which led to substantial improvements.

References

Bonham, S. W., Deardorff, D. L. & Beichner, R. J. (2003). Comparison of student performance using web and paper-based homework in college-level physics. *Journal of Research in Science Teaching, 40*, 10, 1050–1071.

Brown, B. W. & Liedholm, C. E. (2002). Can web courses replace the classroom in principles of microeconomics? *American Economic Review, 92*, 2, 444–448.

Cann, A. J. (2005). Extended matching sets questions for online numeracy assessments: a case study. *Assessment & Evaluation in Higher Education, 30*, 6, 633.

Cassady, J. C., Budenz-Anders, J., Pavlechko, G. & 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.

Department of Education, Employment and Workplace Relations (2008). 'Students 2007 [full year]: selected higher education statistics; ethnicity related data'.

Dinov, I. D., Sanchez, J. & Christou, N. (2008). Pedagogical utilization and assessment of the statistic online computational resource in introductory probability and statistics courses. *Computers & Education, 50*, 1, 284–300.

Engelbrecht, J. & 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, 217–231.

Gibbs, G. (1999). 'Using assessment strategically to change the way students learn'. In S. Brown & A. Glasner (Eds), *Assessment matters in higher education* (pp. 41–53 (chapter 4)) Buckingham: S.R.H.E. and Open University Press.

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

Henly, D. C. (2003). Use of Web-based formative assessment to support student learning in a metabolism/nutrition unit. *European Journal of Dental Education, 7*, 3, 116–122.

Hoskins, S. L. & Hooff, J. C. V. (2005). Motivation and ability: which students use online learning and what influence does it have on their achievement? *British Journal of Educational Technology, 36*, 2, 177–192.

Kibble, J. (2007). Use of unsupervised online quizzes as formative assessment in a medical physiology course: effects of incentives on student participation and performance. *Advances in Physiology Education, 31*, 3, 253–260.

Li, Q. & Edmonds, K. A. (2005). Mathematics and at-risk adult learners: would technology help? *Journal of Research on Technology in Education, 38*, 2, 143.

McSweeney, L. & Weiss, J. (2003). Assessing the math online tool: a progress report. *Mathematics and Computer Education, 37*, 3, 348.

Martindale, T., Pearson, C., Curda, L. K. & 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, 349–360.

O'Dwyer, L., Carey, R. & Kleiman, G. (2007). A study of the effectiveness of the Louisiana Algebra I online course. *Journal of Research on Technology in Education, 39*, 3, 289.

Ricketts, C. & Wilks, S. J. (2002). Improving student performance through computer-based assessment: insights from recent research. *Assessment & Evaluation in Higher Education, 27*, 5, 475.

Smith, G. (2007). How does student performance on formative assessments relate to learning assessed by exams? *Journal of College Science Teaching, 36*, 7, 28.

Stillson, H. & Alsup, J. (2003). Smart ALEKS ... or not? Teaching basic algebra using an online interactive learning system. *Mathematics and Computer Education, 37*, 3, 329.

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, 5–7 July.