Australasian Journal of Economics Education
Volume 12, Number 2, 2015, pp.50-74

STUDENT ENGAGEMENT AND EXAM PERFORMANCE: IT'S (STILL) ABILITY THAT MATTERS MOST*

Mary R. Hedges
Centre for Longitudinal Research,
University of Auckland
and
Gail A Pacheco
Department of Economics,
Auckland University of Technology

ABSTRACT

The use of online quizzes has become more popular in introductory economics courses in recent years, however their efficacy in improving student engagement and performance has seldom been evaluated. This paper attempts to undertake an evaluation of that efficacy by combining individual engagement and exam performance data with demographic information for a cohort of students enrolled in a first year economics course. In a novel extension to this literature, we also employ bivariate probit modelling to allow for a sequential process between student engagement and performance, and to permit the investigation of determinants of exam performance after conditioning results on the factors that influence engagement. This empirical analysis yields interesting differences by gender.

Keywords: threshold concepts, student engagement, ability, bivariate probit, constructive alignment

JEL classifications: A22.

1. INTRODUCTION

In 2009 the Bachelor of Commerce (BCom) at the University of Auckland was subject to a major curriculum review. The outcome of this review was the introduction of two new, integrated first year business papers (courses) into the core of the degree (University of Auckland Business School, 2009). One of the costs of this change

* Correspondence: Gail Pacheco, New Zealand Work Research Institute, Auckland University of Technology, Private Bag 92-006, Auckland 1142, New Zealand, Email: gail.pacheco@aut.ac.nz.

ISSN 1448-4498

was that economics would be reduced from two compulsory papers in the first year to only one (hereafter referred to as 191). The second implication of this wider curriculum review was that this new first year core paper would be common to all three degrees offered in the Business School: Bachelors of Commerce, Property, and Business Information Management (BBIM). These changes were introduced for the 2011 academic year starting with semester 1 in March that year.

Prior to 2011, students did a full semester of microeconomics (101) and macroeconomics (111), the former being a prerequisite for the latter. In contrast 191 was developed as a mixed micro- and macroeconomics paper. For students not majoring in economics this would probably be the only economics paper that they would take in their business degree. Therefore it was imperative that the course would teach students the basic principles of economics and how they could be applied to a wide range of real world situations. At the same time the new course needed to meet external accreditation requirements including NZICA (New Zealand Institute of Chartered Accountants) and international business school accrediting programmes (AACSB and Equis). For students who do major in economics the course had to prepare them to progress to higher level economics papers without having to take an additional course. Having to take an additional course would make economics a less attractive major option compared to all other majors available which could impact on enrolments.

Whereas many course re-developments begin by considering the course content that needs to be added or removed, our course re-design was treated as a green-fields process the starting point of which was consultation with industry, faculty and accrediting agencies. Through these consultations a list of learning outcomes was developed. This process not only considered the topics and technical skills that were required but also how the students would be able to apply that content to a range of real world situations. That is, the focus was strongly on capability building rather than just content acquisition. It was only after this consultation stage that mapping the new outcomes and content against the existing ones took place.

The resulting course was one in which the ideas and concepts of constructive alignment (Biggs 2003) were applied. Consequently, the content, teaching methods, application of theory and assessment items were all aligned with the main purpose of meeting the core learning

objectives; i.e. engaging student learning by developing the students' application of economic concepts and critical thinking skills. Furthermore, the assessment programme became a key teaching tool instead of acting primarily as a measurement tool (Rowntree 1987). That is, the goal was to use the assessment strategically to encourage the type of engagement and integrative learning that we were seeking (Gibbs 1999).

In designing the assessment structure, there were several factors to consider. For instance, balancing the pressure of a high number of assessments in a single semester (12 week) course, with the need to keep students on task, with their economics course at the top of their priority list. By employing the concepts of constructive alignment, we wanted students to:

- get early feedback on their understanding, and reflect on this feedback;
- be able to apply theory to real world problems (done to a limited extent previously);
- recall what they had learned (post completion of course).

Given that one of the core learning objectives of the new economics course was to engage students more in their learning, a key purpose of this article is to explore whether the constructively aligned approach taken and emphasis on engagement, had a positive outcome on student results, with respect to their final exam mark. We also identify whether the role of demographic and other individual characteristics on exam performance differ depending on whether the student has above or below average engagement in their economics paper. Finally, we condition our results on the factors that are associated with engagement, to better understand what predictor variables remain significant, in terms of their association with exam performance. This analysis is also carried out separately for males and females, as initial findings indicate clear gender differences in both level of engagement and performance in economics courses.

The remainder of this article is organized as follows: First we provide a brief literature review defining and linking the concepts of constructive alignment and student engagement with exam performance. This is followed by an outline of the data used, and details of the specific assessment structure developed for this economics course. We then describe the methodologies employed, followed by a discussion of the results and conclusions.

2. LITERATURE REVIEW

The principles of constructive alignment (Biggs 1999, 2003) necessitate that a course (and wider programme) should seek to engage the students in a range of learning activities that involve compatibility between curriculum, teaching methods and assessment procedures (Taylor 2002). However, to constructively align the curriculum to achieve the course aims and objectives is subject to a number of significant constraints. In this particular environment these constraints included:

- expected large numbers (up to 1500 students per semester once the new structure was fully introduced);
- resourcing limits:
    - contact time limited to 3 hours of large lectures (up to 600 students in each lecture);
    - potential for a one hour tutorial per week (class sizes up to 30 students);
    - marking assessment constraints including:
        - assessments in an essay format would be difficult for many markers to deal with effectively, consistently and in an appropriate time frame (markers are predominantly post-graduate students within the economics department);
        - timeliness of feedback given large size of class;
        - early feedback to ensure concepts are understood;
- lecturer constraints to cover the number of streams and maintain consistency between streams;
- core theory coverage to enable progression to higher level economics papers.

It quickly became obvious that the only way to achieve our aims, within the bounds of the constraints noted above, was to take a 'less-is-more' approach. This was based on the pedagogy of threshold concepts (Davies 2003; Davies & Mangan 2005, 2008; Land, Cousin, Meyer & Davies 2005; Land & Meyer 2006; Meyer & Land 2005), where there is a large body of literature (within the economics discipline) to draw on. Central to many of these studies is the 'Embedding Threshold Concepts' project funded jointly by the Higher Education Funding Council for England (HEFCE), the Institute for Education Policy Research, Staffordshire University and Enhancing Teaching-Learning Environments in Undergraduate Courses projects

in the UK (Staffordshire University 2008). This project had produced a number of resources to aid curriculum design based on these ideas. Also, fortunately, economics has a very standard global body of such concepts that can be applied at the local level, albeit with a New Zealand perspective.

Considering the curriculum in terms of threshold concepts meant actively considering what these concepts are and how they differ, or remain similar to fundamental concepts which are what the standard economics courses tended to be based on. Threshold concepts are considered to be a 'portal' or 'conceptual gateway' (Meyer & Land 2003) to a discipline and are characterised as being:

- bounded;
- integrative;
- transformative;
- potentially troublesome;
- probably irreversible.

They have been placed in this specific order due to the path dependency required in developing some of these ideas (Hedges 2011).

In contrast, fundamental concepts are considered 'building blocks' or 'core ideas' that are considered essential to progress learning in the discipline. These are often used in curriculum design and were certainly the basis of 101/111. Their use may result in a 'theory first' approach that sees complex ideas simplified and risk 'rote' learning of methods that lose of sight of deeper ideas. The focus on fundamental concepts often divorce understanding from experience or at best relates theory to experience as an example (Davies & Mangan 2005). In contrast, threshold concepts keep the application as central to the concepts taught. Foundational concepts are very useful in distinguishing one type of economics from another but less useful in distinguishing economic thought from other disciplines (Davies 2003). Clearly focussing on threshold concepts in the re-design of 191 was a major pedagogical shift from how the previous courses were developed.

For threshold concepts to be integrative and transformative the student must be actively engaged in their acquisition. Student engagement is variably defined as:

> [M]aking a psychological investment in learning. [Students] try hard to learn what school offers. They take pride not simply in earning the

> formal indicators of success (grades), but in understanding the material and incorporating or internalizing it in their lives.
>
> (Newmann 1992, p.3)

> [Students being] involved in their work, persist despite challenges and obstacles, and take visible delight in accomplishing their work.
>
> (Schlechty 1994, p.5)

> [A] willingness, need, desire and compulsion to participate in, and be successful in, the learning process promoting higher level thinking for enduring understanding.
>
> (Bomia *et al.*, 1997, p.3)

Clearly student engagement overlaps with both student motivation and achievement but is not the same as either. Many measures of student engagement such as the National Survey of Student Engagement (NSSE) or the College Student Experiences Questionnaire (CSEQ) are focussed on the time a student reports spending on certain types of behaviours that have been observed to be highly correlated with many desirable learning and personal development outcomes (Axelson & Flick 2010). This research seeks to add to that literature by focussing on a more objective measure of student engagement behaviour, and its relationship with exam performance.

The quiz component of the new course provided a range of potential measures of engagement including number of attempts and time spent, on each attempt and in total. These measures were recorded in real time as the student was logged into the quiz software. This overcomes many of the difficulties of students reporting their engagement ex post and often coloured by their performance outcome. Reports of differential reporting by gender where females tend to under-report and males tend to over-report are also overcome (Deaux & Farris 1977). It should be noted that while there is an improvement in terms of the objectivity of the measurement, time spent cannot provide information on the study methods employed or level of engagement during the quiz time. For example we cannot separate out the different quiz strategies discussed by Brothen & Wambach (2001) as 'prepare-gather, feedback-restudy' versus 'quiz to learn'. However, in both cases the time spent on the activity will increase therefore time spent on quizzes can be seen as an all-encompassing measure of engagement.

This research will seek to ascertain the role of engagement on exam performance, and whether students are making efficient use of the online quizzes with respect to the desired learning outcomes. The

range of data we have available allows us to look at exam performance and control for both ability; engagement; and more importantly control for each of them separately and also in a sequential manner (via the biprobit analysis in section 4). Controlling for the underlying ability of the student is of paramount importance. In his keynote address to the *15th Australasian Teaching Economics Conference* (Kennedy 2010), Peter Kennedy pointed out that from his experience as the editor of the statistics section of the *Journal of Economic Education*, the only reliable predictor of student success seems to be their initial ability and controlling for this is important in any analysis of teaching innovation.

3. DATA

We employ data from the first cohort in the newly designed 191 economics course. In addition to the summative assessment results we have very rich data on student quiz participation, including number of attempts on each quiz, time spent on each attempt and marks gained. This course data were then supplemented with matched enrolment data to enable controls for demographic and other individual characteristics of each student such as ability. This process of matching data sets was done by using the student ID number, which was then removed from subsequent data sets, to ensure anonymity under the conditions of the ethics approval. The descriptive statistics and definitions of all key variables are shown in Table 1.

The sample consists of 555 students and is relatively evenly split across gender, with just over 55% of the group being male. Relative to the control group for ethnicity (Pakeha), there are two other groups large enough to enable effective and meaningful analysis; Maori and Pacific Peoples (MaPP), and Asians, who account for approximately 10 and 56 per cent of the sample respectively. The observed underlying ability of the student (regardless of course structure) is measured via their cumulative Grade Point Average (GPA) which indicates a wide range of students in this sample, with GPAs ranging from 0 to 8.75 out of a possible 9.

In terms of the assessment indicators, we have information on all three components of assessment (Tutorials, Quizzes and Test) prior to their exam performance. As a general rule, any student who did not complete at least three of the four assessment items was removed from the sample. This has the potential to upwardly bias the final grade distributions, however, comparison of these distributions with those from previous semesters show that they are similar, and well within

Table 1: Descriptive Statistics

Variable	Description	Mean (St Dev)
Final Exam Mark	Bounded variable: 0 – 50 (Final exam accounted for 50% of assessment)	30.553 (9.005)
Male	Dummy variable: 1 = Male; 0 otherwise	0.557 (0.497)
Age	Age in years at the start of the semester	19.466 (3.479)
MaPP	Dummy variable: 1 = Maori or Pacific Peoples; 0 otherwise	0.101 (0.301)
Asian*	Dummy variable: 1 = Asian; 0 otherwise	0.559 (0.497)
Domestic	Dummy variable: 1 = NZ citizen or Permanent resident in NZ; 0 otherwise	0.786 (0.411)
Bbim	Dummy variable: 1 = Enrolled in Bachelor of Business Information Management degree; 0 otherwise	0.153 (0.360)
Other degree**	Dummy variable: 1 = Enrolled in other degree; 0 otherwise (these are predominantly Bachelors of Arts or Property)	0.218 (0.413)
Ability	Continuous variable of student's cumulative GPA: 0 – 8.75 (0 equates to a D average; 9 equates to an A+ average)	3.575 (2.124)
Papers taken	Number of papers enrolled in semester: Continuous variable (2 – 5)	4.013 (0.303)
Quiz mark	Bounded variable: 0 – 10 (Quizzes accounted for 10% of assessment for the paper)	7.731 (2.139)
Total Time Spent on Quizzes	Number of hours spent on quiz attempts over the entire semester	11.868 (9.177)
Tutorial mark	Bounded variable: 0 – 10 (Tutorials accounted for 10% of assessment for the paper)	8.109 (2.542)
Test mark	Bounded variable: 0 – 30 (The test accounted for 30% of assessment for the paper)	18.764 (4.941)

* Ethnicities other than Maori, Pacific Peoples and Asian serve as the control group (this is predominantly Pakeha); ** Enrolment in a Bachelor of Commerce (BCom) degree serves as the control group.

the Faculty guidelines. It is also important to remember that the objective of the course re-design was not to influence grade distributions *per se*, but to improve critical thinking skills, application and recall of course content, as measured by performance in the final exam.

Assessment Structure

To these ends the assessment structure was developed as an integral part of the entire course. While online quizzes had been previously offered as a formative option in some first year economics papers, the take-up of them tended to be very low once assignment pressure started. This is consistent with the findings of McKeown & MacLean (2010) where they tracked student participation in online quizzes across semesters where there were no grade incentives versus when there were grade incentives. At the same time, Cameron (2010) suggests that even when items carry very low mark weightings, they can act as a great stimulator for student engagement and completion. If the assessment programme was to be used to help teach key skills and capabilities as well as to assess their learning, marks were therefore required to show the priority of all tasks. This need had to be balanced with the unsecured nature of the quizzes and tutorials.

In developing this integrated assessment programme the teaching and learning objectives were explicitly aligned and this structure explained to the students. The structure was also consistent with the explicit capability development goals of each item. Weekly quizzes were designed to encourage mastery of the core ideas and theories. Fortnightly tutorials then assumed this mastery and developed application capability through the analysis of real world problems in a systematic and overt way. The test then expected students to apply this capability to different real world problems they had not seen before. Based on this it would be expected that students that had mastered all three of these components would be much better prepared to succeed in the final exam which was similar in structure to the test.

The weekly quizzes (run through the publisher platform, *My Econ Lab*) were designed to ensure students were understanding the core theory and ideas and able to apply them to simple text book situations (Buckles & Siegfried 2006). Feedback would be automated and immediate on these quizzes. This structure managed the very real resource constraints faced by large first year papers where up to 1,500 students were enrolled per academic year. The five, fortnightly

tutorials were where more complex, real world examples would be addressed, discussed and solutions found using the theory taught and tested through the quizzes. These tutorials were an hour long and had up to thirty students per tutorial. While the preference would have been for weekly tutorials, room and tutor constraints meant this was unrealistic given the size of this course. The mid-term test and final exam would then require students to analyse and apply the material to real world problems that they had not seen in this context before. The details of the programme settled on was:

- 11 weekly quizzes (10 marks):
  - Best ten counted;
  - Unlimited attempts but only the best would count;
  - Assesses key theory and basic understandings;
  - 10 day window to complete each week's test;
- 5 tutorials (10 marks):
  - Applied theory to real world problems;
  - Stepped through process of application but then had to attempt another one their own and hand it in for marking and feedback;
- Mid-semester test (30 marks);
- Final exam (50 marks).

Both the mid-semester and final exam would present students with a mix of recall of key concepts and the application of those concepts to current issues. The weighting between these components was toward the application (70%) with the recall questions being present primarily to build confidence in the time constrained test/exam environment.

4. METHODOLOGY

The first model employed is a linear regression (using ordinary least squares) with exam mark as the dependent variable and various demographic and assessment variables as predictor variables. The exam mark was chosen as the dependent variable in preference to the final course mark (an addition of all assessment items), as it is expected to depend on student's level of engagement, and thus their subsequent ability to recall and apply course material in different contexts.

The key independent variables of interest are tutorial mark, test mark and engagement, as measured via use of weekly quizzes. There were a total of eleven weekly quizzes available for students, each out of 15 marks, however only their top ten quiz marks would count. This

built in redundancy was to minimise issues if students had technical or personal problems that would result in them missing a quiz. Rather than having to deal with these issues the option of the extra test meant that they were not penalised for one bad week. There was no time limit on each quiz attempt but each quiz was only available for ten days. For each quiz there was a quiz bank of approximately 150 questions on that week's material. These questions were a combination of questions taken from the publisher test bank related to the text book and additional questions developed by the teaching team. Each bank of questions was split into 15 pools that matched the topics covered that week. Each randomly generated quiz had one question from each pool. This structure ensured that every quiz generated covered all of the material required for that week. It would not be possible for a student to 'get lucky' and get questions on only one or two of the ideas covered in lectures. In addition to this layer of randomisation, the order of the answers for most of the multi-choice questions would change each time a quiz was generated. Thus, although a student may get a question that they had seen before the order of the possible answers would still be different. Some questions were calculation questions based on algorithms so that the answer would be uniquely different each time the student saw the question.

Due to the structure of the quiz assessment programme with unlimited attempts and the highest mark recorded, quiz marks were very high and less informative of engagement than the underlying activity, namely time spent on quizzes. Following the approach taken by McKeown & MacLean (2010) a more appropriate metric to gauge level of engagement is length of time spent on quizzes. Given that the descriptive statistics presented in Table 1 clearly indicate that this variable is non-normal, we make use of the log transformation of this explanatory variable in all subsequent empirical analysis. A similar transformation is done for tutorial mark, but is not required for the test mark, which follows a close to normal distribution.

The initial method that will be employed in this analysis is linear regression, where the focus will be on the size and significance of the coefficient on the natural log of quiz time [*ln* (quiz time)]. This will provide an indication of the importance of student engagement with respect to exam performance and will be conducted separately for males and females. Previous research indicates that females tend to be more engaged, relative to their male counterparts (Kuh 2010). However, there is other work that suggests that being male is

advantageous in the economics discipline (Agnew 2010; Anderson, Benjamin & Fuss 1994). The gender sub-group analysis undertaken here will determine whether the higher level of engagement by females is significant in terms of its influence on exam performance or whether this is offset by better performance in general in economics by males.

The second empirical model employed in this study is a bivariate probit, which assumes the data takes the sequential format shown in Figure 1. This analysis examines whether there are associations between a range of individual characteristics on engagement (as measured by the natural log of quiz time) and exam performance. A distinctive feature of this analysis is that our methodology allows us to model the role of the control variables on both exam performance and student engagement at the same time.

Figure 1 reveals that 41.62% of the sample performed below average in terms of our indicator for engagement. It appears clear that those that performed above average were more likely to also perform above average in their exam (62.96% versus 49.78%). While not reported in Figure 1, it is important to note that we do find evidence of females being more engaged in their economics course. While 58.38% of the total sample spent longer on quizzes than the average of *ln* (*quiz time*), the comparable figure for females was 67.07%.

The overall scenario presented in Figure 1 involves the analysis of two dependent, and sequential variables to model. This permits marginal effects to be obtained where $P(\textit{Exam mark = 1}|\ \textit{A or B})$, i.e. $P(\textit{Exam mark = 1}|\ \textit{Student engagement = 0})$ and $P(\textit{Exam mark = 1}|\ \textit{Student engagement = 1})$, where one denotes above average, and zero equates to below average. Given these marginal effect estimates from these two conditional probabilities (i.e. comparing route *C* with route *E*) it is possible then to identify whether the drivers of exam performance differ depending on whether student engagement = 1 or 0. This model therefore allows empirical investigation of whether the determinants of above (below) average exam performance are measurably and statistically different, depending on whether the student undertakes above or below average engagement.

More formally, let y_{1i} be a latent variable that denotes the probability that a student has above average exam performance, which is determined by a range of explanatory variables, X_{1i}. Also let y_{2i} be a latent variable for the probability the student has above average engagement asmeasured by *ln* (*quiz time*), which is also determined

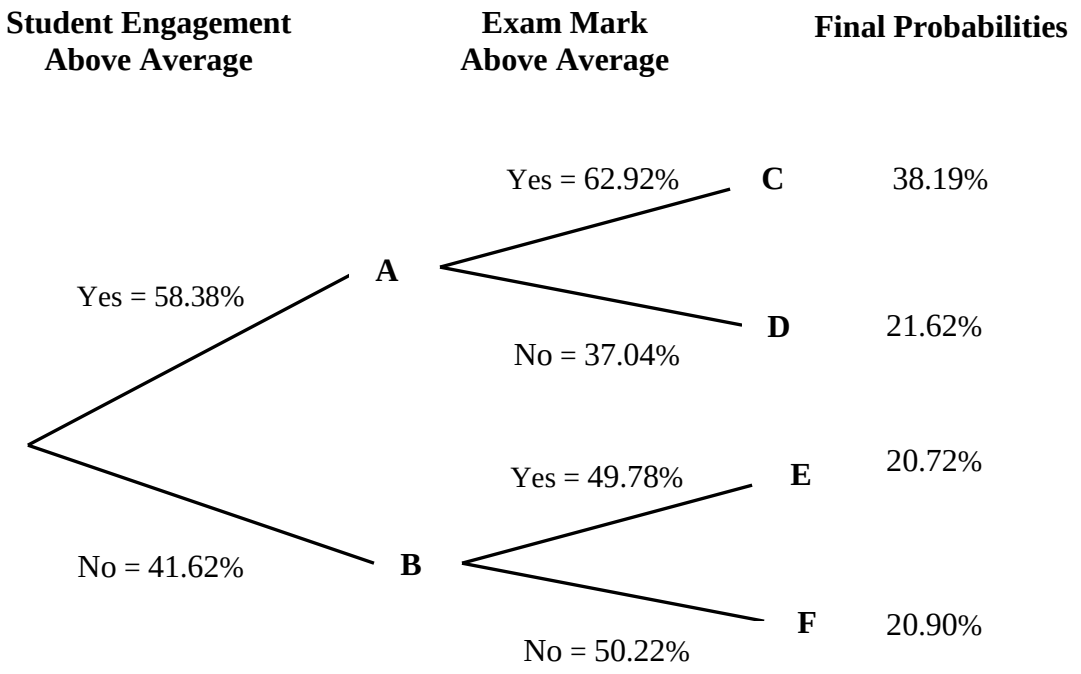

Figure 1: Tree Diagram - Engagement Measured by *ln* (*quiz time*)

by a range of explanatory variables, X_{2i}. The model is represented as follows:

$$y_{1i} = \beta_1 X_{1i} + \varepsilon_{1i} \quad (1)$$

$$y_{2i} = \beta_2 X_{2i} + \varepsilon_{2i} \quad (2)$$

where y_{1i} is observable for all students in the sample and related to the following binary dependent variables, on the basis of the following conditions:

$$\begin{aligned} Exam_i &= 1 \quad if \ y_{1i} > 0 \\ &= 0 \quad if \ y_{1i} < 0 \end{aligned} \quad (3)$$

and

$$\begin{aligned} Quiz_i &= 1 \quad if \ y_{2i} > 0 \\ &= 0 \quad if \ y_{2i} < 0 \end{aligned} \quad (4)$$

where $Exam_i = 1$ denotes that the student has performed above average in their exam and within their cohort, and $Quiz_i = 1$ denotes that the student has above average values for *ln* (*quiz time*). The errors (ε_{1i}, ε_{2i}) are assumed to have the standard bivariate normal distribution, with $E(\varepsilon_{1i}) = 0 = E(\varepsilon_{2i})$. The bivariate probit model has full observability as both $Exam_i$ and $Quiz_i$ are observed in all four of their possible outcomes (i.e. $Exam_i = 1$, $Quiz_i = 1$; $Exam_i = 1$, $Quiz_i = 0$; $Exam_i = 0$,

$Quiz_i = 1$; and $Exam_i = 0$, $Quiz_i = 0$), which leads to the most efficient estimates (Ashford & Sowden 1970; Zellner & Lee 1965).

5. RESULTS

Table 2 presents the results of three linear regression models. In all specifications, individual characteristics and the type of degree the student is enrolled in are included as determining factors behind exam performance. In model (1), our key variable of interest is our measure of engagement [*ln (quiz time)*]. In the subsequent regressions, the model builds to add our other assessment metrics, natural log of tutorial mark and then test mark (models (2) and (3) respectively); in an order the reflects the items role in capability development as discussed in sections 3 and 4.

In general, males appear to have performed better in their 191 exam, relative to females & this result is statistically significant at the 5 per cent level in the first two of the regression models. This finding is consistent with evidence provided by Krieg & Uyar (2001), who also find that males outperform females in their analysis within the economics discipline. At first this result may seem strange, because other studies find that females tend to be more engaged (Kuh 2010), and we would expect that increased engagement would lead to improved exam performance. It is important to note that this link may depend on the nature of the exam. Hickson (2010) found that when exams were in the format of constructed response answers, women did better while males tended to do better when the questions were multi-choice or numerical. This exam was a mix of question types but with an application focus throughout. Therefore it is possible that while males still performed better as found by Krieg and Uyar, this gap may have narrowed through greater engagement by females.

There appears to be a negative relationship between age and exam performance, albeit statistically insignificant. This effect is consistent with other literature that draws attention to the range of reasons that mature students may not perform as well and tend to have a higher drop-out rate than their younger counter-parts. These include: a lack of preparedness for higher education; changing personal circumstances or interest; financial matters; the impact of undertaking paid work; and dissatisfaction with the course or institution (Duff 2004). A future direction of this research could be to control for full-time or part-time status and see if this relationship holds or is mitigated by study mode.

Table 2: Linear Regressions

Variables	(1)	(2)	(3)	(3) Males	(3) Females
Constant	18.137*** (5.471)	15.483*** (5.936)	5.891 (5.077)	1.564 (7.069)	10.843* (6.101)
Male	1.011** (0.504)	1.012** (0.500)	0.719 (0.496)	-	-
Age	-0.127 (0.152)	-0.125 (0.155)	-0.068 (0.134)	-0.080 (0.186)	-0.051 (0.128)
MaPP	0.096 (0.998)	0.527 (0.982)	-0.016 (0.963)	1.569 (1.260)	-1.986 (1.435)
Asian	-0.268 (0.557)	-0.263 (0.559)	-0.979* (0.561)	-0.804 (0.757)	-1.215 (0.837)
Domestic	-1.214* (0.681)	-1.186* (0.677)	-0.366 (0.629)	0.072 (0.888)	-1.104 (0.924)
Bbim	-0.111 (0.773)	-0.028 (0.763)	0.096 (0.708)	-0.574 (0.991)	0.741 (1.052)
Other degree	0.231 (0.590)	0.255 (0.580)	0.536 (0.573)	-0.001 (0.867)	0.906 (0.807)
Ability	3.176*** (0.134)	3.043*** (0.135)	2.112*** (0.214)	1.787*** (0.286)	2.534*** (0.322)
Papers taken	0.661 (0.868)	0.689 (0.882)	1.100 (0.807)	2.248* (1.249)	-0.070 (0.902)
ln(quiz time)	0.622* (0.376)	0.327 (0.371)	0.459 (0.337)	0.312 (0.409)	0.862* (0.521)
ln(tutorial mark)		1.733* (0.994)	1.702* (0.926)	1.509 (1.036)	2.036 (1.586)
Test mark			0.524*** (0.087)	0.636*** (0.115)	0.359*** (0.136)
R squared	0.585***	0.590***	0.590***	0.579***	0.703***

Notes: Reference groups = Female, Pakeha, International students, and BCom degree. N = 555 (309 males and 246 females). Robust standard errors shown in parenthesis. ***, ** and * represent statistical confidence at the 1%, 5% and 10% levels.

Both ethnic minorities (MaPP and Asian) fared slightly worse in their 191 exam, relative to the control group of Pakeha and this is statistically significant at the 10 per cent level for the Asian group in the fullest specification. Interestingly the coefficient is positive for Maori and Pacific Peoples in the second specification. This specification includes discussion and collaborative learning with their peers within the tutorials which has been linked to cultural

learning norms for these ethnic groups pedagogy. The tutorials also provide an greater opportunity for relationships to be developed with tutors that are not always possible in the large lecture formats and are known to support the learning in these groups in particular (Gorinski & Abernethy 2007). Although these results are consistent with other studies they are generally not significant. Perhaps related to these findings is that domestic students do slightly worse relative to international students in all specifications of the model but this is only significant at the 10 per cent level in this first two specifications.

The student's underlying ability (as measured by their cumulative GPA) is a strong and consistent predictor of exam performance, significant at the 1 per cent level in all three specifications. Numerous studies find that underlying ability is a key determinant of future success (Anderson *et al*. 1994; Davies & Guest 2010; Galizzi 2010). In the upcoming empirical analysis, where a bivariate probit model is employed, this study can help untangle the complex relationship between engagement and exam performance, and investigate the extent to which (if any) higher engagement can compensate for lower ability.

Student engagement, measured via quiz time is positive and statistically significant in the first specification. While it seems that the significant influence of quiz time diminishes as other assessment measures are added to the regression (models 2 and 3), when the fullest specification is broken down by gender, we find that engagement remains a significant predictor of exam performance for females. Additionally, these results are qualitatively robust to the use of alternative engagement measures, such as average time spent per quiz, or average number of attempts per quiz.

Bivariate Probit

In the linear regressions contained within Table 2, student engagement is treated as an independent variable associated with exam performance. While that approach may be valid, a possible further perspective is that engagement by a student is a necessary prerequisite step prior to the role that other determinants play with respect to exam performance. It is possible that allowing this sequential process between student engagement and exam performance illustrates whether the student's level of engagement exacerbates or diminishes the relationship between other covariates and exam result. Consequently, the bivariate probit model presents an empirical investigation into whether the determinants of exam performance

Table 3: Coefficient Estimates in Bivariate Probit Model

Variable	(1) Student engagement	(2) Exam performance
Constant	--2.507** (1.010)	-2.229* (1.178)
Male	-0.354*** (0.115)	0.225 (0.142)
Age	0.008 (0.017)	-0.008 (0.020)
MaPP	0.287 (0.220)	0.089 (0.257)
Asian	0.582***(0.132)	0.208 (0.165)
Domestic	-0.102 (0.154)	-0.192 (0.194)
Bbim	-0.192 (0.161)	0.030 (0.174)
Other degree	-0.067 (0.149)	0.203 (0.199)
Ability	0.038 (0.032)	0.711*** (0.062)
Papers taken	0.134 (0.208)	0.122 (0.235)
ln (tutorial mark)	0.909*** (0.208)	-0.183 (0.172)
N	555	
Log pseudo likelihood	-536.321	
Rho	0.022 (0.093)	

Rho suggests positive correlation between regressions (Prob > chi2 = 0.000). Notes: Robust standard errors shown in parenthesis. ***, ** and * represent statistical confidence at the 1%, 5% and 10% levels.

(above or below average result) differ depending on the level of engagement a student experiences in their 191 course. Essentially, this involves estimating equation (1) and (2) together (based on the belief they are related), and then extending this to the sequential form detailed in Figure 1, as the marginal effects of equation (1) are estimated conditional on the results of equation (2).

Application of the bivariate probit model obtains results are presented in Table 3. The first column corresponds to the determinants of above average student engagement (equation (2)). These results are in line with expectations; males are less engaged relative to their female counterparts; Asians are more engaged relative to Pakeha (Kuh 2010); and underlying ability and the other assessment measure of *ln (tutorial mark)* have a positive relationship with student engagement. The second column of Table 3 illustrates the determinants of exam performance, comparing routes *C* and *E* with *D* and *F* on the tree diagram in Figure 1. It appears that the only

Table 4: Marginal Effects

Variables	(1) Exam performance given Engagement = 1	(2) Exam performance given Engagement = 0
Male	0.084* (0.052)	0..085 (0.054)
Age	-0.003 (0.008)	-0.003 (0.007)
MaPP	0.031 (0.091)	0.031 (0.093)
Asian	0.073 (0.062)	0.074 (0.062)
Domestic	-0.068 (0.067)	-0.069 (0.068)
BBim	0.012 (0.063)	0.012 (0.064)
Other degree	0.072 (0.068)	0.074 (0.070)
Ability	0.259***(0.021)	0.263*** (0.022)
Papers taken	0.044 (0.086)	0.044 (0.087)
***ln* (*tutorial mark*)**	-0.071 (0.064)	-0.072 (0.067)

Notes: Robust standard errors shown in parenthesis. ***, ** and * represent statistical confidence at the 1%, 5% and 10% levels.

significant variable is the student's underlying ability after conditioning on factors that are associated with engagement. The biprobit also identifies whether the regressions are related. This is given by the rho, which is statistically significant at the 1% level.

Table 4 presents the estimation of conditional marginal effects of the explanatory variables on exam performance once we've estimated the role of the explanatory variables on student engagement. This is essentially a comparison of routes *C* and *D* on the tree diagram in Figure 1, with routes *E* and *F*, i.e. the determinants of exam performance with above and then below average student engagement. There are two key findings in this table. Firstly, the determinants of exam performance are relatively stable, irrespective of whether or not the student experiences above or below average engagement via their online quizzes. Secondly, is the empirical investigation of whether the determinants of above (below) average exam performance are measurably and statistically different, depending on whether the student undertakes above or below average student engagement. It appears that only underlying individual ability, as observed via their cumulative GPA, remains as a determinant of exam performance once we have conditioned our results on the variables that are associated with engagement. This means that, based on this cohort of

students, when two students have the same level of engagement the only thing that remains strongly associated with their exam mark is their underlying ability. This finding is consistent with extant research by Anderson *et al.* (1994), Duff (2004), Fizel & Johnson (1986), and Shanahan, Foster & Meyer (2008).

When the analysis is repeated by gender, it is interesting that the only difference in terms of determinants of student engagement is that ability is a significant predictor of level of engagement for males, but not for females. In terms of the marginal effects of the explanatory variables on exam performance, after conditioning on the variables that are associated with the level of engagement; there are several key findings by gender. These are shown in Tables 5 and 6. First, in both cases, the determinants of exam performance are stable, irrespective of whether or not the student has above or below average engagement. This is shown by the similarity in the size and significance of the marginal effects in column (1) versus (2) in Table 5, and similarly in Table 6. The second key finding is that while the results for the whole sample, and for males, indicates that ability is the only determinant of exam performance once we have conditioned our results on the variables that are associated with engagement, the same cannot be said for females. In particular, there appear to be three explanatory factors that remain as significant predictors of exam performance after conditioning on the variables that are associated with engagement. These include ability, whether the female was a domestic student, and whether they were enrolled in a BBIM degree (relative to a commerce degree). These findings present as future lines of enquiry, as to why domestic versus international (business information management versus commerce degree) have positive (negative) associations with exam performance only for females.

6. DISCUSSION AND CONCLUSION

This paper has empirically analysed the effectiveness of a constructively aligned course structure that uses assessment explicitly as a learning tool for students. The course was designed to encourage greater student engagement and connection with the course material. After controlling for a range of individual characteristics, linear regressions reveal that engagement (as measured by the natural log of time spent on online quizzes) had a positive relationship with a student's exam performance. This result was statistically significant (albeit at the 10% level) for females.

Table 5: Marginal Effects for Males

Variables	(1) Exam performance given Engagement = 1	(2) Exam performance given Engagement = 0
Age	-0.005 (0.009)	-0.005 (0.009)
MaPP	-0.002 (0.121)	-0.002 (0.121)
Asian	0.077 (0.077)	0.076 (0.075)
Domestic	0.043 (0.091)	0.043 (0.091)
Bbim	-0.132 (0.099)	-0.132 (0.097)
Other degree	0.110 (0.084)	0.110 (0.084)
Ability	0.245***(0.029)	0.244*** (0.028)
Papers taken	-0.023 (0.116)	-0.023 (0.115)
***ln* (*tutorial mark*)**	-0.089 (0.085)	-0.089 (0.086)

Notes: Robust standard errors shown in parenthesis. ***, ** and * represent statistical confidence at the 1%, 5% and 10% levels.

Table 6: Marginal Effects for Females

Variables	(1) Exam performance given Engagement = 1	(2) Exam performance given Engagement = 0
Age	0.009 (0.011)	0.010 (0.012)
MaPP	0.055 (0.140)	0.055 (0.149)
Asian	0.071 (0.112)	0.072 (0.117)
Domestic	-0.233***(0.086)	-0.250***(0.097)
Bbim	0.174** (0.077)	0.186** (0.089)
Other degree	0.060 (0.113)	0.064 (0.120)
Ability	0.301***(0.037)	0.316*** (0.039)
Papers taken	0.153 (0.137)	0.159 (0.141)
***ln* (*tutorial mark*)**	-0.111 (0.089)	-0.123 (0.104)

Notes: Robust standard errors shown in parenthesis. ***, ** and * represent statistical confidence at the 1%, 5% and 10% levels.

This use of bivariate probit modelling to examine the relationship between student engagement and exam performance was a novel extension to the empirical analysis in this paper. The results for the full sample indicated that irrespective of whether the student has above or below average engagement in their 191 course, the marginal effects of all other observed individual characteristics are similar in terms of their relationship with the final exam mark. In particular, only the student's underlying ability stood out as having a significant role with respect to exam performance, once we had conditioned our results on the variables that were associated with level of engagement.

When the bivariate probit modelling was repeated by gender sub-groups, it became clear that there were distinct differences between males and females. In particular, it appears that whether a student is domestic or international, or enrolled in business informational management versus commerce, are significant predictors of exam performance for female students, even after conditioning for factors that are associated with level of engagement. Further research is required to focus on these two factors.

Other avenues for future research include employing alternative measures for ability, which was found to be the standout predictor for exam performance, irrespective of engagement. As future cohorts complete this course it may be possible to combine their data with this cohort in order to use their entry level ranking as the proxy for underlying ability. It was not possible in this research because not enough students had this data. In the New Zealand system there are five different university entry qualifications that can be used. Only two of these generate an entry ranking. Two others are historical entry qualifications that are not directly transferrable to the current standards. A larger sample size of students with entry ranking information could also permit analysis of how good school grades are at predicting success in a student's first year at university.

To conclude, it is useful to note that while the analysis presented in this paper is clearly quantitative in nature, it can also be complemented with qualitative findings. Even though no specific questions were posed to students regarding the new assessment format in their formal university course evaluations, the online weekly quizzes were frequently mentioned in students' response to open-ended questions. For instance, with respect to the question 'What was most helpful to your learning?', approximately 60% of the open-ended positive responses were on the usefulness of the quizzes, 30% on the

tutorials and 10% miscellaneous other. Further commentary on these responses can be found in Hedges (2012).

REFERENCES

Agnew, S. (2010) "What is Happening to Girls Studying Economics in Low Decile Schools?" *Paper presented at the 15th Australasian Teaching Economics Conference*, Hamilton, NZ, 28-29 June.

Anderson, G., Benjamin, D. and Fuss, M. A. (1994) "The Determinants of Success in University Introductory Economics Courses", *Journal of Economic Education*, 25 (2), pp.99-119.

Ashford, J. R. and Sowden, R. R. (1970) "Multivariate Probit Analysis", *Biometrics*, 26 (3), pp.535-546.

Axelson, R. D. and Flick, A. (2010) "Defining Student Engagement", *Change: Magazine of Higher Learning*, 43 (1), pp.38-43.

Biggs, J. (1999) *Teaching for Quality Learning at University*, Buckingham: Society for Research into Higher Education/Open University Press.

Biggs, J. (2003) *Aligning Teaching and Assessment to Curriculum Objectives*, York: Learning and Teaching Support Network.

Bomia, L., Beluzo, L., Demeester, D., Elander, K., Johnson, M. and Sheldon, B. (1997) "The Impact of Teaching Strategies on Intrinsic Motivation", *ERIC Clearinghouse on Elementary and Early Childhood Education*, p.294.

Brothen, T. and Wambach, C. (2001) "Effective Student Use of Computerized Quizzes", *Teaching of Psychology*, 28 (4), pp.292-294.

Buckles, S. and Siegfried, J. J. (2006) "Using Multiple-Choice Questions to Evaluate In-Depth Learning in Economics", *Journal of Economic Education*, 37 (1), pp.48-57.

Cameron, M. (2010) "Economics with Training Wheels: Using Weblogs in Teaching and Assessing Introductory Economics", *Paper presented at the 15th Australasian Teaching Economics Conference*, Hamilton, NZ.

Davies, P. (2003) "Threshold Concepts: How Can We Recognise Them?", in Meyer, J. H. F. and Land, J. H. F. (eds.), *Overcoming Barriers to Student Understanding: Threshold Concepts and Troublesome Knowledge*, London: Routledge, pp.70-84.

Davies, P. and Guest, R. (2010) "What Effect Do We Really Have on Students' Understanding and Attitudes? How Do We Know?", *International Review of Economics Education*, 9 (1), pp.6-9.

Davies, P. and Mangan, J. (2005) "Recognising Threshold Concepts: An Exploration of Different Approaches", *Paper presented at the European Association in Learning and Instruction Conference*, Nicosia, Cyprus.

Davies, P. and Mangan, J. (2008) "Embedding Threshold Concepts: From Theory to Pedagogical Principles to Learning Activities", in Land, R.,

Meyer, J. H. F. and Smith, J. (eds.), *Threshold Concepts within the Discipline*, Rotterdam: Sense Publishers, pp.37-50.

Deaux, K. and Farris, E. (1977) Attributing Causes for One's Own Performance: The Effects of Sex, Norms, and Outcome", *Journal of Research in Personality*, 11 (1), pp.59-72.

Duff, A. (2004) "Understanding Academic Performance and Progression of First-year Accounting and Business Economics Undergraduates: The Role of Approaches to Learning and Prior Academic Achievement", *Accounting Education*, 13 (4), pp. 409-430.

Fizel, J. L. and Johnson, J. D. (1986) "The Effect of Macro/Micro Course Sequencing on Learning and Attitudes in Principles of Economics", *Journal of Economic Education*, 17 (2), p.12.

Galizzi, M. (2010) "An Assessment of the Impact of Online Quizzes and Textbook Resources on Students' Learning", *International Review of Economics Education*, 9 (1), pp.31-43.

Gibbs, G. (1999) "Using Assessment Strategically to Change the Way Students Learn", in Brown, S. and Glasner, A. (eds.), *Assessment Matters in Higher Education*, Oxford: Oxford University Press, pp.41-53.

Gorinski, R. and Abernethy, G. (2007) "Maori Student Retention and Success: Curriculum, Pedagogy and Relationships", in Townsend, T. and Bates, R. (eds.), *Handbook of Teacher Education*, Springer Netherlands, pp. 229-240.

Hedges, M. R. (2011) "Embedding Threshold Concepts: The Use of a Practice-Theory-Practice Cycle", *Paper presented at the Western Economics Association 9th Annual Pacific Rim International Conference*, Brisbane, Australia, 26-29 April.

Hedges, M. R. (2012) "Constructive Alignment: Linking to Business Beyond the Classroom", *Paper presented at the 17th Australasian Teaching Economics Conference*, Gold Coast, Queensland, Australia.

Hickson, S. (2010) "The Impact of Question Format in Principles of Economics Classes: Evidence from New Zealand", *New Zealand Economic Papers*, 44 (3), pp.269-287.

Kennedy, P. (2010) "Common Mistakes Made in Classroom Research", *Paper presented at the 15th Australasian Teaching Economics Conference*, University of Waikato, Hamilton, New Zealand.

Krieg, R. and Uyar, B. (2001) "Student Performance in Business and Economics Statistics: Does Exam Structure Matter?", *Journal of Economics and Finance*, 25 (2), pp.229-241.

Kuh, G. D. (2010) "What We're Learning about Student Engagement from NSSE: Benchmarks for Effective Educational Practices", *Change: Magazine of Higher Learning*, 35 (2), pp.24-32.

Land, R., Cousin, G., Meyer, J. H. F. and Davies, P. (2005) "Threshold Concepts and Troublesome Knowledge: Implications for Course Design and Evaluation", in Rust, C. (ed.), *Improving Student Learning Diversity and Inclusivity*, Oxford: Oxford Centre for Staff Learning and Development.

Land, R. and Meyer, J. H. F. (eds.) (2006) *Overcoming Barriers to Student Understanding*, London: Routledge.

McKeown, P. and MacLean, G. (2010) "Modelling Student Learning Behaviour and Marks in Online Quizzes in Econ 101", *Paper presented at the 15th Australasian Teaching Economics Conference: Frontiers in Economics Teaching*, Waikato University, Hamilton, New Zealand.

Meyer, J. H. F. and Land, R. (2003) *Threshold Concepts and Troublesome Knowledge (1): Linkages to Ways of Thinking and Practising within Disciplines Improving Student Learning - Ten Years On*, Oxford: Oxford Centre for Staff and Learning Development.

Meyer, J. H. F. and Land, R. (2005) "Threshold Concepts and Troublesome Knowledge (2): Epistemological Considerations and a Conceptual Framework for Teaching and Learning", Higher *Education*, 49 (3), p.16.

Newmann, F. (1992) *Student Engagement and Achievement in American Secondary Schools*, Teachers College Press.

Rowntree, D. (1987) *Assessing Students: How Shall we Know Them?* London: Kogan Page.

Schlechty, P. (1994) "Increasing Student Engagement", *Missouri Leadership Academy*.

Shanahan, M., Foster, G., and Meyer, J. H. F. (2008) "Associations Among Prior Acquisition of Threshold Concepts, Learning Dimensions and Examination Performance in First-Year Economics", in Land, R., Meyerm, J. H. F. and Smith, J. (eds.), *Threshold Concepts within the Disciplines*, Rotterdam: Sense Publishers, pp.155-172.

Staffordshire University (2008) *Embedding Threshold Concepts*, available at http://www.staffs.ac.uk/schools/business/iepr/etc/index.htm.

Taylor, R. (2002) *Designing Undergraduate Degree Programmes Handbook for Economics Lecturers*, Bristol: Economics Network.

University of Auckland Business School (2009) *Undergraduate Curriculum Review*, Auckland: University of Auckland.

Zellner, A. and Lee, T. H. (1965) "Joint Estimation of Relationships involving Discrete Random Variables", *Econometrica*, 33 (2), pp.382-394.