Effects of the introduction of an online practice exam utilising Confidence Based Marking

Gavin Buskes
The University of Melbourne
Corresponding Author Email: g.buskes@unimelb.edu.au

CONTEXT
Multiple-choice tests and exams are widely used as means to assess students' knowledge, particularly due to their efficiency for large classes. Their highly structured format also means that they can provide useful statistics to educators. However, the binary nature of the per-question feedback statistics (e.g. correct or incorrect) tend to mask gaps in students' knowledge due to the uncertainty in not knowing how confident students were in selecting a particular alternative. Indeed, assessing the reliability of one's knowledge and inference is a key academic skill which has particularly dire consequences in Engineering when it fails. This paper disseminates the results of introducing a Confidence Based Marking (CBM) online practice exam in a large first year Engineering subject, where students were required to indicate their confidence level for each answer and were provided feedback on the reliability of their assessment of their knowledge.

PURPOSE
This study investigated whether the introduction of a CBM online practice exam enabled students to more accurately judge their knowledge in order to better prepare them for the final exam, resulting in improved academic performance and whether there were any insights gained from differences in performance and confidence measures between different subsets of students.

APPROACH
During semester, an online multiple choice quiz question repository system was made available that required students to indicate the confidence of their answers according to a three-point CBM scale. Having become familiar with the system, students then sat an assessed online multiple choice CBM practice exam in the final week of semester that resembled the format of part of the final exam. Feedback was provided to them on the reliability of their assessment of their knowledge in order to identify their strengths and weaknesses in the subject material. Data was collected on the performance and confidence levels for each student for each question and analysed to identify relationships. Final exam data was collated and related to the online practice quiz data in order to measure any improvement in student performance.

RESULTS
Student feedback indicated that the use of the quiz question repository during semester significantly improved feedback and their confidence over time with the subject material. Results from the online practice exam revealed some differences in performance and confidence levels across particular subsets of students. Students' results in the final exam were correlated with their CBM online practice exam performance and indicated a small improvement in overall results compared to previous years

CONCLUSIONS
CBM assessment has been shown to provide a valuable measure of the reliability of students' knowledge, especially in Medicine. Applying such a methodology to a large Engineering subject, in the form of an online quiz question repository and online practice exam, has provided feedback to students to allow them to better judge their knowledge and improve in the key areas they need to do so. It is anticipated that further refinement of the system will translate to improved retention of knowledge and improved final exam performance.

KEYWORDS
Assessment, knowledge, confidence-based marking.

Introduction

Since their inception over 100 years ago (Goodenough, 1950), tests and exams constructed using Multiple Choice Questions (MCQs) have been widely used as a form of assessment in higher education due to their high level of reliability, versatility, efficiency and ease of marking (Roediger III & Marsh, 2005). Despite the wide usage of MCQs and detailed published guidelines for their construction (Haladyna, Downing, & Rodriguez, 2002), there remain inherent limitations such as the difficulty of detecting the guessing of answers (Burton, 2001), the inability to test higher-level cognitive functions and the lack of opportunity for a student to show the working used to obtain the selected answer in order to obtain partial credit (McAllister & Guidice, 2012).

The basic, most common form of MCQ, the “one correct option” style, has been adapted over the years to spawn a range of types that attempt to address the inherent limitations in such a scheme (Albanese & Sabers, 1988; Rowley & Traub, 1977). Alternative grading schemes have been utilised such as allowing more than one correct answer, weighting answers according to their quality and negative marking for incorrect answers (Siddiqui, Bhavsar, Bhavsar, & Bose, 2016). Despite these measures there is still a perceived unfairness in MCQs (McCoubrie, 2004), where students can still obtain a correct answer through an amount of guessing and process of elimination, rather than through the intended cognitive process being tested.

An alternative method to combat some of the aforementioned shortcomings in the use of MCQs is to require an extra metric to be provided with the response to each question indicating the confidence that the student has in their selected answer. Confidence-Based Marking (CBM), in which a student must indicate their confidence level in each answer and be graded according to a suitably designed marking scheme, helps to encourage reflection, justification and rigour (Gardner-Medwin, 1995). It rewards both justification to the point of high confidence and the ability to identify reasons for reservation about an answer, and therefore encourages a more rigorous approach both to learning and assessment (Bryan & Clegg, 2006). Students’ misconceptions are highlighted when they receive the feedback and the system is perceived as more realistic and fair in that it eliminates a large amount of guessing.

This paper describes the implementation of a CBM MCQ test in a first-year engineering subject through an online practice exam, held in the final week of semester. Feedback was provided to students on their accuracy, measured as the number of correctly answered questions, and the reliability of their assessment of their knowledge, through a CBM-weighted mark and calculated bonus. The goal of the practice exam was to identify individuals’ strengths and weaknesses in the subject material in order to guide their study in the Study Without Teaching Vacation (SWOT Vac) period and during the exam period. Data was collected on the performance and confidence levels for each student for each question and analysed to identify relationships and differences across sub groups of students.

Motivation

ENGR10003 Engineering Systems Design 2 is a first-year, multidisciplinary engineering subject that comprises approximately 800 students in semester 2. This subject is compulsory for most Engineering students and consists of three distinct modules - Digital Systems, Mechanics and Programming. The subject is largely project-based within each module with various opportunities provided for assessment and feedback throughout the semester via assignments, short in-class quizzes and demonstration of project work. However, most of the assessment is completed in a team setting and the feedback provided tends to mask the individual contribution of each team member. Consequently, students were not getting an accurate picture of their individual levels of understanding of the subject material when receiving feedback on their assessments and this was potentially translating into poor exam

performance. Furthermore, with the subject being segmented into three discipline-focused modules, knowledge acquired during the first module may not be retained by the time the exam is sat several months later, and misconceptions that were not dispelled during a particular module may manifest themselves in the final exam. The exam is weighted as 60% of the subject assessment and more importantly, is a hurdle requirement for successfully passing the subject; such misconceptions could prove costly in this context.

To enable students to practice and test their acquired knowledge during semester an online MCQ repository was set up on a server accessible by all students enrolled in the subject and at any time. Students could attempt small quizzes that were constructed from pools of questions, broken into topic areas within each subject module. Some questions were parameterised in order to provide variety and prevent rote learning. Completing quizzes from the repository was optional and did not contribute to the assessment for the subject. The selection of an MCQ format for the quiz system was predicated on the large numbers of students potentially accessing the system and the rapidity and efficiency of automatic marking that MCQs can provide.

In order to take a snapshot of the students' knowledge and provide them with this feedback, they were required to complete an online practice exam worth 5% of the total subject assessment in the final week of semester. Questions were of a similar style to the repository questions and the format of the practice exam closely resembled the actual end of semester exam, which is comprised of 40 one correct answer MCQs (40%) and several short answer questions (60%). Completing past exam papers is a relatively common observed method of study for students, however most tend not to complete them under strict exam conditions, for example in study groups. Furthermore, as a matter of policy the School of Engineering does not provide solutions to past exam papers which may shape students' study habits with past exams. Students were encouraged to complete the practice exam under realistic exam-like conditions and would receive direct feedback, obviating the need to providing solutions to past exams. In addition, becoming accustomed to the format, length and difficulty of past exam papers acts as a means for students preparing themselves for the final exam. The added benefit is that taking a test generally improves students' performance on a later test; this is referred to as the "testing effect" (Kuo & Hirshman, 1996)

Confidence Based Marking

To alleviate some of the inherent shortcomings of MCQs and increase the quality of the feedback provided to the students heading into the exam study period, the online practice exam implemented CBM. When using CBM with MCQs, students are asked to state with each answer their level of confidence, *C*, in the correctness of their decision – *C* = 1 (Low), 2 (Medium), or 3 (High). If the answer is correct, then this is the score awarded. An incorrect answer leads to a score of zero for level 1, and -2 or -6 for levels 2 and 3 respectively. Level 2 gives equally weighted negative marking for wrong answers. Students are told to choose level 2 unless they are very confident (>80% chance of being right), when they should choose level 3, or rather hesitant (<67% chance of being correct), when level 1 is appropriate. This strategy is optimal to maximise their expected scores. The expected average CBM mark for a given percentage correct, or accuracy level, for the three confidence levels is given in Figure 1. A student can never expect to gain systematically by either overestimating or underestimating confidence. High marks are gained firstly, of course, by getting the answer correct.

A bonus is added to the simple accuracy as a measure of how well the student categorises responses as uncertain or reliable (Bryan & Clegg, 2006). The bonus is positive or negative, proportional to the amount the average CBM mark is above (or below) the average that would be obtained if the student had used the same optimal C level for all of their answers as shown in Figure 1. Negative bonuses are common in self-tests when students often have misconceptions (confident errors), but students should aspire to gaining positive bonuses of 2-5%

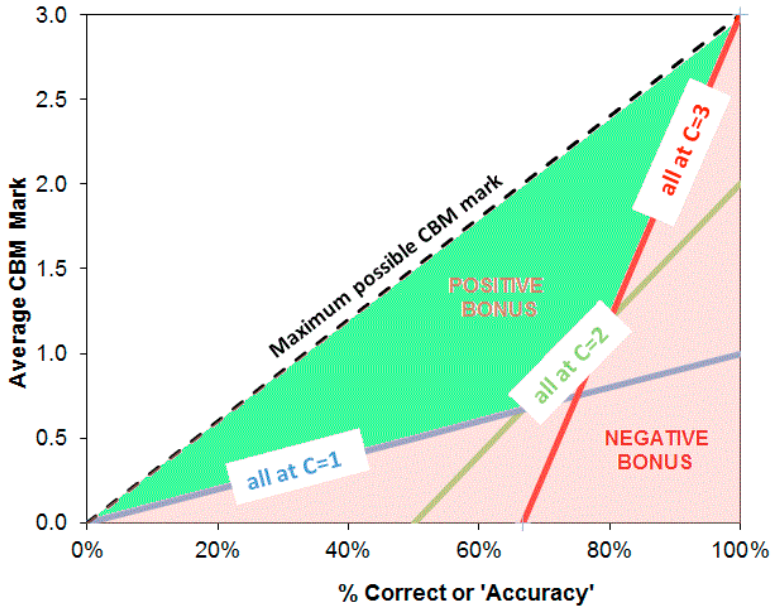

Figure 1 : Average CBM Mark versus Accuracy

Implementation of the Online Practice Exam

The online practice exam was designed to closely resemble the MCQ section of the final exam, which comprised 40% of the total final exam mark. Like the final exam, there were 40 questions, each with one correct answer and worth one mark each, selected from five possible alternatives. The practice exam was run in the final week of semester with no limits on the time taken to complete it within this time frame. Students could review their choices and change their answers at any time until it closed. In order to reduce instances of collusion in completing the practice exam, the ordering of questions and answer alternatives was randomised and some questions contained variables that were parameterised.

After the practice exam was closed, students were provided with feedback that comprised of an overall accuracy score, an average CBM mark, a bonus-adjusted accuracy score and the correct answers to all of the questions indicating their selections. The bonus-adjusted accuracy score comprised 5% of their final subject assessment.

Results

Overall Practice Exam Performance

Figure 2 shows the performance of the 751 students that fully completed the online practice exam (defined as having answered all 40 questions). The dashed line indicates the maximum possible CBM score for a given accuracy level, which would be obtained if students (unrealistically) selected high confidence for every correct question and low confidence for every incorrect question. The solid line indicates the average CBM score that would be obtained if the student had used the same optimal confidence level for all of their answers. Accuracy scores would receive a bonus proportional to the amount the average CBM mark is above (or below) this line. It is clear that most of the CBM scores are above this line, although there are several extreme outliers.

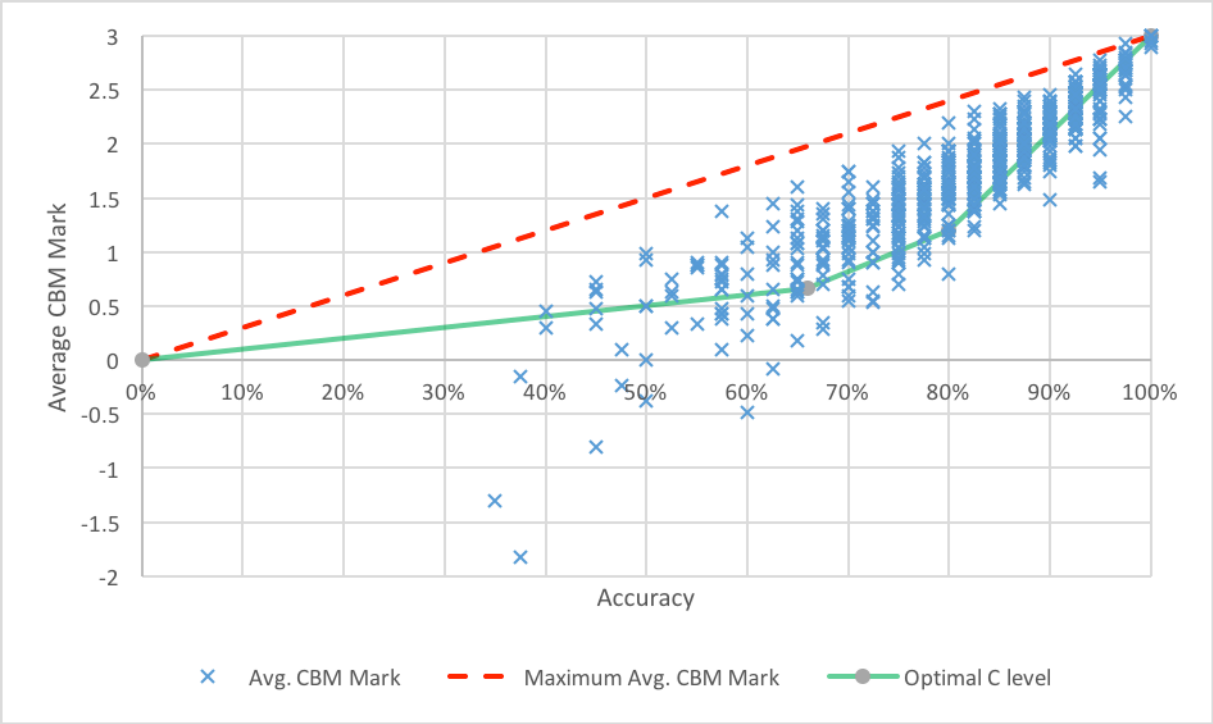

Figure 2 : Practice Exam Average CBM Mark (*n* = 751)

Figure 3 shows the CBM adjusted accuracy score, where it is clear that most students received a positive bonus (point lies above the dashed line), indicating that to a large degree their confidence levels reflected their accuracy.

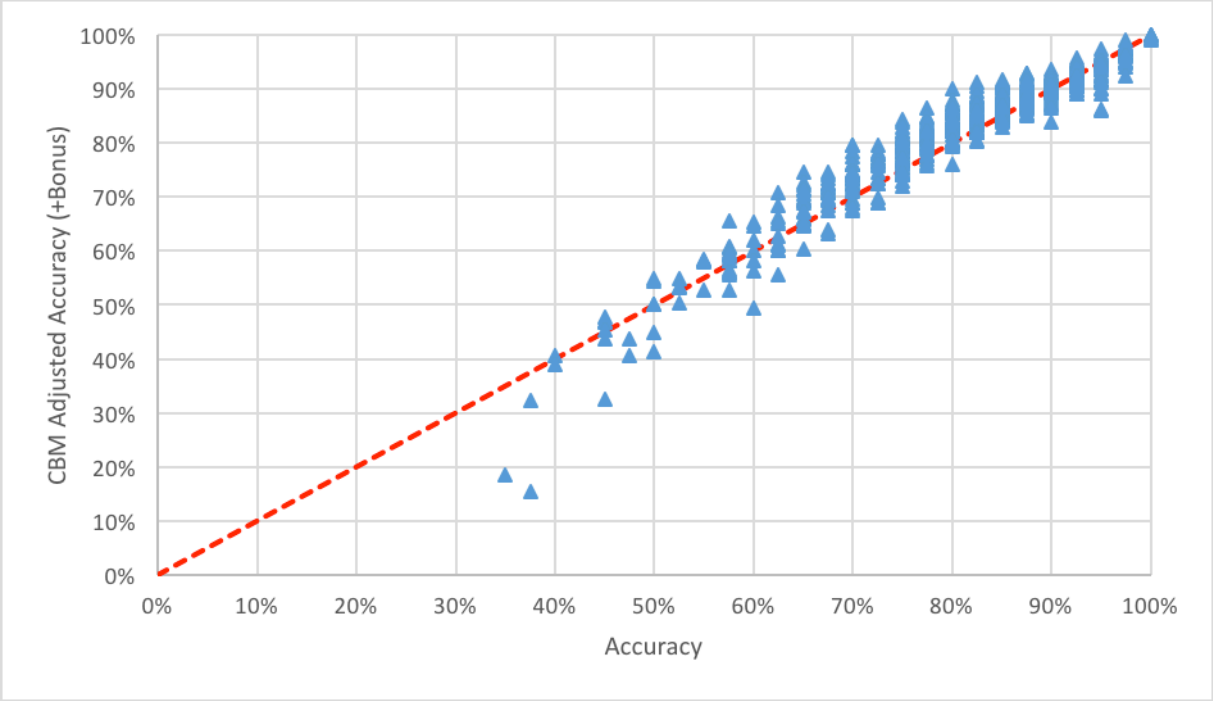

Figure 3 : Practice Exam CBM Adjusted Accuracy (*n* = 751)

Practice Exam – Comparisions between sub groups

Gender performance

The results for the average CBM score and accuracy were split according to the gender of the students and are given in Table 1. Note that while males performed slightly better than females in terms of the mean of both the average CBM score and the accuracy, there appears to be a significant difference in the variance of the two groups. In particular, Bartlett's test for equal variances was performed on both sets of data, with $\chi^2 = 8.19$, $p < 0.01$ for the average CBM score and $\chi^2 = 9.55$, $p < 0.01$ for the accuracy score, indicating that the difference in variances is statistically significant. Interestingly, a *Z*-test of the means of the two groups indicated no statistical significance in their difference for both average CBM score and accuracy, to a standard significance level of 0.05.

Table 1 : Practice exam statistics for male (*n* = 594) and female (*n* = 157) students

	Average CBM Score		Accuracy	
	Male	**Female**	**Male**	**Female**
Mean	1.7640	1.6653	0.8202	0.8025
Variance	0.3775	0.5357	0.0124	0.0180

Table 2 shows the accuracy versus the confidence level according to gender. Males marginally outperform females at all confidence levels with the largest difference of 3% being observed at the C = 1 (Low) confidence level. Chi-squared tests indicated that these differences were not statistically significant.

Table 2 : Practice exam accuracy for males and females at each confidence level

	Percent correct (accuracy)		Significance (df = 1)
Confidence Level	**Male (*n* = 594)**	**Female (*n* = 157)**	
C = 1 (Low)	47%	44%	$\chi^2 = 2.52$, $p = 0.11$
C = 2 (Medium)	72%	70%	$\chi^2 = 0.89$, $p = 0.35$
C = 3 (High)	91%	90%	$\chi^2 = 2.15$, $p = 0.14$

Local and International student performance

The results for the average CBM score and accuracy were split according to the categorisation of students being either local or international and are given in Table 3.

Table 3 : Practice exam statistics for local (*n* = 457) and international (*n* = 294) students

	Average CBM Score		Accuracy	
	Local	**International**	**Local**	**International**
Mean	1.7081	1.7982	0.8051	0.8342
Variance	0.4151	0.4024	0.0155	0.0101

Note that the means of both measures for international students are higher than the means of the local students. A *Z*-test was performed on the average CBM score, indicating no statistically significant difference in the means. A *Z*-test was also performed on the accuracy

(Z=3.520, p < 0.001) indicating that the difference in means is statistically significant. So while there is a significant difference in accuracy between local and international students, when the average CBM score is calculated, the significance of this difference has disappeared. This could imply that local students better estimated their level of confidence.

Final Exam Performance

Of the 751 students that completed the online practice exam, 744 sat the final exam. These students' Part A exam performance (measured as an accuracy percentage) versus their average CBM score is shown in Figure 4. A linear regression was performed (R^2= 0.3033) indicated by the dotted linear trend line. This appears to indicate a trend that performance in the practice exam average CBM score is a basic indicator of final Part A exam score.

In order to test the predictive quality of the practice exam for performance on the final exam, the correlation between the Part A (MCQ) performance on the final exam versus each of the practice exam accuracy score, practice exam average CBM score and continuous assessment total was evaluated. The results are given in Table 4, indicating that the practice exam average CBM score was the better predictor of final exam Part A performance – more so than the practice exam accuracy score. The continuous assessment performance was the least accurate measure of final exam performance.

Table 4 : Comparing effects of assessments on final Part A exam score (n = 744)

Assessment Item	Correlation with Part A exam score
Continuous assessment	0.5055
Practice exam accuracy	0.5354
Practice exam average CBM score	0.5507

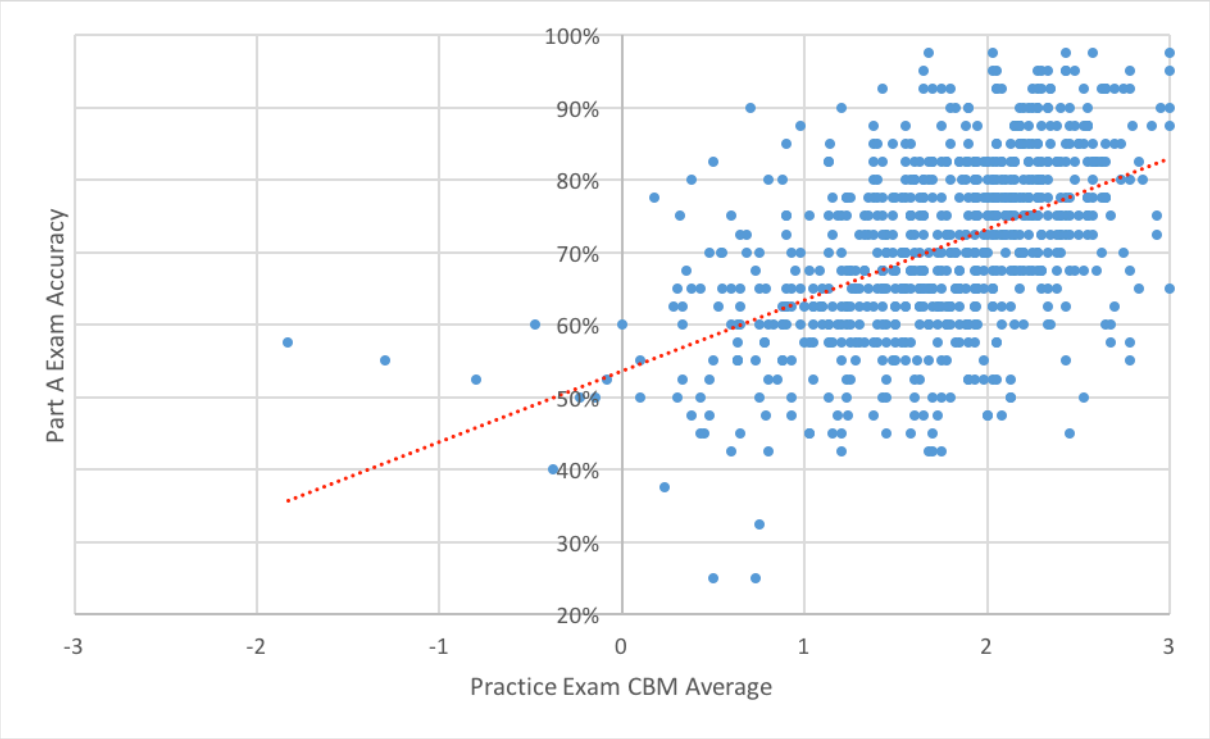

Figure 4 : Part A Exam Accuracy – linear regression trend line (dotted) (R^2 = 0.3033, n = 744)

To measure the effect of introducing the online practice exam, the Part A exam performance of students was compared with the previous two years. The results are given in Table 5. While appearing to indicate a slight improvement in the mean for 2015, a *Z*-test indicates that this was not statistically significant to an alpha level of 0.05 over both 2013 and 2014 results.

Table 5 : 2013-2015 Part A Exam Results (score out of 40)

	2013 (*n* = 725)	2014 (*n* = 754)	2015 (*n* = 860)
Mean	28.046	27.969	28.167
Variance	25.640	25.893	25.876

Student feedback

The end of semester Subject Experience Survey (SES) presented a statement, Q7, "Focusing on my own learning in this subject, I received valuable feedback on my progress", that students responded to using a 5-point Likert scale, with 5 representing "strongly agree". These results are compared for the past four years in Table 6. A student *t*-test indicated a statistically significant ($p < 0.01$) increase in the mean from 2014 to 2015, when the online practice exam was introduced.

Table 6 : SES question on receiving feedback (Q7)

	2012	2013	2014	2015
Mean	3.5	3.5	3.5	3.9
Std Dev	1.05	1.08	1.01	0.92
Responses	268	268	316	345

Discussion

While the implementation of the online practice exam appeared to show a slight improvement on the final exam Part A mean scores over 2013 and 2014, there was not a high enough level of statistical significance to definitively assert this. This could be due to the fact that for this first implementation of the practice exam, students only received feedback in the form of the correct answers and their average CBM score. They did not receive worked solutions, nor any guidance on what to revise or where to find the relevant theory if they did not get a particular question correct. It is felt that a more comprehensive feedback system incorporating such elements will better guide students to tailor their study and consequently perform better in the final exam.

The online practice exam was open for an entire week with unlimited time to complete it within this period. It is unclear how many students approached the practice exam as a true test of their knowledge under exam-like conditions as suggested, or decided to try to maximise their marks by utilising resources such as text books and lecture notes to achieve a high accuracy score. This would obviously affect the assigning of a confidence level to each question and potentially also the predictive quality of the practice exam with respect to performance on the final exam. For the next iteration of the subject, a time limit will apply to the practice exam in order to force students to judge their confidence level under more exam-like conditions.

Student feedback on the online practice exam, as exhibited on the end of semester SES, was generally positive. Many students indicated that they considered it a valuable assessment and feedback tool and liked the freedom in being able to complete it in their own time. A more in-depth survey about how the feedback received from the practice exam was

used by the students to shape their study programme would provide further insight into its effectiveness.

Conclusion

The implementation of an online MCQ practice exam utilising CBM was introduced to allow students to critically assess their knowledge in certain topic areas in order to better focus their study efforts in preparation for the final exam. The practice exam has clearly improved the level of feedback for students heading in to the SWOT Vac and the exam period as evidenced by the Q7 SES results. Overall exam results showed a slight improvement over previous years, although it was not deemed statistically significant to a high enough level. This improvement in exam performance might be increased via the delivery of more detailed feedback upon completing the practice exam. Some interesting differences in sub groups of students have also been highlighted, in particular the variance in practice exam scores between genders and the performance of international students versus local ones.

References

Albanese, M. A., & Sabers, D. L. (1988). Multiple True-False Items: A Study of Interitem Correlations, Scoring Alternatives, and Reliability Estimation. *Journal of Educational Measurement, 25*(2), 111-123.

Bryan, C., & Clegg, K. (2006). *Innovative Assessment in Higher Education*: Routledge.

Burton, R. (2001). Quantifying the Effects of Chance in Multiple Choice and True/False Tests: Question selection and guessing of answers. *Assessment and Evaluation in Higher Education, 26*(1), 41-50.

Gardner-Medwin, A. R. (1995). Confidence assessment in the teaching of basic science. *Research in Learning Technology; Vol 3, No 1 (1995)*.

Goodenough, F. L. (1950). Edward Lee Thorndike: 1874-1949. *The American Journal of Psychology, 63*(2), 291-301.

Haladyna, T., Downing, S., & Rodriguez, M. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied measurement in education, 15*(3), 309-333.

Kuo, T.-M., & Hirshman, E. (1996). Investigations of the Testing Effect. *The American Journal of Psychology, 109*(3), 451-464. doi:10.2307/1423016

McAllister, D., & Guidice, R. M. (2012). This Is Only a Test: A Machine-Graded Improvement to the Multiple-Choice and True-False Examination. *Teaching in Higher Education, 17*(2), 193-207.

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical Teacher, 26*(8), 709-712.

Roediger III, H., & Marsh, E. (2005). The Positive and Negative Consequences of Multiple-Choice Testing. *Journal of experimental psychology. Learning, memory, and cognition, 31*(5), 1155-1159.

Rowley, G. L., & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement, 14*(1), 15-22.

Siddiqui, N. I., Bhavsar, V. H., Bhavsar, A. V., & Bose, S. (2016). Contemplation on marking scheme for Type X multiple choice questions, and an illustration of a practically applicable scheme. *Indian Journal of Pharmacology, 48*(2), 114-121. doi:10.4103/0253-7613.178836