MASTERY OF LEARNING - DOES IT MAKE A DIFFERENCE TO STUDENTS' ONLINE ENGAGEMENT AND PERFORMANCE IN A FIRST-YEAR STATISTICS UNIT?

Karol Binkowski
Macquarie University, Australia
karol.binkowski@mq.edu.au

A large enrolment introductory statistical unit was developed within a Mastery Learning educational philosophy framework. The method states that students must demonstrate mastery of a concept before moving on to the next topic. The unit captures this by an online Learning Management System statistical quizzes structure, with random seed-generated questions in R and practice quizzes before the real one, where students need to max out scores to move on to more advanced tests. The focus is on individualised learning, and students progress through the material at their own pace. We study results across several semesters, relationships between various attempts, scores, final marks, and the online engagements of students in the unit, including time spent watching pre-recorded lecture notes. We identify elements in the assessments and rules around re-taking failed or missed quizzes that we can change. The result tells us the design's strengths and weaknesses throughout several sessions.

INTRODUCTION

With an average enrolment exceeding a thousand students of two sessions each year, the Introductory Statistics unit is a collection of four cohorts with a composition of three undergraduate cohorts and one postgraduate, with a slightly varying assessment structure.

The Faculty of Science and Engineering offers a generalist degree *Bachelor of Science*, with several majors run by departments and schools across the university. The most common courses are *Bachelor of Congnitive-Brain Sciences*, *Cyber Security*, *Human Sciences*, *Information Technology*, *Medical Sciences*, and *Bachelor of Science*, or a dual degree of some of those. We also have a number of external students from *the Open University Australia Non-Award* course. Among the postgraduates, the one with the highest number of students is *the Master of Data Science*.

The unit is a core requirement in the degree and the majors. Bilgin et al. (2020) provide more background information and the history of this and other first-year statistics service units. The unit comprises five modules with no final exam and is considered to have a high online student engagement rate, as evidenced by Thurn (2023).

Since the early work of Bloom (1968), there have been many implementations of Mastery Learning for STEM courses. Recently Perez & Verdin (2022) conducted a thorough review of undergraduate engineering courses, including various assessments, such as quizzes, homework, assignments, projects and final exams. In this paper, we described the Mastery Learning online unit and its quiz-based assessment structure and how effectively students use provided support in the form of practice tests and pre-recorded content in their study. The primary objective of this paper is to tackle the challenge of handling large datasets and transforming them into meaningful metrics to investigate the impact of engagement on students' performance in assessment tasks within a fully online unit.

Sabbag & Frame (2021) conducted a similar analysis of Moodle log data in a randomization-based introductory statistics course for small cohorts of undergraduate social science students. Their findings suggest a connection between student achievement and using course resources, particularly video-related materials.

Concept and implementation of the unit

A former Introductory Statistics unit with over a thousand students had traditionally suffered from high failure rates primarily due to the final exam performance. To alleviate this problem, an aim was formulated for students to master the fundamentals before moving on to more complex questions and to ensure that their grades reflected the attainment level. New assessment questions were crafted considering Bloom's cognitive taxonomy (1968), progressively increasing complexity to challenge students with more demanding concepts. The approach involved building upon a foundation of knowledge, focusing on progressing sequentially through topics and ensuring mastery of each one before moving on to the next.

In: EM Jones (Ed.), *Fostering Learning of Statistics and Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
July 2023, Toronto, Canada.

Unit Structure

The Introductory Statistics unit is structured into five topics, each covered in two-week intervals. Each topic is assessed by three individualised quizzes that add up to 100 marks for each topic, ranking from pass-level questions up to two more difficulty levels. A student must get full marks in a lower-level quiz before moving to higher levels. There are three attempts for assessment quizzes but unlimited practice quizzes. The quizzes contain a large number of question-answer scenarios in the question bank that are individually randomly chosen each time: no student receives the same quiz. All quiz results are automatically marked. The practice quizzes act as learning tools, allowing students to test their understanding before taking the graded quizzes. Students are required to pass every one of the five topics in order to pass the unit. Altogether the assessment structure comprises 15 quizzes. There are also 9 non-graded quizzes called Lecture Participation quizzes (LPQ), related to the lecture's content in a given week.

The assessment was created using the R language (RMarkdown), based on a set of scenarios devised, and a large set of questions was generated, on which individualised quizzes are based. The RMarkdown was compiled into XML files and uploaded to our Learning Management System (LMS) based on Moodle, called iLearn. Following the initial extensive setup work, the quizzes are relatively low maintenance, with automatic grading.

An automated feedback mechanism is built into the quiz page, where students can compare their answers with model answers after the due date of each module test. We do not provide instant feedback, however, we have moderated student forum for students to ask questions if they encounter any difficulties. Reasonable adjustments, such as time extensions for tests based on declared disabilities, are accommodated, ensuring individualised quiz duration (default is 30 minutes). Each module is worth 20% of the final mark, and as a result, these modules can be seen as a formative assessment.

The unit acts as a foundation unit for *the Bachelor of Science* and has an additional employability skills component that spans three weeks of self-study and includes seven hurdle non-graded quizzes. Those quizzes do not cover statistical content, so our analysis has excluded their results.

The emergence of Covid-19 and the ensuing lockdowns have accelerated the advancement of online delivery modes. Le (2022) studied the efficacy of pre-recorded and live online lectures in improving student performance. Our unit adopts a hybrid approach by supplementing pre-recorded content with live online Questions and Answers (Q&A) sessions, enhancing the learning experience. We encourage students to get familiar with the material and prepare questions for our interactive Q&A lecture. Students have on-campus or online Small Group Teaching Activities (SGTA) where concepts are practised and Practical classes in computer labs where students apply statistical concepts to datasets with samples in Excel.

DATA AND METHODOLOGY

For our analysis, we employ simple linear regressions and visual scatterplot assessments of various metric pairs obtained from the LMS based on student performance and engagement. The metrics under investigation include both raw data, such as quiz scores or final marks, and derived data, for example, the number of attempts, average time spent on taking the attempts, average marks achieved on practice basic quizzes, and time spent on watching the pre-recorded lectures.

The first session usually sees an enrollment of over one thousand students, while the second session's enrollment is slightly below that number. In Session 2 2022, the unit was undertaken by four cohorts with a composition of undergraduate STAT1170 (80%, 706 students), FOSE1015 (8%, 73), FOSX1015 (5%, 44), and postgraduate STAT6170 (7%, 62).

The left panel of Figure 1 displays courses with student representation exceeding three per cent, color-coded according to the average overall mark achieved in each course. The list of all possible courses and double degree combinations is extensive, with a relatively low number of students representing each.

Preparation and anonymisation of data for analysis.

This study received ethical approval and a waiver of consent to use grade book de-identified data. The dataset utilised for analysis was constructed from three distinct sources: the grade book, quiz attempt records, and video content logbooks. The unit grade book initially comprises students' personal identifiers, such as IDs, names, email addresses, and quiz attempt scores. The video content logbook is a collection of time each student watched each of the provided pre-recorded material.

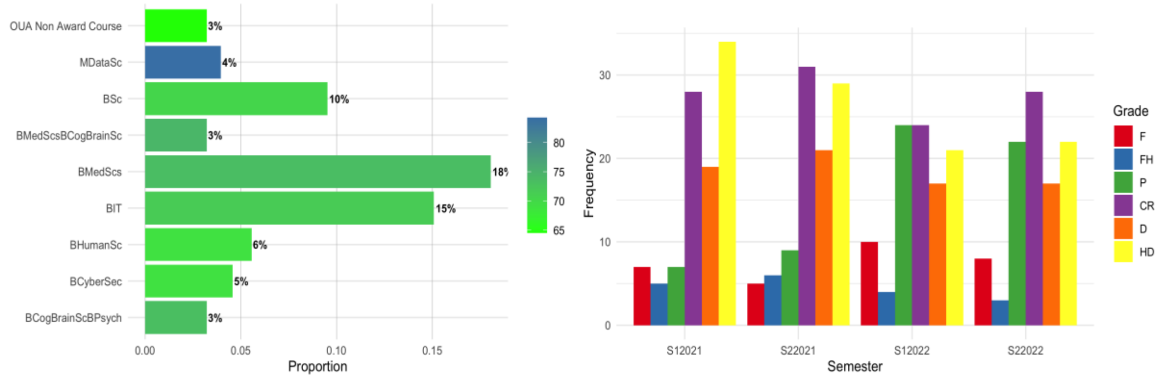

Figure 1: Left: The bars are colored based on the average overall mark for each corresponding course. Right: Grade distribution of the largest courses in the unit over sessions S1 and S2 during the 2021-2022 academic year, F 0-50%, P 50-64%, CR 65-74%, D 75-84%, HD 85-100% and FH.

Students' personal information was removed to preserve anonymity, and only the quiz and final scores were retained for analysis. Any students who failed to complete all basic quizzes and thus did not meet the course hurdle requirements were subsequently excluded from the dataset. The LMS gathers data on every student's attempt at the unit quizzes. Using this data, we matched each student with their number of attempts on each module's basic practice quiz and their average grade and time taken for each quiz attempt. From these metrics, we then derived cumulative metrics across all five modules. Initially, there were 884 records in the grade book. However, we removed students who did not take all the required quizzes in the unit, including those who failed a hurdle or did not log in to the LMS system at all; the number decreased by less than ten per cent to 809. The decision about removing students from the sample for incomplete participation may introduce some bias, as we potentially lose information.

We distinguish two slightly different datasets in the analysis. The first consists of a subset of students who were "committed" to watching the video content. The second dataset is the entire group of students, where the video-watching time was not considered.

RESULTS

In this section, we first present the changes in grade distributions over the past four sessions. We then explore the time spent by students watching pre-recorded content and examine the impact of basic practice tests on students' performance.

Changes in grade distribution across multiple sessions.

The figure 1 (right) illustrates a comparison of grade distributions between four successive sessions, denoted as S1 and S2, for 2021 and 2022. The university final grades consist of Fail (F, less than 50%), Pass (P, 50-64), Credit(CR, 65-74), Distinction (D, 75-84), High Distinction (HD, 85-100), and Fail Hurdle (FH). When FH is applied, the final mark is reduced to 49 marks.

The reflection reports were presented and deliberated upon completion of each session during the examiners' meeting. Both 2021 sessions were run online, with purely online Q&A lecture sessions. Some on-campus classes were opened in S2 2021. Session S1 2022 started with a couple of changes. A new pre-recorded content was added, introductory studio-recorded videos were introduced each week, and the lecture slides walkthrough was divided into shorter videos. The lecture Q&A sessions were back on-campus theatre. The distribution of marks for Basic-Intermediate-Advanced tests was adjusted from 64-20-16 to 60-20-20, to reduce the number of High Distinction grades. The decision to implement this change was based on observations that a high proportion of High Distinction grades were inconsistent with students' previous performance, as indicated by their Weighted Average Marks (WAM). The adjustment decreased the percentage of HDs awarded from 34% and 29% in the previous years to 21-22% in the following year. As a side effect, there was a slight increase in the proportion of students who failed.

There has been an effort to support students and, in hindsight, to decrease F and FH rates. We used to send an individualised bulk email reminder about the basic hurdle quiz. In 2022, the university introduced an app called MyLearn, which provides live information to enable students to monitor their assessment tasks. That helped students keep track of all the quizzes in the unit. As of session S2 2022, we removed a hurdle of Lecture Participation quizzes, which helped to diminish the number of Failed Hurdle students. We also introduced a second chance on the hurdle basic test for any student who missed it or failed, in the form of an automatic extension of one week.

How much time did students spend on watching pre-recorded content?

Figure 2 (top-left) may indicate that a considerable amount of time spent watching the pre-recorded content guarantees a “baseline” (red-line) overall mark. For example, students who watched for over 15 hours would not receive less than 75 final marks. However, this trend is weak.

Figure 2 (top-left) revealed a sizable cluster of students who had watched very little pre-recorded content and still achieved high scores, suggesting they may have learned effectively through practice and reading the lecture notes alone. As a result, we excluded this group of students from the analysis focused on time spent watching content. With 11 hours of pre-recorded content spread across 40 videos of varying lengths, we set a threshold of 1 hour. The exclusion of this group reduced the dataset from 809 to 313 students. This decision was made based on two observations. First, we obtained relatively higher R-squared statistics when regressing the final marks on time watched without including students who watched for less than one hour. The second is that for those students, the overall marks span the entire range, especially between 50 and 100 marks in an inconsistent manner. Considering that the number of attempts on the basic practice test is negatively correlated with the final mark (Figure 3 (top-left)), we conjectured that the large fluctuation of marks is due to students relying on a trial-and-error approach to passing the basic practice quizzes.

Following this, we explored students enrolled in various non-statistical courses and their online engagement. We chose the most represented courses in that dataset by setting a threshold above 3 per cent. Based on the scatterplot in Figure 2 (top-right), we noticed that, on average, courses vary in the amount of time students spend watching pre-recorded lectures and that there is a weak positive correlation between time watched and final marks for certain courses. Specifically, students who enrolled in the double course *Bachelor of Cognitive Sciences* and *Bachelor of Psychology* and the postgraduate students in *Master of Data Science* tended to watch the content for the longest duration, on average.

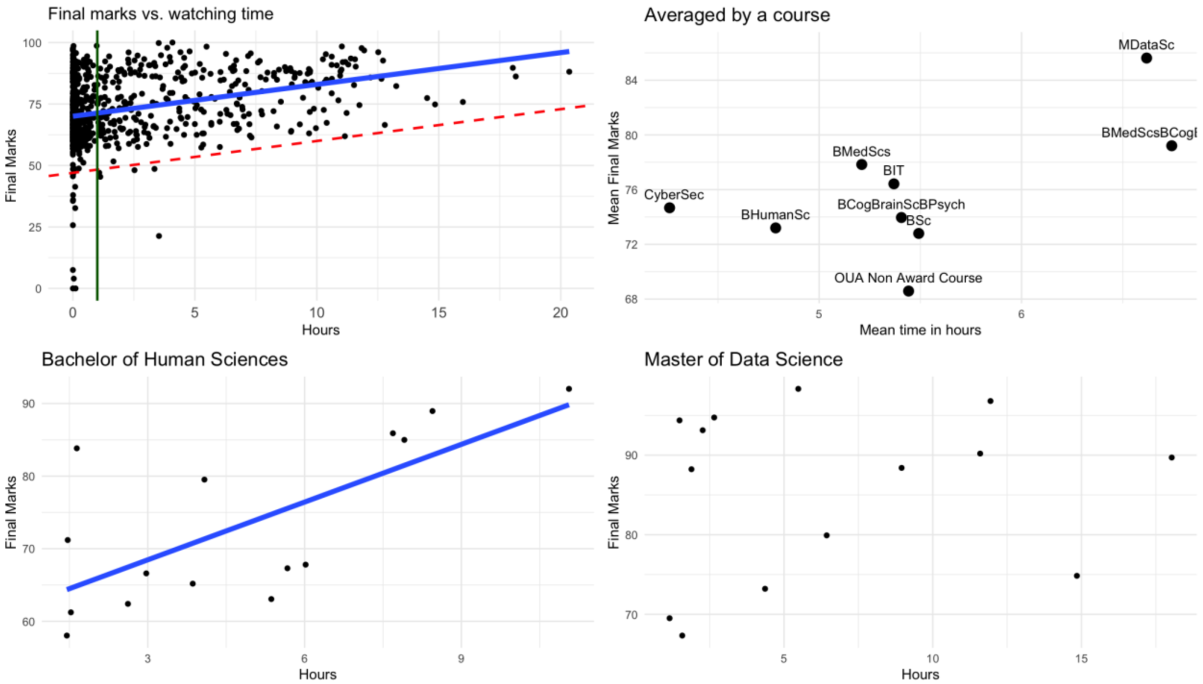

Figure 2: Scatterplots of pre-recorded content watched vs final marks.

In the bottom-left plot of Figure 2, we found a correlation between the overall marks and the time spent watching pre-recorded lectures for *Bachelor of Human Sciences* students, with an R-squared value of 0.49. On the other hand, in the bottom-right plot, we observed a random pattern for *Master of Data Science* students, suggesting no significant relationship between the time watched and the final marks. The observation made here is that the performance of those students was slightly higher on average compared to other courses, which aligns with the notion that postgraduate students in co-taught units tend to excel more than undergraduate students.

How do basic practice tests contribute to students' performance?

We calculated the total number of attempts made by each student across all five practice tests, the average mark achieved on all practice basic tests, and the average time taken to complete each attempt for each module. We used all 809 students for this analysis and ignore the time spent watching the content. Our focus is solely on the metrics pertaining to the basic practice tests and the final marks.

In Figure 3 (top-left), we noticed a negative correlation between the number of attempts made on the basic practice tests and the final marks. This finding is counterintuitive and suggests that a group of students tried to pass the test by guessing rather than actively learning.

The time between attempts on practice tests was set to zero. That means that students could retake practice quizzes straight away. To rectify this, we have implemented a time delay in the current session, and students need to wait five minutes before re-attempting. Also, we note that we do not offer practice quizzes for intermediate and advanced tests to prevent guesswork.

A positive linear trend exists between the average marks achieved on the basic practice tests across all modules and the final marks, as shown in Figure 3 (top-right).

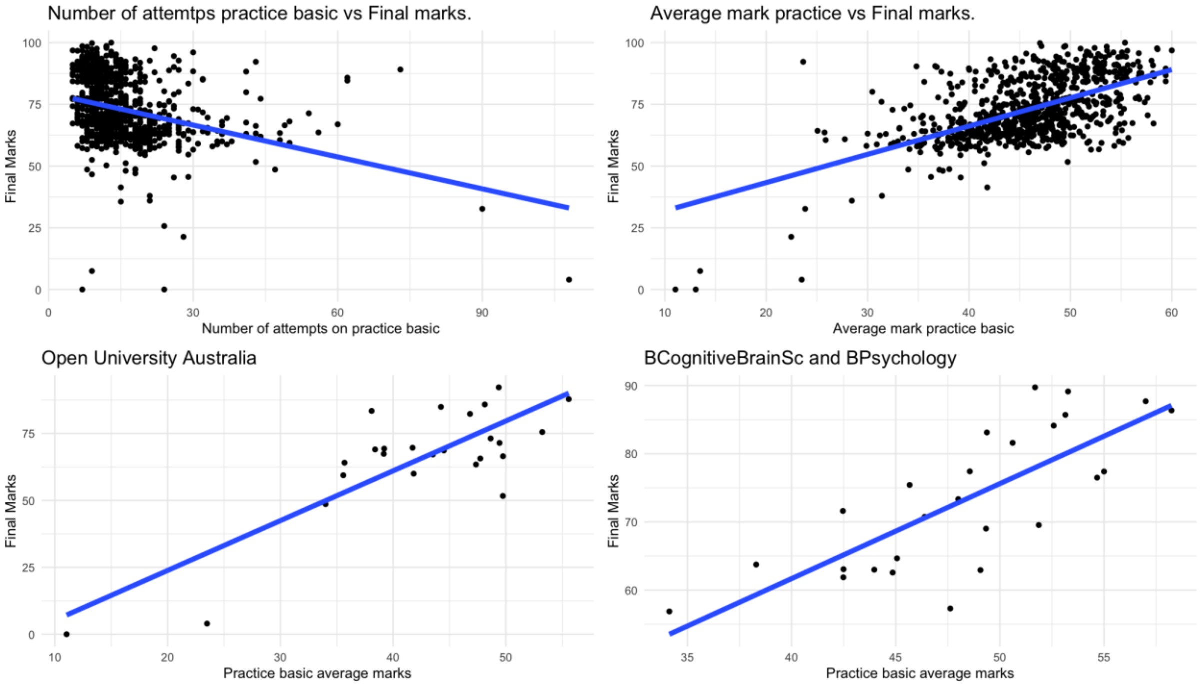

Figure 3: Scatterplots of practice basic test metrics vs final marks.

We focused on the most represented courses to further investigate this group of students. We selected courses with over 3% representation of the total enrolled students, which reduced the dataset to 537 students.

The group of students with the highest R-squared statistics of 0.67 are those enrolled in *the Open University Australia Non-Award Course*. This result is not unexpected, as this group cannot access on-campus classes and must rely solely on the online content and practice tests (see Figure 3, bottom-left). The second highest R-square value of 0.58 is for students enrolled in the double degree program *Bachelor of Cognitive Brain Sciences* and *Bachelor of Psychology* (see Figure 3, bottom-right). *The Bachelor of Science* comes next on the list, with an R-squared value of R-sq 0.46, followed by the

Bachelor of Information Technology with R-sq 0.45, *Bachelor of Medical Sciences* with R-sq 0.31, and *Masters of Data Science* with R-sq 0.29.

We also report that although there was a strong correlation between attendance in SGTA and Practical classes, we did not observe any clear relationship between these variables and either the time spent watching pre-recorded content or the overall marks achieved. Similarly, concerning Lecture participation quizzes, we did not find any significant relationship to overall marks or watching time.

CONCLUSIONS AND DISCUSSION

Using scatterplots and simple linear regression, we analysed the relationship between students' performance on assessments, their time spent watching pre-recorded content, and their effort on practice quizzes. Our analysis revealed two distinct student behaviours. The first type of student diligently watches the lecture content and benefits from it. The strength of the relationship between their watched time and assessment performance varies across different courses. The second type of student behaviour still performs well despite putting only minimal effort into watching the lectures. This suggests a need to tailor additional content to specific cohorts of students. For instance, creating more relevant examples for students in programs like *the Bachelor of Cyber Security* could increase engagement in pre-recorded lectures.

Another identified gap relates to the gamification of assessment tasks, particularly the basic practice test and the tendency of students to keep guessing until passing the quiz. Based on a positive trend, dedicating effort to practice tests by achieving higher marks may have contributed to improved performance on graded assessments among students in the entire cohort. However, we observed a negative correlation between the number of attempts on practice tests and final marks, suggesting that guessing does not lead to better performance. Some suggestions are to reduce the number of attempts allowed on the basic practice test and increase the delay between re-attempts to penalise such behaviour.

In the future, we plan to expand this work in a couple of directions. The first is evaluating derived features of machine-processed language data on the unit discussion forums. These forums are categorised into assessment, content, and general admin topics and are heavily utilised with threads in hundreds. Examining the relationship between student engagement on these forums and their performance would provide valuable insights into their learning processes. We may also include a more extensive analysis of Moodle event logs, especially content video clicks. The second is to analyse future cohorts' datasets using an updated assessment design. As part of a re-design of *the Bachelor of Science*, we are considering changing the assessment by invigilating two out of five module tests and adding a statistical report based on a randomised dataset profiled to each student's course discipline. Each semester brings about subtle variations, and we eagerly anticipate further research and building upon this initial exploration.

ACKNOWLEDGEMENTS

Special thanks to Assoc. Prof. Ayse Bilgin for encouragement, and Dr Connor Smith and Dr Nicholas Tse from Macquarie University for their valuable discussion(s) assisted in developing this work. Any errors or mistakes are solely my responsibility.

REFERENCES

Bilgin, A. A. B., Bulger, D. & Fung, T. (2020). Statistics: your ticket to anywhere. Statistics Education Research Journal. 19, 1, p. 11-20 10 p.

Bloom, B. S. (1968). Learning for mastery. Instruction and curriculum. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1. Evaluation comment, 1(2), n2.

Le, Kien, (2022). Pre-recorded lectures. Live online lectures and student academic achievement, MPRA Paper 112171, University Library of Munich. Germany.

Perez, C. L., Verdin, D. (2022), Mastery learning in undergraduate engineering courses: a systematic review, 129th ASEE Annual Conference and Exposition, Conference Proceedings.

Sabbag, A., & Frame, S. (2021). Learning design and student behavior in a fully online course. Technology Innovations in Statistics Education, 13(1).

Thurn F. (2023). Enhancing 'teacher presence' to improve online engagement, TECHE blog, https://teche.mq.edu.au/2023/01/enhancing-teacher-presence-to-improve-online-engagement/