Evaluating Student Engagement With and Perceptions of a Flipped Classroom Design for a Large Statistics Subject

Paul WT Fijn
University of Melbourne
paul.fijn@unimelb.edu.au

Alba Santin Garcia
University of Melbourne
alba.santingarcia@unimelb.edu.au

We present the preliminary results from a project investigating a large statistics class designed for and taught using a flipped classroom model, with pre-recorded videos. The study, undertaken in 2021 during the Covid-19 pandemic, utilised student surveys in conjunction with metadata on their engagement with electronic resources. This preliminary investigation focuses on the quantitative aspects of the study. We aimed to identify which particular resources or activities were most beneficial to student learning. Overall, all engagement with the course, whether assessed or not assessed, contributed positively to students' learning.

Introduction

Active learning, where students are participants in the learning process, has been shown to be beneficial in a STEM (science, technology, engineering and mathematics) context (Freeman et al., 2014), and also specifically in statistics (Kalaian & Kasim, 2014). This occurs through focusing the responsibility of learning on the learners, and engaging students in deep learning along with thinking about what they are doing and learning. One model to increase active learning when teaching staff are present is a "flipped classroom" model (see e.g., Brame, 2013).

A typical classroom dynamic at our university involves content delivery during lectures, active learning during small-group classes (typically tutorials or computer labs) through working on questions with peers and supported by a tutor, and additional practice done individually. The "flipped classroom" model used in the course for this study reverses this dynamic, through delivering the content as short videos and quizzes completed individually, prior to attending classes, followed by active learning in an interactive lecture, tutorial and a computer lab. This allows students to work through content at their own pace, and students benefit from peer and staff interaction while completing higher level tasks. The interactive lectures are largely inspired by Eric Mazur's (see e.g., Crouch & Mazur, 2001) "peer instruction", using structured multiple-choice questions, combined with peer discussion, in a lecture environment.

Evaluation of the flipped classroom

Several studies have evaluated flipped learning in tertiary mathematics and statistics courses. Commonly these have been quasi-experimental designs comparing the flipped classroom in one semester with a traditionally taught version in another (see e.g., Cilli-Turner, 2015), or parallel classes in the same semester each taught using either a traditional or flipped classroom design (Gundlach et al., 2015; Guererro et al., 2015). The results from these studies are mixed as to the effectiveness of a flipped classroom for educational outcomes, as measured by final grade. They are also limited by confounding factors such as time (different semesters), student agency in selecting which teaching model they enrol in, and differences in teaching staff. Even careful study design cannot eliminate some key issues identified in previous studies, such as differing staff expertise in teaching the two models (see e.g., Simmons et al., 2020).

Unlike previous research, for this study there was no natural comparison to a previous or concurrent iteration of the course taught traditionally, since the course was designed for and has only been taught using the flipped classroom model. Staff have also been able to refine the course and their teaching over time, with these data being collected the fifth time the course

(2025). In S. M. Patahuddin, L. Gaunt, D. Harris & K. Tripet (Eds.), *Unlocking minds in mathematics education. Proceedings of the 47th annual conference of the Mathematics Education Research Group of Australasia* (pp. 141–148). Canberra: MERGA.

was taught. With an established course designed specifically for this teaching model, and staff who have had opportunity to develop their skills in delivering these classes, our focus was on identifying areas for future refinement and development.

Some studies have undertaken a more fine-grained evaluation of flipped learning. Survey-based studies typically have small sample sizes of less than 40 participants (see e.g., Ng et al., 2019; Johnston, 2017). Ng and colleagues (2019) attempted to evaluate the time students spent on various tasks in a course (including video and quizzes) and the students' perception of these tasks. The study was limited by a small sample size (15 students of 19 undertaking the course), only used student estimates of time spent on various tasks, and all student responses fell within a very narrow range (hence comparisons between components were not possible). Our project aimed to examine which aspects of a flipped learning model are effective for learning, using a mixture of metadata and survey responses. The research questions examined were:

- Which activities and/or resources are the most beneficial to students' success?
- Which activities and/or resources do students perceive to be beneficial to their learning in this course?

Determining the most beneficial activities and resources would provide evidence on how to improve the implementation of a flipped learning model for large-cohort teaching that uses video-based asynchronous learning. Success/effectiveness was defined here in terms of overall grades. Students' confidence in statistics and perceptions of learning were also investigated; however, these were not considered appropriate for the analysis presented here due to a lack of data for the statistical methods used for this preliminary analysis.

Research Design

Overview of methodology

Mixed-methods approaches are a powerful tool for education research (see e.g., Johnson & Ongwuegbuzie, 2003) and we sought to combine quantitative and qualitative methods suited to exploring the diversity of student experiences of learning. The full study combines phenomenography (see e.g., Marton, 1981) with a range of quantitative methods for both exploratory analysis and hypothesis testing. This is a preliminary report of the research and focuses on the quantitative results only.

Context

The course being investigated is a second-year introductory applied statistics course with a biology pre-requisite, but no mathematics pre-requisite, except that required for entry into a science degree at our institution. The course was designed to utilise a flipped learning model and has been taught in this manner since it was created in 2017. The course is primarily taken by students studying biology, statistics, environmental engineering and data science, with a small number from other fields. In 2021, 188 students completed the course.

The assessment in the course has four components: weekly quizzes (5%), fortnightly tests (15%), written assignments (25%) and an exam (55%). The weekly quizzes are short (three to five multiple-choice questions), students have unlimited attempts and are given full credit for getting at least one question correct. Fortnightly tests consist of a mixture of multiple-choice questions and scaffolded calculations/data analysis; students have a single attempt (without a time limit) and are graded automatically. All other assessment items are graded by teaching staff, and the grade is contributed proportionately. Quizzes focus on the key concepts for each week and are primarily used for informing the interactive lectures; tests assess lower-level understanding and recall; assignments focus on application of data analysis and interpretation; and the exam covers conceptual and interpretation questions.

The data were collected during 2021 and were necessarily affected by the Covid-19 pandemic. The course was offered in 'online only' and 'hybrid' modes, due to a range of limitations. 'Online only' students attended all classes as a synchronous online activity. 'Hybrid' students attended on-campus tutorials and computer labs, with all interactive lectures held online. Students normally attending their classes on-campus would occasionally attend online if they were unable to attend campus (e.g. if they had symptoms of Covid-19).

Data collection

The data collection combined survey results and a range of metadata available from the learning management system used by our university.

Survey data were collected voluntarily (implied consent) on two occasions: all enrolled students were invited to complete one survey halfway through the semester (April 2021, N=38, 17.9% response rate), and one immediately after the end of classes (June 2021, N=38, 20.2% response rate). The two surveys were identical and aimed to ascertain student perceptions of the various resources (e.g. online videos) and classes (interactive lectures, tutorials, computer labs) within the course. The design of the survey questions was based on the researchers' previous experiences with evaluation surveys; no prior piloting or validation was undertaken. These surveys asked students about their academic background, confidence with statistics and statistical thinking, how useful they found each type of class/resource, and to respond to some specific prompts about key aspects of the classes. There were also open-ended prompts which are part of the qualitative research and not further discussed here. Details of the survey variables, each corresponding to a single survey item, can be found in Table 1. As there were no derived scales used, it is not possible to evaluate the reliability of these variables.

We also collected metadata on student interactions and engagement with all electronic resources and activities, and student results data, under an opt-out ethics approval. Details of these variables observed can be found in Table 2.

Table 1

Survey Variables Measured

Short name	Question/prompt	Response type
Recordings useful	How useful did you find the pre-recorded online lecture videos?	6-point Likert
Interactives useful	How useful did you find the interactive (synchronous) lectures?	6-point Likert
Tutorials useful	How useful did you find the tutorials?	6-point Likert
Computer labs useful	How useful did you find the computer labs?	6-point Likert
Interactives helped	The interactive lectures helped me understand concepts better	5-point Likert
Tutorials helped	The tutorials helped me answer questions and apply knowledge	5-point Likert
Teaching helped	The teaching in this subject helped me to learn	5-point Likert

Statistical Analysis

A range of statistical analyses were planned, all to be implemented using standard statistical software, *R* (version 4.3.2, 2023). They belong to three distinct categories: checking for bias in sampling, applying linear (multiple regression) models to both the survey and metadata collected, and conducting principal component and cluster analysis on the metadata for the whole cohort. Data from various sources (metadata, survey results) were linked using a unique identifier created for the purpose of this study.

Table 2

Metadata Variables Measured

Short name	Variable	Measurement
Quizzes	Quizzes completed	Number of quizzes completed (up to 10)
Assignments	Assignment score total	% awarded, average of 3 assignments
FortnightlyTests	Fortnightly test total	% awarded, average of best 5 test scores
FinalGrade	Final grade for subject	% awarded
PageViews	Accessing materials or resources via LMS	Number of times a page is viewed
Attendance	Attendance at interactive lectures	% attendance
Participations	Participate in LMS activity	Number of assessment attempts
VideoViews	Videos watched (individual videos)	Number of videos watched
VideoViewPercent	Videos watched (total length)	% of total length watched
EngageTotal	Engagement with resources (all documents)	% of resources accessed
EngageWeighted	Weighted engagement with resources (only resources still available at end of semester)	% of resources accessed
Mode	Learning mode	Online-only or hybrid

To check for bias, final grades for various cohorts were compared. Initially these were planned to be independent samples *t*-tests; however, due to skewness in the data, Mann-Whitney *U* tests were conducted instead. One concern was differences in the mode of learning, whether online-only or hybrid, as these students may have substantial differences in the ways in which they interact with the course. Another concern is response bias, necessitating comparison between students completing at least one survey, with non-responders.

Linear models are a standard technique for assessing the contributions of a large number of potential predictors to a single outcome, and have been used for similar studies previously (see e.g., Guerrero et al., 2015). The distribution of final grades in the course exhibited a negative skew; typically, this would be remedied by transformation of the data. Diagnostic plots for the fitted models indicated this was not necessary in this case. Models were fitted to the whole dataset (using metadata only), to the subset of students who completed the initial survey (using metadata and initial survey responses), and to the subset of students who completed the final survey (using metadata and final survey responses). In each case, data were pre-processed to standardise the predictors prior to model fitting. A stepwise process based on Akaike's Information Criterion (AIC), provides an efficient way to compare a variety of potential models, incorporating the quality of the fit (measured here using R^2, the proportion of variability explained by the model) while remaining parsimonious (favouring fewer predictors).

Cluster analysis is a useful technique to identify any groups of students in the data, based on similarities across all the potential variables in the study. To further investigate any groups identified by this method, a principal component analysis (PCA) was performed. PCA allows the data to be viewed in a much smaller number of dimensions, through analysis of the correlations between the variables in the data. As techniques only suited to large samples, cluster analysis and PCA were employed on the metadata only. Student interactions with the course were separately analysed. This consisted of formal interactions (such as individual assessment tasks) and student metadata (including viewing pages on the LMS, watching lecture recordings, etc.) with a total of 17 variables observed for the full cohort of 212 students (which includes students for whom only partial data was available). Excluding students with partial data did not alter the findings of the following analyses, so all are included for completeness.

Results

First, the 'online-only' and 'hybrid' cohorts were compared to determine if there were differences due to the mode of delivery. No significant differences were found (Mann-Whitney U=4284.5, P=0.79) based on teaching mode. The overall sample size was large (N=188); however, the number of students completing the surveys were much smaller (initial survey N=38, final survey N=38, with N=17 students completing both surveys). No significant differences were found between those completing at least one survey and the remainder of the cohort (Mann-Whitney U=3383.5, P=0.27). The small number (N=17) completing both surveys limited the ability to fit models utilising data from both surveys and metadata as predictors.

Linear models

Prior to fitting models, correlations between all variables were assessed. Individually, each predictor was positively correlated with the overall grade (correlations ranging from 0.211 to 0.763), as well as all predictors being positively correlated with each other (correlations from 0.199 to 0.986). With the large number of variables (21 for the metadata only, 38 each for those incorporating survey results) individual reporting of the correlations and their significance is not insightful. A stepwise process using AIC identifies the variables which provide the most information to predict the grade, while omitting those which are redundant. Linear (multiple regression) models were fitted to the metadata alone (all students who did not opt out, N=188) and also to the metadata combined with the survey data (only the students who voluntarily completed the surveys, N=38 in each case). This identified models with reasonable predictive power for all three of these scenarios. The models are summarised below in Table 3.

Table 3

Coefficients for Predictors in Best Models for the Complete Dataset and Two Subsets.

	N	***R^2***	**Predictors in best model identified**	**Coefficient**
Metadata only	188	62.8%	Assignments	1.37****
			FortnightlyTests	1.16****
			Quizzes	-2.68***
			VideoViewPercent	0.063*
Metadata + initial survey	38	75.3%	Assignments	3.04****
			Tutorials helped	-4.88**
			FortnightlyTests	-3.00**
			Quizzes	-4.86**
			Attendance	-14.7*
			VideoViewPercent	-14.3*
			EngageTotal	86.9*
			VideoViews	35.4*
			+ 7 others (P>0.1)	
Metadata + final survey	38	59.7%	Assignments	1.27***
			Teaching helped	5.42*
			Tutorials useful	3.09*
			Recordings useful	-2.86*
			+ 11 others (P>0.1)	

Predictors ordered by significance; level of significance indicated: ****p<0.001, ***p<0.01, **p<0.05, *p<0.1.

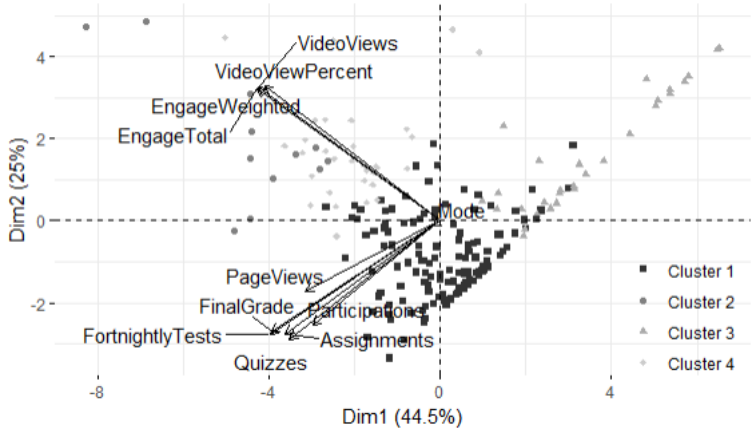
The most salient information in Table 3 is the predictors identified. The exact number of predictors is not important, as there is a tendency for overfitting on small datasets (partially mitigated by utilising a stepwise AIC process). Coefficients are provided for completeness; however, due to the high collinearity between all predictors, some coefficients are negative even though all predictors are positively related to the final score.

Cluster and principal component analysis

Principal component analysis (PCA) is an alternative analysis of the correlations between variables in a manner which complements the linear model analysis above. Cluster analysis found four distinct clusters of individuals: three larger groups (with 139, 30 and 31 individuals) and one smaller cluster (12 students). PCA was used to characterise the clusters, identifying two components being substantially more influential than all others in these data, accounting for a combined 60.1% of the variability in the data. Components based on a slightly simplified model (combining assessment items of the same type) showing the same relationships and accounting for a slightly higher 69.5% of the variability are presented here, to reduce clutter and enable easier interpretation of these graphs. The relationship between the observed variables, the components and the clusters can be seen in Figure 1.

Figure 1

Biplot Showing Principal Components and Clusters



Note. This figure shows the relationship between the clusters and the variables, using the principal components to form a two-dimensional view of the data. Points are individual students, and the arrows indicate the contribution of observed variables to the components.

The PCA in Figure 1 shows two clear groups of variables; within these groupings, variables are highly correlated with each other. The first group, towards the lower-left of Figure 1, includes all assessment items and general engagement with the course (such as number of page views on the LMS, number of assessment attempts/submissions). Students in this direction tended to do these activities more frequently than average, whereas those in the opposite direction (towards the upper-right) are those who did so less frequently than average. The second group, towards the upper-left of Figure 1, are more specific measures of engagement with content on the LMS: watching videos, and engagement with other resources. Considering both the clusters and the PCA, as can be seen in Figure 1, it appears that Cluster 1 (the largest, *N*=139) is students who were generally engaged in the course and assessment, but tended to engage less with resources and videos. Cluster 2 (the smallest cluster, *N*=12), is a group of students who heavily engaged with both assessment and all online resources and videos. Cluster 3 (*N*=30) appears to be students who tended not to engage with any aspects of the course. Finally, Cluster 4 (*N*=31) are students who tended to utilise the online resources and videos more than average, alongside below average general engagement and participation in assessment. Further work is required to fully understand these clusters.

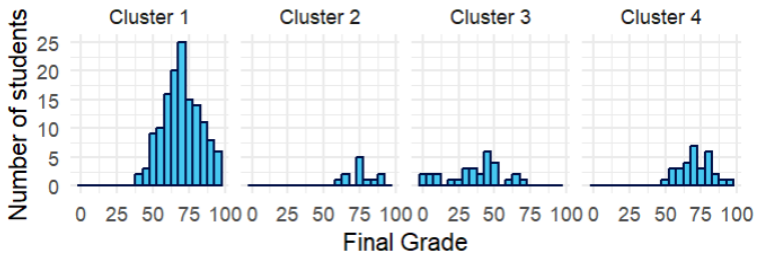
As with previous analyses, students' mode of learning (online or hybrid) was not found to have any significant impact. This can be seen in Figure 1, where Mode does not contribute to either component (in the centre of the graph near (0,0)).

Preliminary investigations of the clusters have identified that one group (Cluster 3) primarily consists of students who did not achieve high grades overall and very few obtained a

passing grade, as seen in Figure 2. These students submitted very few assessment items, viewed the LMS fewer times, and did not watch lecture recordings. This accords with previous research (see e.g., Freeman et al., 2014; Barlow & Brown, 2020), linking disengaged students with lower achievement. The differences between the other clusters show that there are a range of ways for students to successfully engage with the subject: this complicates the aim of this research to determine the most beneficial activities, as some students will benefit more than others.

Figure 2

Comparison of Final Grade for Students within the Clusters



Discussion

Positive correlations between all measures of assessment and engagement indicates that all activity within the course tended to be positively related to the students' overall grade. This does indicate that all activities and assessments were of some value to students. When considering the complete dataset, it is notable that, except for assessment, only the video percentage completed was identified as a significant predictor, although this relationship was weak. Video percentage is perhaps the most direct measure of progress through the course content available in the metadata. Notably this does not measure whether or not students are engaged with (or even watching) the videos, which is a limitation of the available data. The principal component analysis reinforces these findings, where there are two primary dimensions: one relating to engagement with assessment and the course more generally, and a second which relates to watching (or re-watching) lecture videos.

At face value, the clusters identified appear to group students based on behaviours: assessment focused students who view the LMS reasonably frequently (cluster 1); students who are highly engaged in all aspects of the course (cluster 2); students who disengage from most aspects of the course (cluster 3); and students who use a lot of resources but tend to engage less (cluster 4). The results presented here suggest this is a reasonable interpretation, but do not support this as a conclusion. Further work, such as considering engagement patterns throughout the semester rather than overall, will enable this to be investigated more fully. Understanding the differences between these groups will help with identifying which students are most likely to benefit from improving specific resources.

The primary aim of this research was to identify which aspects of the course are most beneficial. Assessed items were generally highly relevant to doing well. This is not surprising, as the final grade is at least partially composed of these results. The weekly quizzes, very low stakes assessments which are not scored proportionately but only on completion, were generally as useful as other assessment predictors, indicating that completion of regular tasks is beneficial to student learning. Notably, static resources and videos were not heavily used by most students (particularly the largest group, cluster 1). Given the substantial time and resources required to produce these relative to other learning aides, extensive work on these is unlikely to be of as much benefit. This is not a strong recommendation, as it is still unknown whether this may systematically disadvantage some students who tend to use these resources more.

The results incorporating the final survey demonstrate that students who had positive experiences of the teaching in the course (by expressing agreement with "The tutorials helped

me answer questions and apply knowledge" and "The teaching in this [course] helped me to learn") performed better on average. This could be due to different reasons, such as students who liked the flipped classroom design being more engaged, or simply that students who responded well to the teaching both learned more and had more positive experiences. Analysis of student comments may be able to elucidate this further – we have chosen to omit any comments here in the absence of a full and rigorous analysis. The current study is only the first step in analysing these extensive data. Future plans for this research include a range of qualitative and quantitative analyses, including: further comparisons based on highly engaged/disengaged student groups; student confidence with statistics as an alternative outcome measure; analysis of student comments as to which aspects of the course they found most useful; and analysis of consistency in completing activities over the course of the semester.

Ethics Statement

Ethics approval (Project ID 13665) was granted by the University of Melbourne, and participants gave informed consent.

References

Barlow, A., & Brown, S. (2020). Correlations between modes of student cognitive engagement and instructional practices in undergraduate STEM courses. *International Journal of STEM Education*, *7*, 1-15. https://doi.org/10.1186/s40594-020-00214-7

Berrett, D. (2012). How 'flipping' the classroom can improve the traditional lecture. *The chronicle of higher education*, *12*(19), 1-3.

Brame, C. (2013). Flipping the classroom. *Vanderbilt University Center for Teaching*. https://doi.org/10.1016/b978-0-12-814702-3.00009-3

Cilli-Turner, E. (2015). Measuring learning outcomes and attitudes in a flipped introductory statistics course. *Primus*, *25*(9-10), 833-846. https://doi.org/10.1080/10511970.2015.1046004

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American journal of physics*, *69*(9), 970-977. https://doi.org/10.1119/1.1374249

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the national academy of sciences*, *111*(23), 8410-8415. https://doi.org/10.1073/pnas.1319030111

Guerrero, S., Beal, M., Lamb, C., Sonderegger, D., & Baumgartel, D. (2015). Flipping undergraduate finite mathematics: Findings and implications. *Primus*, *25*(9-10), 814-832. https://doi.org/10.1080/10511970.2015.1046003

Gundlach, E., Richards, K. A. R., Nelson, D., & Levesque-Bristol, C. (2015). A comparison of student attitudes, statistical reasoning, performance, and perceptions for web-augmented traditional, fully online, and flipped sections of a statistical literacy class. *Journal of Statistics Education*, *23*(1). https://doi.org/10.1080/10691898.2015.11889723

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational researcher*, *33*(7), 14-26. https://doi.org/10.3102/0013189x033007014

Johnston, B. M. (2017). Implementing a flipped classroom approach in a university numerical methods mathematics course. *International Journal of Mathematical Education in Science and Technology*, *48*(4), 485-498. https://doi.org/10.1080/0020739x.2016.1259516

Kalaian, S. A., & Kasim, R. M. (2014). A meta-analytic review of studies of the effectiveness of small-group learning methods on statistics achievement. *Journal of Statistics Education*, *22*(1). https://doi.org/10.1080/10691898.2014.11889691

Marton, F. (1981). Phenomenography—describing conceptions of the world around us. *Instructional science*, *10*(2), 177-200. https://doi.org/10.1007/bf00132516

Ng, W. L., Teo, K. M., Wong, K. Y., & Kwan, K. L. (2019). The delivery role and assessment role of computer-based technology in a flipped university mathematics course. In W.C. Yang, & D. Meade (Eds.), *Electronic Proceedings of the 24th Asian Technology Conference in Mathematics.*

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://doi.org/10.32614/r.manuals

Simmons, M., Colville, D., Bullock, S., Willems, J., Macado, M., McArdle, A., Tare M., Kelly J., Taher M.A., Middleton S., Shuttleworth M. & Reser, D. (2020). Introducing the flip: A mixed method approach to gauge student and staff perceptions on the introduction of flipped pedagogy in pre-clinical medical education. *Australasian Journal of Educational Technology, 36(3),* 163-175. https://doi.org/10.14742/ajet.5600