Learning analytics as a tool for exploring student learning patterns

Bethany Rognoni

BSc(Hons)

School of Agricultural, Computational and Environmental Sciences

University of Southern Queensland

October 2017

Submitted in partial fulfillment of the requirements of the award of Bachelor of Science (Honours)

Abstract

Learning analytics can provide a statistical insight into the learning behaviours of students through the utilisation of datasets retrieved from online learning systems (OLS). These datasets are often large and contain mixed data types, potentially making both collation and analysis of the data complex. This research uses demographic, assessment and OLS data from a large undergraduate service course in statistics, taught at an Australian university. It provides an exemplar of how the application of learning analytics might be performed, using the R statistical package to implement multivariate statistical analyses. The research focuses on both the collection and preparation of educational data for analysis, and the application of both basic and multivariate statistical methodologies (cluster analysis and principal components analysis) to identify relationships between different sources of data. It was found that the data collation process is time- and resource-intensive, but valuable as the integration of different data sources allows a deeper insight into the nature of student interaction within a course. Both cluster analysis and principal components analysis were found to provide useful interpretations of the data. The major relationships identified include: external (online) students achieve higher grades than on-campus students; external students access OLS resources more frequently than on-campus students; students obtain lower grades in the invigilated examination than the open assignments; and students who do not access the OLS resources tend to perform poorer on course assessments. Suggestions for potential interventions with the aim of improving the academic performance of students based on these trends included making early contact with students who are not accessing course resources, and introducing an additional invigilated assessment item to the course assessment structure.

Certification of Thesis

The work presented in this thesis is, to the best of my knowledge and belief, original work of Bethany Rognoni, except as acknowledged in the text. The material in this thesis has not been submitted, either in whole or in part, for a degree at this or any other university.

B. Rognoni — 27/03/2018

Bethany Rognoni — Date

27/3/18

Rachel King — Date

C A. McDonald — 27/3/18

Christine McDonald — Date

Acknowledgements

This thesis would not have been possible without the tireless and expert guidance from my supervisors, Rachel King and Christine McDonald. Your support and guidance over the past two years has been amazing, from the beginning of data entry to the polishing of the final dissertation. I am so grateful for your assistance not only with the thesis itself, but also with my development of organisational and time-management skills. I now know much more about how to use R, RStudio, knitr, LaTeX, and BibTeX – I never thought I could get to this level when you first offered this honours project to me two years ago. I am extremely lucky to have had the pleasure of going through this with both of you, it's been awesome!

I am also very grateful to Enamul Kabir, Taryn Axelsen, and the markers of the statistics course that is the focus of this project, for your assistance with acquiring assessment and OLS data for this research.

Thank you also to my friends, family and colleagues who have been with me through this journey – I really appreciate all the support and enthusiasm you have shown for what I've done!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Learning analytics is defined as the analysis of educational data for the purposes of identifying behavioural trends and understanding how students interact with educational material (Siemens, 2012). The potential for analytics to be used in educational contexts has increased over the past few decades along with the application of advances in technology and computing capabilities within education delivery. However, the field of learning analytics is still relatively young (You, 2016), with papers focused on introducing the field only being written as recently as five years ago (Siemens, 2013; Soby, 2014). Consequently, a generalised framework for the implementation of learning analytics has not yet been fully established. This may, in part, be because the development of such a framework requires a flexible statistical analysis technique as it is heavily dependent on the context in which it is applied (Hernández-García & Conde, 2014).

1.1 Literature Review

Foundational ideas and methods applied in learning analytics to date have frequently been inspired from older fields, such as citation analysis, social net-

work analysis and cognitive modelling (Siemens, 2013). The overarching aim in these older fields is to link behaviour with outcomes by analysing data. This has provided a benchmark to develop applied learning analytics techniques within a defined course (or subject or unit) of educational study (Siemens, 2013).

There are many drivers that motivate current interest in learning analytics. Prior to education being offered extensively online, teacher-student interactions were predominantly in-person, allowing the teacher to monitor student learning behaviour and its impact on academic progression simultaneously. This analysis was often a qualitative assessment based on personal interaction and not necessarily a data driven process. Following the widespread adoption of online delivery of course content, direct contact between teacher and student is becoming less reliable as a method for assessing student engagement and progression. The platforms used to deliver content online also automate the collection of data, aspects of online behaviour of course participants being recorded (Slade & Prinsloo, 2013). This results in the passive collection and storage of large datasets containing course/subject/unit specific information that can potentially inform educators, even in cases when direct personal contact with students is limited or non-existent. Clow (2013) has identified that there is an increasing push toward the collection and quantification of information, as opposed to qualitative evaluation. The quantification of some aspects of education such as learning behaviours, student achievement and teaching methods will enable effective decisions to be made regarding the operation of a course within an institution, unclouded by the potentially biased judgement coming from both learners and educators.

Another motivating factor around learning analytics relates directly to the usability of the large datasets now being collected. Interpretation of these educational datasets using different statistical methods is now possible, due to the advances in big data in other fields (You, 2016). One example includes

using statistical data reduction techniques prior to or as part of the analysis of big data to help reduce the dimensionality or size of these datasets. This has allowed the application of learning analytics to become feasible in cases where huge numbers of attributes are measured for each student (de Freitas *et al.*, 2015). The accessibility of these methods to educators and researchers has improved as the computational capacity needed for implementation of these methods has become readily available within education fields (Slade & Prinsloo, 2013). However, care must be taken by researchers who do not have the statistical knowledge necessary to implement, understand and interpret these methods. One major downfall in this area is that the advancements in computing ability may lead to the use of advanced statistical methods by researchers simply because they are available, resulting in improper implementation of the techniques or misinterpretation of the results (Greller & Drachsler, 2012).

In the case of formal education through secondary or tertiary education facilities, learning analytics was unable to advance further than simple analyses with small sets of data until the advent and widespread implementation of e-learning. E-learning incorporates the use of an online learning system (OLS) for course delivery, which tracks and collects data of students' interaction within the system (Buttner & Black, 2014; Hu *et al.*, 2014; Siemens, 2013). Each time a student performs an action such as accessing a resource, watching a lecture, submitting an assessment item, or posting on discussion forums, a log is typically automatically generated within the OLS, recording details of the action. These logs of interactions can form massive datasets which are much more comprehensive than what was available prior to e-learning and OLS implementation (Gibson & de Freitas, 2016). In the past, surveys of students regarding their interaction, records of final grades, and sometimes records of class attendance were generally the only indications of student interaction that educators and researchers had access to, which limited the advancement of learning analytics. Now that access logs are generated for each student's

individual interactions with the course content, the potential applications of learning analytics have expanded significantly (Romero-Zaldivar *et al.*, 2012; You, 2016), with no extra effort required by the course educators for data collection to occur.

The challenge now is how best to realise the full potential of these new and potentially large data resources. The approach commonly taken is to identify trends in the data after the delivery of educational content has concluded (de Freitas *et al.*, 2015; Strang, 2016). This is useful to educators and researchers, as it provides a method for identifying areas where students may tend to struggle, or may reveal modules (subsets of content) that have lower rates of interaction (Soby, 2014; Yassine *et al.*, 2016). This analysis can then inform changes to course content or delivery for future cohorts of students. However, since OLS data is generated in real-time (that is, it is generated and accessible as soon as student interaction has occurred) there is also the potential for real-time analysis to inform educators about their current cohort of students and provide them with tailored feedback and interventions (Gasevic *et al.*, 2016). Real-time analysis does pose the additional challenge of needing to complete analyses and implement changes quickly so that the current cohort is benefited, which may require extra resources (time, people) to be allocated to this work (Romero-Zaldivar *et al.*, 2012). Whichever approach is taken, the educator or researcher must be aware of the challenges innate to using the data itself.

The value of the automatically generated OLS data is not always clear in its raw form. Currently, much of this data is left unused, due to the absence of generally applicable guidelines or a framework for effective and efficient techniques of OLS data management and appropriate methods of statistical analysis. Some OLS platforms do provide tools for data management and data summarisation, however the functionality of these tools is rudimentary at best, or too specific (Ferguson, 2012; Hernández-García & Conde, 2014; Yassine

et al., 2016). Developments in the data science field and the propagation of big data across many discipline areas in the last decade has stimulated the development of a range of statistical methods for dealing with large datasets (Giannakos *et al.*, 2016). These forms of analysis may assist in overcoming initial barriers of data usability and accessibility in learning analytics (Clow, 2013). In cases where data management tools offered through the OLS are not sufficient, the educator/researcher must build a framework for the application of statistical methods specific to the data of interest.

Initially the most important records of interaction need to be identified for any application of learning analytics to allow the development of strategies for data selection and collection (Hernández-García & Conde, 2014; Serrano-Laguna *et al.*, 2014). This allows for more insightful analyses to be performed when starting with variables that are specifically targeted to the hypothesis (Bainbridge *et al.*, 2015). However, when data is automatically generated by an OLS, the particular form that the database takes is decided during the creation of the OLS, and the data may not necessarily be tailored for easy use in statistical analyses. Unfortunately, as highlighted by Dringus (2012), the limitations on the data collected by a particular OLS in turn limits the educator's or researcher's ability to obtain a complete and accurate picture of the way students interact with the learning environment.

An important aspect of interpreting OLS data raised by Bainbridge *et al.* (2015) was that the selection of particular interactions for analyses should be based on a desire to understand and recognise genuine learning, rather than just collecting indications of simple participation. Nyland *et al.* (2016) identified that "transaction-level data" offers the most detail concerning students' interactions with content. This gives it "high validity" according to Bainbridge *et al.* (2015) as it is as close as possible to representing the true engagement of a student with content. An example was given by Nyland *et al.* (2016), in which the focus was restricted to a particular assessment question that was very con-

fined in its nature. In its assessment, students were expected to complete a set number of tasks within a set workspace (Microsoft Excel), and the assessment was arranged such that transaction-level data (for example, records of time spent between attempts at each step of the problem) were feasible to collect. The study concluded that using more detailed achievement data provided a better understanding of "knowledge gaps" than overall achievement, however it is cautioned that this may not be feasible for courses with different assessment types. Another example by Romero-Zaldivar *et al.* (2012) promoted the use of a self-contained, purpose-built educational setting to observe learner behaviour, allowing the educator or researcher to tailor the aspects of the learner behaviour recorded. Despite transaction-level data enabling a better understanding of genuine learning, this data also tends to be the most difficult data to analyse, due to the large number of attributes being measured and the need to manipulate the data in preparation for analysis.

A study by Cocea & Weibelzahl (2011) addressed directly the task of inferring true engagement of students with educational material based on access logs provided by standard OLS systems. Their aim was to provide a way for disengagement in students to be detected and flagged by the system in a timely manner. They explored approaches to define exactly what "disengagement" means in terms of the data available. Time spent by a student on each section of a module and responses to each question of an assessment item are examples of the type of data that was collected. Human "raters" were asked to rate each student's engagement with each item based on the time spent on each part – if they had spent too little or too much time on a particular section, it was deemed as disengagement. Disengagement in this context was defined as the opposite of genuine learning. This method of data collation, or value-adding to the automatically generated data, is not only time-consuming, but also requires the allocation of a lot of resources including the time taken for raters to sift through data for each student and manually input engagement ratings. Furthermore the study aimed to deliver results in real-time, which

as highlighted earlier also requires additional investment of resources. This approach may be feasible for contexts where the number of students is small (in the Cocea & Weibelzahl (2011) study there were only 48 students), however it would not be appropriate for situations where larger cohorts of students are involved. The application of learning analytics to detect trends of genuine learning often involves finding a workable balance between resource investment and useful output, and this balance becomes increasingly important as the size of the dataset increases (Gasevic *et al.*, 2016).

Researchers must also consider which OLS records are more useful than others for the identification of associations between learner behaviour and academic success. Hu *et al.* (2014) aimed to implement an early warning system that could reliably detect early learning behaviours in students that were strongly related to failure later in the teaching period. The study concluded that having access to time-dependant variables (that is, variables that change over time, such as frequency of OLS access, or time spent using the OLS per week) provided a more reliable prediction of academic achievement, as opposed to static variables. It is cautioned, however, that this approach is only suitable if the OLS is the primary method of accessing course resources; if an OLS is only used as a supplement to course delivery, then time-dependent variables lose their meaning. It is vital that researchers consider the course structure, particularly how students are expected to be using particular resources, and determine from this which aspects of the OLS data are likely to be useful for analyses (Gasevic *et al.*, 2016).

Within a university context, Gibson & de Freitas (2016) highlighted that the OLS data available is not restricted to within-course data. The OLS data may include information about students' account activity and their activity on the university's online library, in addition to any voluntary information given in contexts such as course evaluation surveys and interviews. Information such as a student's engagement with their student account or the online library are

not likely to be indicators of engagement with learning materials specifically, so discretion must be used by researchers when determining the usefulness and meaning of this data. There is also the possibility of students accessing resources through other means. For example, if a student is on-campus, they may never log into the online library, opting instead to visit the on-campus library to access information. In cases specific to a particular course, a student may only be interested in attending lectures on-campus, and may not access many of the OLS resources due to this. For a student studying primarily online, an OLS record stating that they downloaded a document, for example, does not capture *how* the student interacted with this resource. It could be that they downloaded it and never looked at it again, instead of truly engaging with it. In this way, OLS data can be misleading, as interpretation of the true meaning of access records can be complex (Bainbridge *et al.*, 2015). The challenge then lies in the interpretation of a student's access records with regards to inferring true engagement and learning.

The useful application of learning analytics to determine true engagement with material typically requires additional pertinent data such as assessment and demographic data for each student. For this reason, learning analytics is often restricted by either not having access to, or needing to invest time and effort into acquiring this extra data (Ellis, 2013). Demographic data most often comes from the university's databases of student information (Gasevic *et al.*, 2016; Martin & Whitmer, 2016; Strang, 2016), but can also be gathered from external sources, such as the Australian Census (de Freitas *et al.*, 2015). Overall assessment data can often be gathered from the university's grades database. However, if more detailed assessment data is required, such as part-marks attributed within each section of an assignment, manual collection is often required as this level of detail is not generally stored when assessment items are marked. Bainbridge *et al.* (2015) also highlighted that grades by themselves are not sufficient to understand the true learning journey of a student, and the combining of assessment data and OLS data can add value to

each dataset, to create a more comprehensive picture of student engagement and learning.

Several researchers have also highlighted the privacy concerns surrounding the automatic collection and collation of student activity data (Clow, 2013; Dringus, 2012; Pardo & Siemens, 2014; Slade & Prinsloo, 2013; West *et al.*, 2016). Students are often not made explicitly aware that their interactions with university systems are being recorded. Additionally, if course instructors choose to analyse this data, students are often left unaware as to how their data is being used, and for what purposes (Pardo & Siemens, 2014). Slade & Prinsloo (2013) emphasised that clear ethical boundaries should be placed on how OLS data is interpreted, managed, and kept private. Dringus (2012) also expressed concern that transparency may be overlooked during the learning analytics process due to the massive amounts of data readily available through the OLS via passive collection. A study on the ethical concerns of learning analytics was conducted by West *et al.* (2016), where students and staff from several universities in Australia were interviewed on their opinions surrounding how ethical considerations are identified and managed within learning analytics application and research. They concluded that the notion of ethical considerations in regards to learning analytics is still a relatively new concept in most institutions. This is likely due to learning analytics itself being a very recently developed field (West *et al.*, 2016). General unawareness of passive data collection by an OLS, coupled with the absence of strict ethical guidelines or frameworks specific to learning analytics, means that it is vital that educators and researchers implement open and explicit ethical practices and ensure that these meet existing ethical standards regarding data management and privacy (Romero-Zaldivar *et al.*, 2012).

Many researchers have acknowledged that given the range of data types available for analysis from an OLS, not only the data type but also any assumptions about the data distribution must be considered when choosing an appropriate

statistical method (Gasevic *et al.*, 2016; Strang, 2016). The importance of cleaning the data as a pre-analysis process has been addressed by several researchers, as this is where any mistakes, errors or inconsistencies in the dataset need to be removed or corrected (Gibson & de Freitas, 2016; Nyland *et al.*, 2016). Once the data has been cleaned, statistical methods are also often used for further data reduction.

Data reduction has two main components: feature reduction and dimension reduction (Han *et al.*, 2012). Feature reduction refers to the removal of extraneous variables not of interest to specific analyses. It is often a necessary step when dealing with automatically generated data where data for a general set of variables are collected and this set cannot be restricted or limited prior to collection. Bainbridge *et al.* (2015) also suggest only using a particular access record for a course resource where at least 80% of the student cohort have interacted with it. If this threshold is not met, it can be left out of the analysis. This will reduce the data to only those variables which provide meaningful information about the cohort. Dimension reduction also applies to variables (not cases) but requires the application of statistical methods to reduce the number of variables that form the basis of results for interpretation. Often dimension reduction techniques are applied to reduce the impact of problems such as multicollinearity in the dataset, and to ensure that analyses are meaningful. Methods such as cluster analysis and principal components analysis are examples of dimension reduction methods (Papamitsiou & Economides, 2014).

Exploratory and descriptive data analysis is used to understand the characteristics of the cases themselves (Bainbridge *et al.*, 2015; Cocea & Weibelzahl, 2011). In learning analytics, this process reveals characteristic trends in the cohort of students being studied. For example, it may be interesting to know the proportions of the cohort with respect to gender, age, or academic background (You, 2016). Exploratory analysis also draws the researcher's attention

to broad trends in the dataset that may be important to be aware of before proceeding with the main analyses (Giannakos *et al.*, 2016). This may entail testing several different statistical techniques on some variables in the dataset to see which best suit the situation, or performing some basic statistical tests (such as chi-square tests or t-tests) to quickly identify if there are any potentially overwhelming trends (Buttner & Black, 2014; Cocea & Weibelzahl, 2011).

One of the most popular statistical methods used to directly address research hypotheses related to student engagement and progression in learning analytics is regression analysis. This method provides information on the nature of a relationship between one dependent variable and one or more independent variables (Bainbridge *et al.*, 2015; Gasevic *et al.*, 2016; Gibson & de Freitas, 2016; Hu *et al.*, 2014). Multivariate statistical techniques, such as discriminant function analysis (useful in finding predictors of categorical dependent variables), cluster analysis, principal components analysis and multivariate regression analysis, can also be used to explore relationships where there is more than one dependent variable (Manly, 2005). However, few studies have been found which have explored the utility of these multivariate methods specifically in the context of learning analytics. Sutton & Nora (2008) used factor analysis for dimension reduction, however then uses multiple regression (as opposed to multivariate regression) for the final analysis. Strang (2017) tried using cluster analysis, however found that a definitive model was not able to be produced in the particular circumstance, and so the results were not presented.

Ellis (2013) cautions that statistical analyses need to be performed with the right goals in mind. Only focusing on identifying trends that make students achieve very highly or put students at risk of failing, for example, disregards the analysis of the bulk of average-performing students. It is important that the analyses are focused appropriately to adequately answer the research questions. As mentioned earlier, it is also important to ensure that the data used for

the analyses is meaningful, which can entail introducing thresholds such as minimum participation requirements (Bainbridge *et al.*, 2015). Ultimately, it is important to have the overarching aims of the project in mind when making decisions relating to the data and which analytical techniques to use.

1.2 Research Aims

This project will focus on collecting and analysing data from a first year undergraduate service course in statistics taught at an Australian university, with three offers (semesters) delivered over the full 2016 academic year. Some of the reasons for choosing this course are that many students show high levels of anxiety and low levels of quantitative skills when entering into the course, which can lead them to fall behind early in the course or struggle to maintain momentum throughout the course. Drawing comparisons between achievement data and OLS data may help with course delivery and enhance learning outcomes for these students. This course also allows for the comparison of on-campus and external (online) student behaviour. The methods used in this project, if shown to be informative, may then be extended to apply to other courses.

Using assignment and exam achievement data in conjunction with selected OLS activity reports to assess three chosen course topics, the following three aims will be addressed:

- The major characteristics of assessment achievement of each cohort will be identified and summarised, and the relationships between assessment, demographic and OLS variables will be explored;
- Any underlying groups in each cohort will be identified using multivariate statistical methods;
- The information gathered in addressing the first two aims will be used

to inform potential interventions that could be administered to improve student performance.

These aims will be considered both within each of the three teaching semesters in 2016 individually, and among the three offers. They will also be considered when comparing on-campus and external (online) students. Ethics approval has been gained by the project supervisors for this research; the ethics code is H15REA223.

Just as important as these specific research goals is the expansion of the field of learning analytics itself. Importantly, this project will not just analyse OLS, demographic or assessment data separately (which is often the case in learning analytics studies), but will merge them to better understand how they inform each other. If the process of analysis of many types of educational data can be streamlined or at least made easier through the efforts of this project, then these processes themselves are valuable to the university.

Chapter 2

Methods

2.1 Structure of the course studied

The course being investigated in this research is a large service course in statistics that is core to the majority of programs at the university. There is significant student anxiety regarding the course, and as such many levels of support and resources are available for student access. The course content requires progressive learning, as each new topic builds on previously delivered content. Due to this, maintaining regular progress and engagement with the course material can significantly improve student performance.

The assessment for the course consists of four main items: three assignments, the first of which assesses early engagement only and minimal understanding of course content; and a final examination, which is presented in two parts – a multiple choice section (Part A) and a written short answer section (Part B). Even though new assignments and exam questions are developed for each semester, care is taken to ensure that each assessment item is equivalent in content of topics and difficulty across the three semesters. The first assignment and the first part of the final exam (Part A) were not considered as relevant indicators of student progression through the course, as the former did not

assess course content, and the latter was a multiple-choice question assessment which provided no indication of partial understanding of topics. As such, the assessment data collected for this research consisted of specific questions chosen from the second and third assignments, and the short answer section of the final exam (Part B). These assessment items required students to actively engage with the material, and show understanding through practical analysis and/or written answers.

Assignments 2 and 3 cover different content and are to be completed by students at different stages of the semester. The assignments are non-restricted, in that students are able to access any resources they choose to assist them with assignment completion. The same content is generally covered in the final exam. The final exam is restricted and supervised, however students are allowed access to their own A4 page of notes, as well as the formula sheet and statistical tables provided with the exam. The exam also incorporates a second requirement for achieving a passing grade in the course: students must achieve at least 40% on the exam, as well as 50% in the course overall, to be awarded a passing grade. Although final grades are not considered in this research, it is important to note that this second hurdle may affect student anxiety and achievement in the exam assessment. Using the assignments and the final exam, it is possible to explore changes in student achievement in particular topic areas over the course of the semester.

Prior to collection of the assessment and OLS data, three course topics that were assessed in one of the written assignments and in the final exam were chosen as the topics of interest for this research. Regression, binomial distribution and hypothesis test topics were chosen as they represent key course content and concepts, and are delivered at different stages throughout the teaching period. It is hoped that patterns of student engagement and achievement may be found by mapping achievement across the assessment items for each of the chosen topics, and pairing this with OLS access data for each topic. See Table

2.1 for an outline of the topics chosen and their appearance in the assessment items.

Table 2.1: Question mapping across semesters

Assessment	Regression	Binomial	Hypothesis Test
Sem 1 A2	Question 3	Question 6	
Sem 1 A3			Question 2
Sem 1 Exam	Question 2	Question 3	Question 6
Sem 2 A2	Question 6	Question 5	
Sem 2 A3			Question 2
Sem 2 Exam	Question 5	Question 6	Question 3
Sem 3 A2	Question 6	Question 2	
Sem 3 A3			Question 3
Sem 3 Exam	Question 5	Question 2	Question 1b (4 parts)

2.2 Data Collection

Student demographic data, including information such as study mode, citizenship status, and degree and major of study, was downloaded from the OLS. The final demographic database only included those students who had remained enrolled for the entirety of the semester, which was acceptable as the scope of this project includes only students whose progress throughout the teaching period is trackable up to and including the final exam.

The collection of the assessment data began once Semester 1 2016 grades had been finalised, with data being entered into a spreadsheet in Microsoft Excel by hand. Only final grades for each assessment item are entered into the OLS, not the total marks awarded for each question or part of each question. This data, therefore, needed to be manually recorded as part of this project. This manual time-intensive process is identified as a major barrier to the accessibility and

usefulness of learning analytics for educators and researchers, where the goal is to integrate assessment and OLS data. Minor errors in the assessment marks were corrected during collation (e.g. small errors in summing of marks across an assessment item by markers).

As the OLS generates an access log for each object or activity that can be accessed by students, the access logs of interest were those that would provide insight into student engagement with material relevant to the three topics of interest (Table 2.1); OLS objects which required interaction from both on-campus and external (online) students best suited this purpose. For example, student interactions with online lectures were not considered useful as on-campus students could access this content through class attendance rather than through the OLS. Tutorial solutions, in contrast, were accessible only within the OLS for all students, and students who accessed the solutions were likely to have been genuinely interacting with the tutorial questions and, subsequently, the topic content. For this reason, the tutorial solutions relevant to the three topics of interest were chosen as OLS objects of interest for this research. It was also decided that the example exams access logs would be used, as they indicate engagement with revision material close to the final examination. The OLS logs for the three chosen topics' tutorial solutions and the example exams were downloaded for each semester separately.

2.3 Data Cleaning

All data files were imported into the statistical software R version 3.3.2 (R Core Team, 2017) and all data cleaning and analyses were performed in R via RStudio version 0.99.902 (RStudio Team, 2015). The package "knitr" was used to incorporate this work into the final LaTeX document (Xie, 2017). The R code for data cleaning and merging was extensive and is therefore not included in the Appendix, however full code is available upon request to the author.

Cleaning was not required for the demographic data prior to merging and analysis with the other datasets.

Prior to the commencement of analysis, all students with any missing assessment data were removed from the assessment database. This was necessary because no mapping of topic understanding across the semester could be performed for students who did not complete all assessment items. In addition, the methods of multivariate analysis used in this project require that cases with missing values for any variable need to be removed.

The downloading of OLS logs for Semesters 2 and 3 2016 was straightforward, however for the Semester 1 2016 OLS data, the downloaded files only included access records dated on or after 6 May 2016. The OLS had stored the access logs for the majority of OLS objects in two separate files: one with logs generated prior to 6 May 2016; and another with logs generated on or after 6 May. This problem was specific to the Semester 1 2016 OLS. These two databases were merged in R prior to merging with other data types (i.e. demographic and assessment data).

Merging the three databases (OLS, assessment and demographic) required a common variable on which to merge. The OLS database did not include student identification numbers, so matching the databases was done by student name. It was assumed at the outset that no two students had the same name, and no issues in merging occurred to suggest that this assumption was false or misleading. Matching databases by name was challenging due to the format of student names in the OLS files being different from the format in the demographic and assessment data files. As part of the merging process, any students who did not appear in the assessment database were removed from the merged database, to prevent missing data in the final analysis database. One overall database was created for each semester, resulting in three databases on which statistical analyses could be performed.

2.4 Preparation of the Data for Analyses

The original demographic data listed each student's degree and major, with over 40 unique degrees listed in each semester. Each of the degree programs studied by students were sorted into one of four categories (Degree Type):

- Business, Commerce, Law and IT (BusComLawIT)
- Psychology
- Science
- Other

These categories align with the general perspectives held by course examiners about the composition of the student cohort, and are coherent from the perspective of the university's program structure. The programs that students were enrolled in differed for each semester, and so the R code used to assign one of the four categories to each student was changed for each semester. Programs listed under the "Other" category included Engineering, Education, Arts and Professional Development; none of these programs have a compulsory requirement for students to complete the statistics course that is the focus of this project, unlike the programs in the other three Degree Type categories.

The assessment data contained scores achieved by students for each question. For example, if a student achieved 10 out of 18 on a question, this mark would be recorded as 10 in the database. These scores were converted into percentage achievement by dividing the score by the question total, to enable comparisons in achievement between questions for the same topic with different totals in different semesters.

Cutoff dates for the access logs were implemented for the overall semester, so that the OLS data contained only interactions that occurred during the teaching period. Any logs of accesses made prior to the first day of the teaching period and after the final exam date were removed from the database.

When a student accesses an OLS object, several access logs are typically gen-

erated within a few minutes of each other. It is unknown if there is a precise cause for this, however given the short time period within which they occur it is not reasonable to interpret these records as unique interactions by the student. To mitigate this complication, any access logs generated on a given day (regardless of how many) were condensed into one single access record of that OLS object by that student. These unique days were then tallied to give a count variable representing the number of times the object was accessed by each student during the semester. This count was then further summarised to a binary categorical variable, indicating whether or not students had accessed the content at any stage during the teaching period.

2.5 Summary Statistical Methods

Summary statistics were performed for each of the three semesters to provide an insight into the makeup of the student cohort for each semester. R code for each of the analyses run can be found in Appendix A.

The distributions of the relevant variables for students in each cohort were explored using contingency tables and bar charts. The demographic variables "Study Mode" and "Degree Type" and the frequency of access of OLS objects were of primary interest.

Boxplots were used to compare the assessment achievements of both on-campus and external (online) students across both assessment items (the assignment and the exam) in the three chosen topics. Another set of boxplots was created to investigate the differences in assessment achievement for the two levels of OLS access (never accessed, and accessed at least once).

To further understand student performance across the teaching period, histograms of the differences in achievement between the assignment and exam questions for each topic were plotted ($Difference = Exam - Assignment$), and Wilcoxon Signed-Ranks tests were performed on these differences to deter-

mine whether or not they were significantly different from zero. The Wilcoxon Signed-Ranks test was chosen as it is a non-parametric test for comparing matched samples, and therefore does not require the underlying populations of assignment and exam scores to be normally distributed.

2.6 Cluster Analysis

Cluster analysis was performed in order to identify the existence of any naturally occurring distinct subgroups in the data. Each semester's database was analysed separately. Cluster analysis methods are based on distance measures where a single distance value is calculated between any two cases (students) based on the variation (or difference) between them across multiple variables of interest. Clusters of cases (students) are then formed on the basis of defining cluster membership which minimises distances between data points within a cluster, while maximising the overall distances between clusters (Manly, 2005).

There are many forms of distance measures available. For this research the Gower distance was used to calculate the distance matrix (a matrix of the multivariate distances between each pair of cases), as the Gower distance allows mixed data types to be used in the analysis. This distance method differentiates between categorical and continuous variables by applying a relevant distance method to each, and then combining the distances into one final value representing the aggregate distance between two cases (Kalisch, 2012).

To find the distance between two cases (two students) on a continuous variable, the following equation is used:

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{R_f},$$

where $d_{ij}^{(f)}$ is the distance between case i and case j on variable f. The variables

x_{if} and x_{jf} are the scores for case i and case j on variable f, and R_f is the range of possible scores for variable f. The difference between the scores for both cases is divided by the range of possible scores so that the distance value calculated is between 0 and 1 (Kalisch, 2012).

The equation for the calculation of the distance between two cases on a categorical variable is given below:

$$d_{ij}^{(f)} = \begin{cases} 1, \text{ if students are different on variable } f \\ 0, \text{ if students are the same on variable } f \end{cases}$$

Once the distances between the two cases on each variable are calculated, these distance values are aggregated into the final Gower distance using the following equation:

$$d(i,j) = \frac{1}{p} \sum_{f=1}^{p} d_{ij}^{(f)},$$

where $d(i,j)$ is the aggregate distance between cases i and j, p is the total number of variables included in the analysis, and f is the index of summation. The sum of all distance values is divided by the total number of variables, again to ensure that the final distance metric lies between 0 and 1. The distance matrix is then constructed from these final distance values, $d(i,j)$. For the creation of the final distance matrix, a linear combination of the distances using weightings for each variable can be calculated, however for this analysis it was decided that the weightings would not be changed from the default of equal weightings for all variables. The final distance matrix is of $(n \times (n-1))$ size, where n is the original number of cases (students) in the semester dataset (Kalisch, 2012).

The clustering algorithm selected was the Partition Around the Medoids (PAM) agglomeration method (Spector, 2011). This algorithm is similar to k-means clustering, except that members of the dataset are chosen as the centre measures for each cluster, rather than means generated based on the data points

in each cluster that may not exist as an actual case in the dataset. It is an iterative procedure where cluster medoids are randomly selected, and then the distance matrix calculated earlier is used to assign each data point to its closest medoid. The algorithm then calculates if any data points in each cluster would provide a lower average distance if they were to be designated as the medoid instead of the original. If any data points are identified as potentially better medoids, then the process repeats for these new medoids, until the best medoids are found (Spector, 2011).

Partitioning around the medoids can be done only when the number of clusters to be generated is first defined. A metric known as silhouette width can be used to determine the level of appropriateness of the clusters that the data have been assigned to (Rousseeuw, 1987). For this cluster analysis, the PAM algorithm was run for many different numbers of clusters, and then the silhouette width values for each scenario were plotted to visually compare and determine the optimal number of clusters to be used. Silhouette width was used as it can be calculated with any distance metric, which is important since Gower distance calculation involves different distance metrics (Rousseeuw, 1987).

As described above, the cluster analysis performed in this research included both quantitative and categorical variables. The quantitative variables used were the achievement variables for the assignment and exam questions for the three chosen topics (six variables in total), while the categorical variables used included the demographic variables "Degree Type" (coded with levels 'BusComLawIT' (combined Business, Commerce, Law and IT factor), 'Psychology', 'Science' and 'Other') and "Study Mode" (coded with levels 'External' and 'On-campus' for Semester 1 and 3, and coded with levels 'External', 'On-campus Toowoomba' and 'On-campus Springfield' for Semester 2) and the OLS access binary variables for the tutorial solutions for the three chosen topics (coded with levels 'No Access' and 'At Least One Access').

The clusters can be visualised through the use of a two-dimensional ordination

plot, which represents the relative distances between cases based on the distance matrix. The scale of the axes of the ordination plot have no relationship with any of the original variables. The ordination plot provides a graphical representation of the approximate (relative) unitless distances between cases, to enable interpretation of how the clusters differ and the variance among cases within clusters. Care must be taken with interpretation, as the definition of clusters is a subjective process and may be influenced by the biases of the researcher.

The packages "dplyr" (Wickham *et al.*, 2017), "cluster" (Maechler *et al.*, 2017) and "Rtsne" (Krijthe, 2015) were used to perform the cluster analysis in this project.

2.7 Principal Components Analysis

Principal components analysis (PCA) is a multivariate data reduction technique that is based on the correlation between variables, rather than the distance matrix used for cluster analysis. It identifies patterns of simultaneous variation by using either the correlation or the covariance matrix of the variables, and constructs a new set of variables by using linear combinations of the original variables (Manly, 2005). These new variables, or principal components, are chosen so that the first accounts for as much variation in the data as possible, the second accounts for as much variation as possible that is not included in the first principal component, and so on until all variation in the data is accounted for.

The principal components are determined by finding the eigenvalues and corresponding eigenvectors of the correlation (or covariance) matrix of the original variables. Each principal component has a corresponding eigenvalue and eigenvector. The eigenvalue represents the variance explained by the principal component, and the elements of the eigenvector represent the loadings (the co-

efficients) of the original variables on the principal component (Manly, 2005). The composition of the principal component is given in the equation below (where Z_i represents the i^{th} principal component, X_j represents the j^{th} original variable and a_{ij} represents the loading of the j^{th} original variable on the i^{th} principal component):

$$Z_i = \sum_{j=1}^{p} a_{ij} X_j$$

Because PCA is a data reduction technique, it is hoped in performing PCA that a relatively small number of principal components account for a large proportion of the total variation in the data. Interpretation of the principal components can be done via ordination plots, specifically bi-plots, which show a two-dimensional plot of the data with the first two principal components on the x- and y-axes, and vectors of the original variables superimposed to allow interpretation of their correlation with the principal components.

Basic PCA, in relying on the correlation (or covariance) matrix to calculate principal components, requires all variables to be continuous, as categorical variables do not correlate with other variables in a traditional sense. There are advanced PCA methods that allow the inclusion of categorical variables in the analysis, such as CATPCA (Linting & van der Kooij, 2012), however in this research the PCA was performed using only the continuous variables, and then the categorical variables were added to the model afterwards as supplementary variables. Their relationship with the principal components can be visualised, and interpreted descriptively, by adding their centroids to a bi-plot corresponding with the factor map of the continuous variables. The centroid of each category is based on the distribution of the cases in that category (Lê *et al.*, 2008).

For this research, student achievement variables were used as the continuous variables for the generation of principal components, and the supplementary variables used included "Degree Type", "Study Mode" and three binary OLS variables indicating whether or not students accessed the tutorial solutions

for each of the three chosen topics (regression, binomial and hypothesis test). The supplementary variables were added to the ordination plot to enable interpretation of their relationships with both the principal components and the variable vectors. PCA was used as it was hypothesised that the student achievement variables would be moderately to highly correlated, and PCA is an effective technique for the visualisation of highly correlated data in two or three dimensions. The R function "prcomp" was used to generate the principal components, and the R package "FactoMineR" was used to add the supplementary variables and generate the ordination plots (Lê *et al.*, 2008).

Chapter 3

Results

Results from the analysis of Semester 3 2016 data are presented here in full as an exemplar. In addition, the results for Semester 1 and Semester 2 2016 are described in relation to corresponding Semester 3 2016 results, while the full set of results for these two semesters are presented in Appendices B and C. Semester 3 consisted of students enrolled externally, as well as on-campus at Springfield (there was no on-campus Toowoomba mode offered in this semester).

3.1 Assessment achievement

Difference in Assignment and Exam Achievement

It is important to understand the general trends of achievement in the course to gain a better understanding of the cohort and the course. Figure 3.1 shows that achievement in the assignments is skewed to the left, with the majority of students achieving between 65% and 100%. On the final exam, however, the distribution is more even across achievement levels, indicating that there are more students performing poorly in the exam than in the assignments.

To further explore the difference in performance between the assignments and

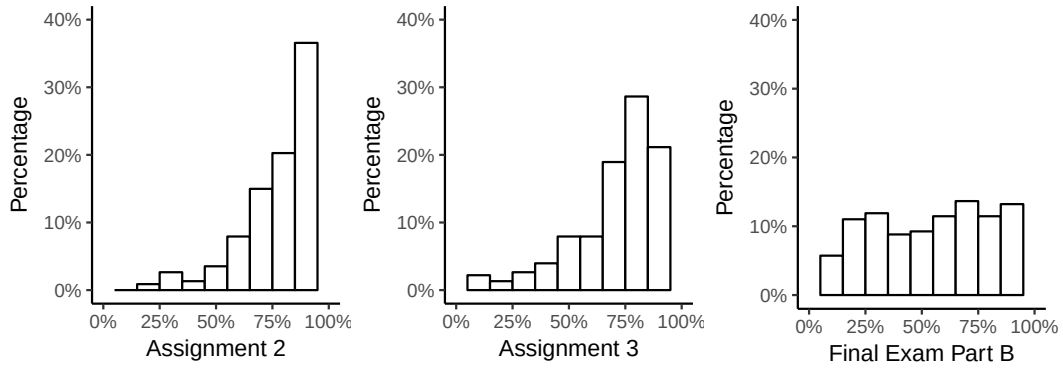

Figure 3.1: Distribution of overall achievement for each assessment item

the exam, Figure 3.2 shows differences in performance for each topic with histograms of the differences in achievement between each topic's Assignment and Exam question (*Difference = Exam − Assignment*). Most students' achievement scores have decreased from the assignment to the exam (the majority of the occurrences are negative). For all topics the tallest bin is between -10% and 10%, indicating that many students retain relatively consistent achievement in each topic across the two assessment items.

It must be noted that since the difference in achievement is calculated by taking the assignment score away from the exam score, if a student has achieved perfect marks (100%) in the assignment, it is not possible for them to improve on this, and so they cannot get a positive difference between the assignment

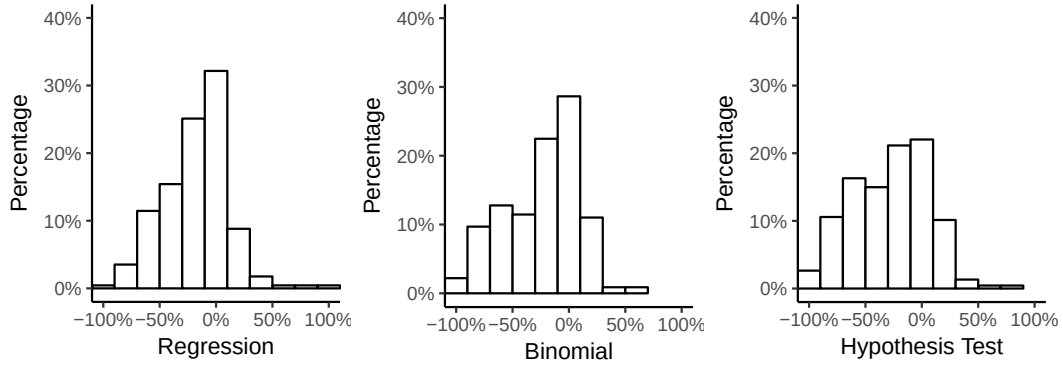

Figure 3.2: Histograms of differences between assignment and exam achievement for each chosen topic (*Difference = Exam – Assignment*)

and the exam. However, if a student has achieved poor marks in the assignment, there is more opportunity to perform comparatively better on the exam. This may be contributing to the low frequencies of positive differences from the histograms. Similar trends to Semester 3 are displayed in Semesters 1 and 2 (refer to Figures B.2 and C.2 in Appendices B and C).

To check whether there is a significant difference between the assignment and exam scores for each topic, t-tests were used. Table 3.1 gives the p-values of each test, both parametric (t-test) and non-parametric (Wilcoxon Signed-Ranks test). The differences between assignment and exam achievement are statistically significant for all topics. This, in conjunction with Figure 3.2, indicates that the performance on the invigilated exam is poorer in comparison to the assignments for a substantial number of students. These findings were reflected in the other semesters (refer to Tables B.1 and C.1).

Table 3.1: p-values for testing of differences between assignment and exam scores in the three topics

Assessment Topics	Parametric	Nonparametric
Regression	1.22E-16	1.43E-16
Binomial	9.73E-23	6.83E-20
Hypothesis Test	5.48E-28	4.14E-23

3.2 Relationships between data sources

Distribution of Demographic Data

The distribution of study mode by type of degree for the 227 students who were enrolled in the introductory statistics course in Semester 3 is presented in Table 3.2. Most noticeable is that the total number of external students enrolled (83%, n=188) is much greater than the total number of on-campus students (17%, n=39). The degrees studied by students are represented in four groups: the combined group of business, commerce, law and information technology (BusComLawIT) (43%, n=98), psychology (13%, n=29), science (27%, n=62) and other degrees (17%, n=38). For all Degree Types except BusComLawIT, the total on-campus enrolment count for each is under ten students, representing only 17% of the cohort in total. The combined BusComLawIT degrees have the highest enrolment of all degree types, with 43% of total enrolment for Semester 3.

When compared with the frequency tables for the Semester 1 and Semester 2 cohorts (refer to Table B.2 and Table C.2 respectively), it can be seen that a smaller proportion of the Semester 3 cohort studied a degree in the BusComLawIT category than that of the other cohorts; in Semester 3, 43% of students studied under the BusComLawIT category, while in Semester 1 this category consisted of 58% of the cohort and in Semester 2 this category made up 63%

Table 3.2: Distribution of Degree Type by Study Mode

Degree Type	External	On-campus Springfield	Total
BusComLawIT	77	21	98 (43%)
Psychology	22	7	29 (13%)
Science	58	4	62 (27%)
Other	31	7	38 (17%)
Total	188 (83%)	39 (17%)	227 (100%)

of the cohort. Conversely, the Other category made up 17% of the Semester 3 cohort, while only comprising 7% of the Semester 1 cohort and 5% of the Semester 2 cohort.

The proportion of external students was similar between Semester 3 (83%) and Semester 1 (78%). However, in Semester 2, this proportion changed substantially, with only 48% of students enrolled externally (the remainder divided between enrolling on-campus at Toowoomba (35%) or Springfield (16%)).

Access Records by Study Mode

Learning analytics in this research context is dependent on the ability to track access to resources by students throughout the teaching period. Of particular interest is exploring the access habits of students in different demographic groups. The most relevant demographic variable available to access habits is study mode, as different study modes may be linked to different levels of engagement with the course. Figure 3.3 shows the distribution of access for each OLS object chosen for this project (tutorial solutions for each topic as well as example exams), with external and on-campus cohorts shown side-by-side for comparison.

Figure 3.3 highlights that a large percentage of both external and on-campus students do not access the tutorial solutions, however a very small proportion of students do not access the example exams. It is also evident that the external cohort (188 students) consistently have higher percentage access rates for the OLS objects compared to the on-campus cohort (39 students). The on-campus access rates for the three tutorial solutions are 49% (Regression), 46% (Binomial) and 54% (Hypothesis Test), indicating that around half of the on-campus cohort do not access the tutorial solutions, even though they cannot access this specific resource via on-campus attendance. However, it should be noted that students who attend tutorials are given feedback in class if they request it.

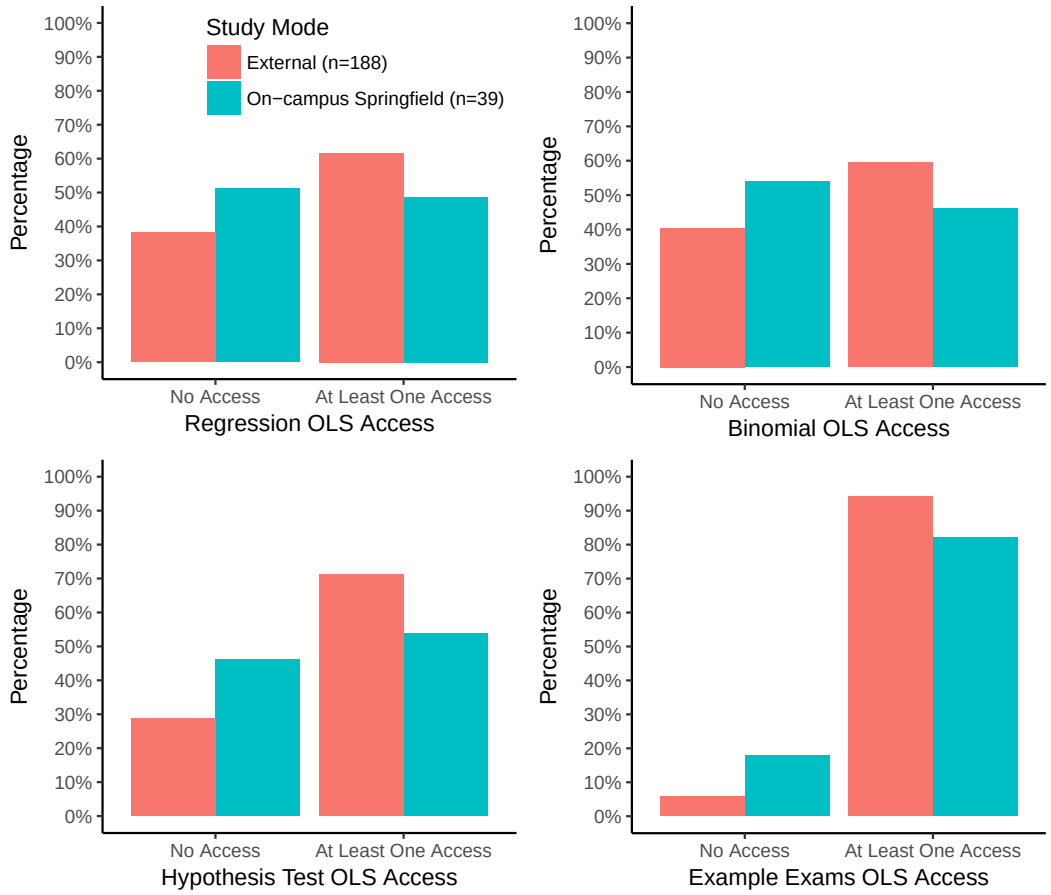

Figure 3.3: Distribution of OLS access by study mode; OLS objects include tutorial solutions for the three chosen topics and the example exams

When comparing these results with that of the Semester 1 and 2 cohorts, most of the results are similar, with a higher percentage of on-campus students having not accessed the resources than external students (refer to Figures B.3 and C.3). A notable difference between the on-campus Springfield cohorts of Semester 2 and Semester 3 is the higher proportion of students in Semester 2 who did not access the OLS objects. In particular, 44% of Semester 2 on-campus Springfield students did not access the example exams on the OLS, which is substantially higher than the 18% of on-campus Springfield students in Semester 3 and 12% of on-campus Toowoomba students in Semester 1 who did not access the example exams.

OLS Access by Demographic Data

It is important for course examiners to know the frequency at which their cohort is interacting with the learning material. Also of interest is the access habits of different subgroups within the cohort, to enable understanding of how students in differing circumstances may interpret the role of the OLS in supporting their learning.

Table 3.3 shows the percentage of students from each degree type who accessed each OLS object at least once throughout the semester. The example exams OLS object was the most frequently accessed OLS object for all degree types, with access being above 89% compared with access to all other objects of

Table 3.3: OLS access (percentage) for each degree type

OLS Objects	BusComLawIT (n=98)	Psychology (n=29)	Science (n=62)	Other (n=38)
Regression	60	62	56	61
Binomial	64	52	48	58
Hypothesis Test	68	72	65	71
Example Exams	92	97	92	89

interest not being higher than 73%. Table 3.3 also indicates that the hypothesis test topic tutorial solutions were the most popular tutorial solutions across all degree types. Out of all the degree types, the students in the Science degree category tended to access the least frequently, with the lowest percentage access for all OLS objects except for the example exams.

When comparing these OLS access frequencies with those from Semesters 1 and 2 (see Tables B.3 and C.3), it can be seen that the only consistent trend is that the example exams are always accessed more frequently than the other OLS objects. Of interest is the stark difference between Semesters 2 and 3 in terms of the access to example exams. For Semester 3, the example exams access rates (regardless of Degree Type) were always above 89%, however for the Semester 2 cohort, the example exams access rates were consistently between 70% and 80%. This indicates a cohort-wide trend of reduced access to the example exams in Semester 2.

Note that Figure 3.3 and Table 3.3 are the only places where access to the example exams OLS object is presented – in all other parts of this project, only the access records for the tutorial solutions for the chosen topics are used.

Distribution of Assessment Data by Study Mode

The boxplots in Figure 3.4 enable comparisons between the external and on-campus cohorts for achievement in each of the three chosen topics in both the assignments and the exam. Note that the external enrolment was 188 students and the on-campus enrolment was 39 students. It can be seen that the median score in all assessment items for the three topics is consistently higher (to varying degrees) for the external cohort than the on-campus cohort. This is also true for both the lower and upper quartiles of achievement scores between external and on-campus students. The largest differences can be seen in the Binomial and Hypothesis Test exam questions, where the median score for external students is about 20% higher than that of the on-campus students.

Also evident from Figure 3.4 is the considerably larger interquartile ranges of the exam question achievement scores when compared to the corresponding assignment questions, indicating that there is more spread in the exam results. Students (regardless of study mode) tend to perform relatively well on assignment questions (with the lower quartile of assignment achievement scores located at or above 60%). In the exam questions, however, students are more likely to perform at a lower standard on average, with the lower quartile often being located between 20% and 40%. The lower quartile of the hypothesis test exam question for on-campus students was located at 0%, indicating that this questions was most likely not attempted by many of the on-campus students.

The spread (measured by the interquartile range) of the binomial assignment achievement scores in Semester 1 is greater than that of the Semester 3 binomial assignment achievement scores, for both on-campus and external students. In addition, achievement in the hypothesis test exam question by the Semester 1 cohort was generally much higher than that of the Semester 3 cohort, again for all study modes. Apart from these differences, all other results are extremely similar between the two semesters (refer to Figure B.4).

The trends found in the Semester 2 achievement data are also similar to those found in the Semester 3 results. A notable aspect of the Semester 2 results (see Figure C.4) is the performance of on-campus Springfield students in the exam questions – most students tended to perform extremely poorly, especially when compared to achievement in the assignment questions.

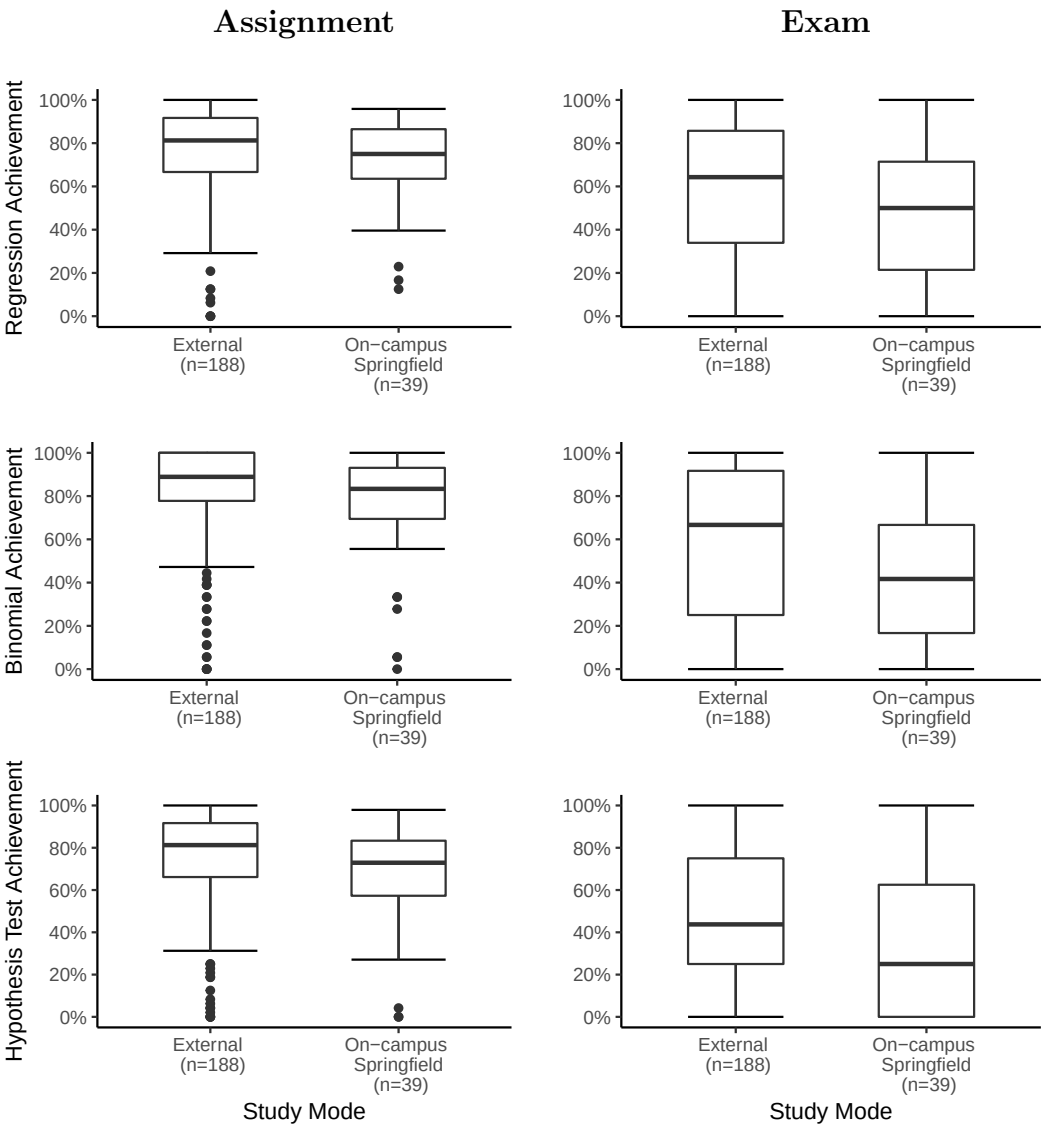

Figure 3.4: Distribution of achievement on each topic for both assignment and exam questions by study mode; assignment questions on the left, exam questions on the right. Note that the example exams OLS access logs are not considered here.

Distribution of Assessment Data by OLS access categories

Figure 3.5 shows similar plots to those in Figure 3.4, however the boxplots are categorised by the number of times the OLS object had been accessed, rather than study mode. The analysis of assessment and OLS data in combination is a critical aim of the application of learning analytics in this research project.

In most of the plots in Figure 3.5, it can be seen that the students who did not access the OLS object at all performed, on average, slightly worse than those who accessed once or more. With the exception of the Binomial Assignment question, the median achievement of students who never accessed the relevant tutorial solutions is lower than that of students who accessed the tutorial solutions at least once.

Similar to Figure 3.4, the interquartile ranges of the exam achievement scores are larger than that of the assignment achievement scores, indicating higher variability in student achievement during the exam. There are many outliers in the lower end of achievement scores for the assignment questions regardless of frequency of OLS access. Also evident is that the difference in median achievement when comparing students who did not access to students who accessed at least once is much larger for the exam questions compared to the assignment questions. This indicates that accessing the OLS tutorial solutions at least once may lead to better performance on invigilated assessment at the end of the semester.

The corresponding figures for Semesters 1 and 2 (refer to Figures B.5 and C.5 respectively) are very similar to the results shown in Figure 3.5. The only major difference is the achievement of students in the hypothesis test exam question between Semesters 1 and 3; students in Semester 1 performed much better on this question than the Semester 3 students.

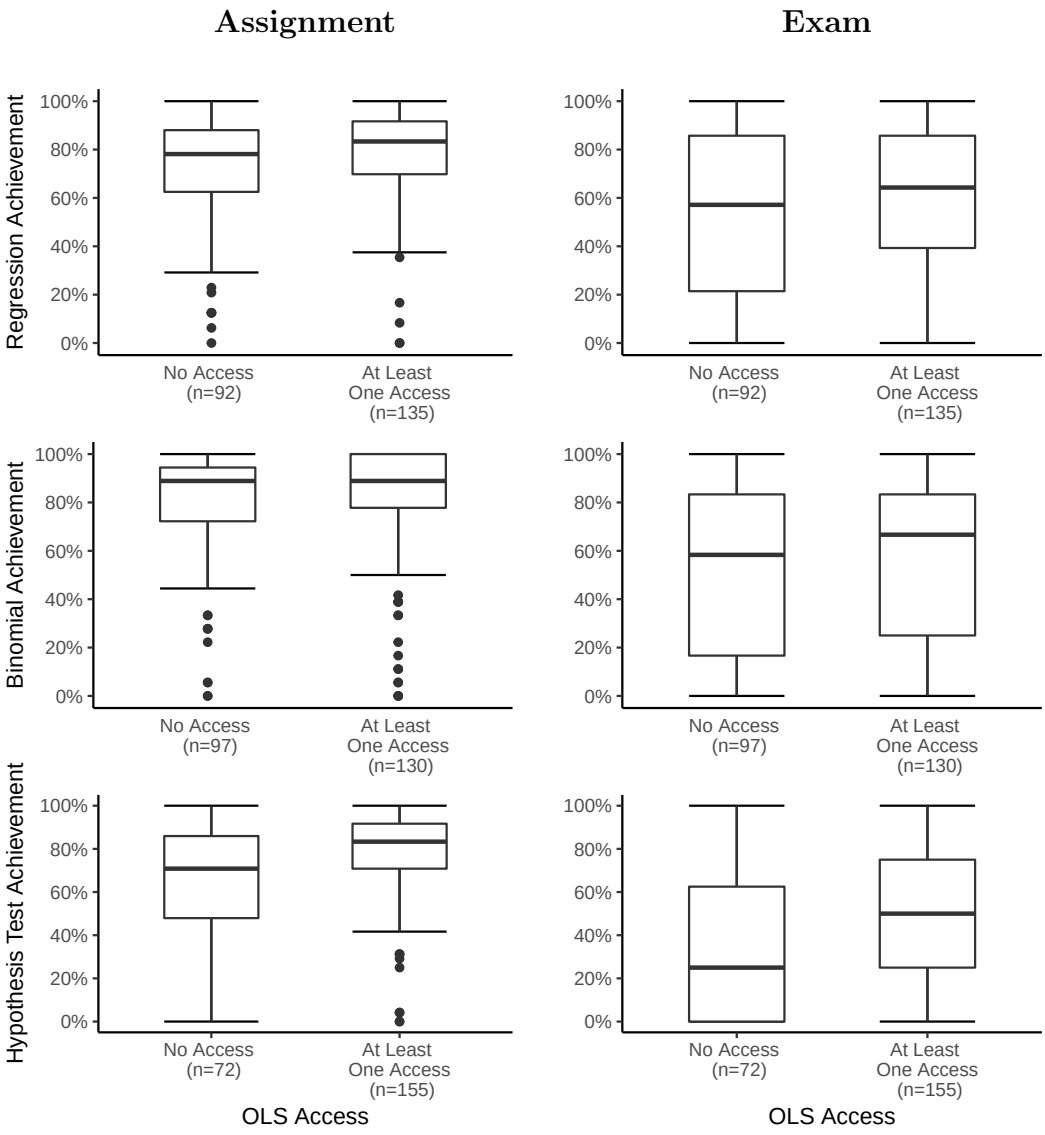

Figure 3.5: Distribution of achievement on each topic for both assignment and exam questions by the frequency of access to tutorial solutions; assignment questions on the left, exam questions on the right. Note that the example exams OLS access logs are not considered here.

3.3 Multivariate Analyses

Cluster Analysis

Figure 3.6 indicates that using two clusters provides the highest silhouette width, so two clusters were used in this analysis. The clusters can be visualised using the 2D ordination plot as shown in Figure 3.7. Although the two clusters neighbour each other closely with no gap in-between, they are still recognisable as separate clusters. There is some overlap of the clusters, with a few cases occurring well within the opposite cluster to which it was assigned. This may be due to variation of those cases from the other cases in the cluster on a couple of the original 6 continuous and 5 categorical variables included in the calculation of the distance matrix. There are very few cases that considerably overlap the cluster boundary compared with the total number of cases.

Within each cluster, the distances between cases in both dimensions are a representation of the average distances between cases among all variables. Therefore, the medoids for each cluster (the cases central to each cluster) will not be reported as they represent the average characteristic of the cluster only and do not provide perspective on the variation within each of the closely neighbouring clusters. It is more appropriate instead to use summary statistics of the clusters as measures of their overall characteristics.

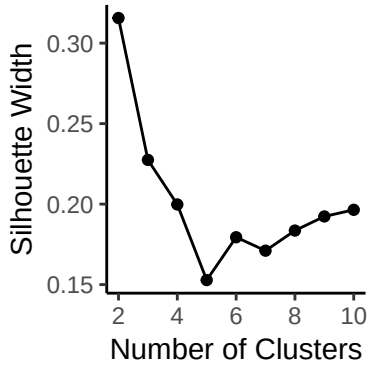

Figure 3.6: Silhouette Width for different numbers of clusters

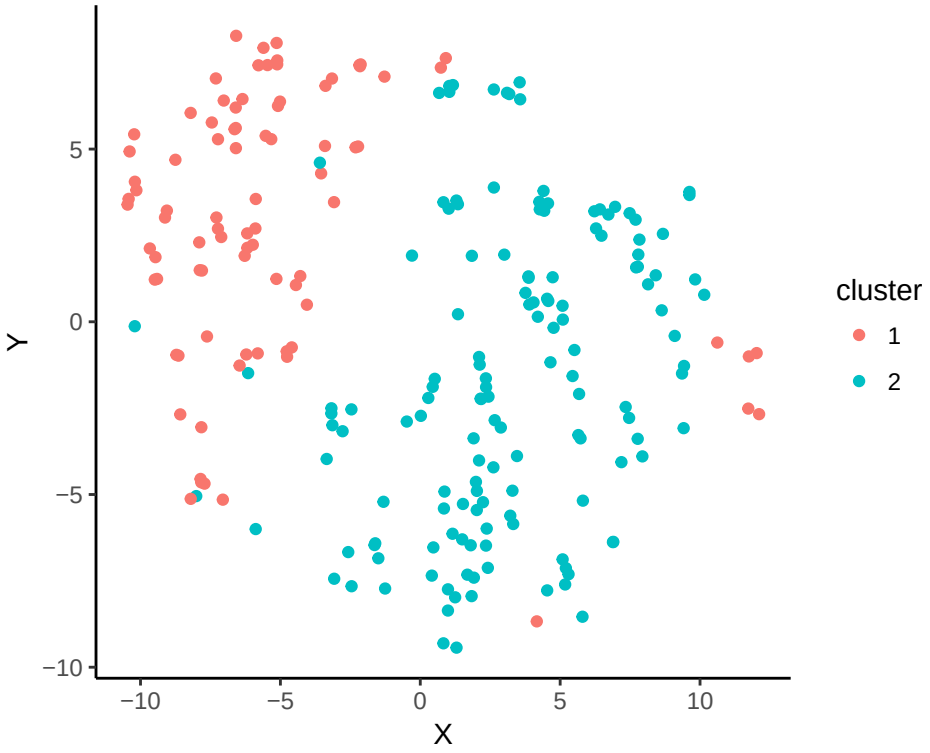

Figure 3.7: 2D ordination plot of aggregate distances between cases, with cases coloured by assigned cluster

Cluster Analysis – Summary Statistics on Clusters

Table 3.4 shows that the clusters seem to be primarily influenced by OLS access; for cases with "No Access" for the three tutorial solutions, the majority of cases are placed in Cluster 1, and for cases with "At Least One Access" for the three tutorial solutions, most of the cases are located in Cluster 2. It is important to note that there were more cases assigned overall to Cluster 2 (n=142) than Cluster 1 (n=85). For most demographic categories, there are noticeably more students in Cluster 2 than in Cluster 1 (the degree types "BusComLawIT", "Psychology" and "Other", and the "External" study mode). For the other demographic categories, the distribution of students in each clusters is roughly equal between the two clusters (the degree type "Science" and the "On-campus Springfield" study mode). It can also be seen that the mean achievement level is always higher for Cluster 2 than Cluster 1, with the Binomial Exam question being the biggest difference in average achievement

(13.1%) between the two clusters.

Figure 3.8 displays the relationship between cluster membership and OLS access. It can be seen that the majority of students in Cluster 1 (around 85% for Regression, around 90% for Binomial and around 70% for Hypothesis Test) are students who never accessed the respective tutorial solutions, and that the majority of students in Cluster 2 (around 85% for Regression, around 85% for Binomial, and around 90% for Hypothesis Test) are students who accessed the respective tutorial solutions at least once.

Table 3.4 also shows that there are noticeably more external students in Cluster 2 than in Cluster 1 (66 students in Cluster 1, 122 students in Cluster 2), further supporting the trend shown in Figure 3.3 that a greater proportion of external students access the OLS resources (since Cluster 2 is mostly comprised of students who accessed at least once, see Figure 3.8).

The cluster analysis on the datasets from Semesters 1 and 2 showed very similar results to the Semester 3 analysis. Two clusters were identified in each dataset, and were shown to be distinct despite closely neighbouring each other (see Figures B.7 and C.7). As with Semester 3, the primary defining factors between clusters in Semesters 1 and 2 were access to the OLS resources, and the cluster comprised mostly of students who accessed the OLS resources always had a higher average achievement in each assessment question, as well as a larger number of external students, than the other cluster (see Tables B.4 and C.4).

Table 3.4: Frequencies of Degree Type, Study Mode and OLS Access, and mean and standard deviation of achievement in assessment questions, by Cluster

Variables	Labels	Cluster 1 n=85	Cluster 2 n=142	Total n=227
Degree Type	BusComLawIT	31	67	98
	Psychology	9	20	29
	Science	30	32	62
	Other	15	23	38
Study Mode	External	66	122	188
	On-campus Springfield	19	20	39
Regression Access	No Access	73	19	92
	At Least One Access	12	123	135
Binomial Access	No Access	78	19	97
	At Least One Access	7	123	130
Hypothesis Test Access	No Access	61	11	72
	At Least One Access	24	131	155
Assessment Questions		Mean (SD)	Mean (SD)	
	Reg Assignment	70.2% (22.5%)	77.6% (20.6%)	
	Bin Assignment	73.7% (29.9%)	82.9% (21.2%)	
	HT Assignment	65.5% (28%)	77.3% (22.5%)	
	Reg Exam	53.2% (32.2%)	59.9% (30.4%)	
	Bin Exam	47.7% (34.8%)	60.8% (32.7%)	
	HT Exam	42.4% (35.3%)	46.6% (34.7%)	

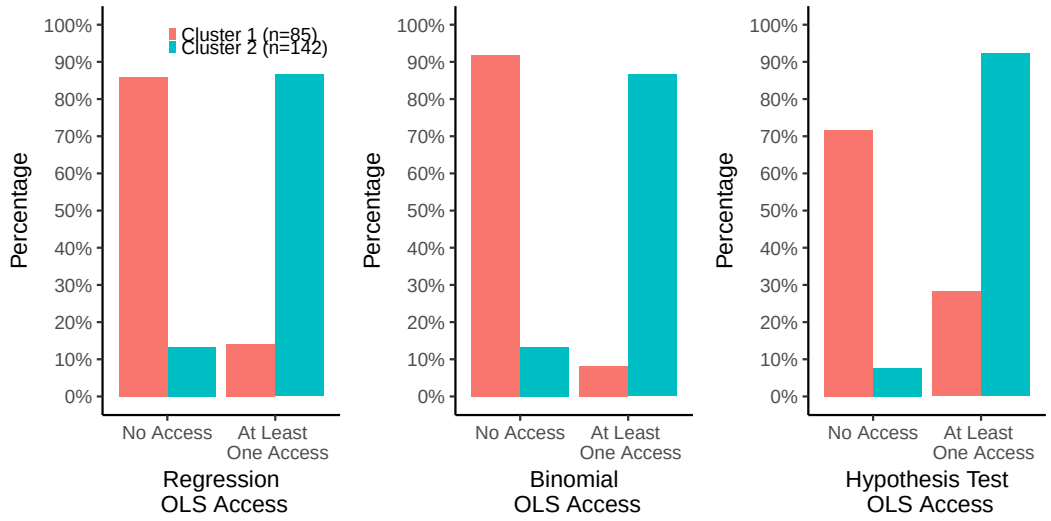

Figure 3.8: OLS access by cluster membership

Principal Components Analysis

The individual and cumulative percentage variation explained by each principal component are given in Table 3.5, and the loadings of each original variable on each principal component are given in Table 3.6. Table 3.5 shows that the first two principal components cumulatively explain 72% of the total variation in the data. The loadings given in Table 3.6 show that each of the original variables are moderately weighted on the first principal component, while for the second principal component the assignment variables are weighted positively, and the exam variables are weighted negatively.

Visual representations of the first two principal components (both the individuals factor map and the variables factor map) are given in Figure 3.9. The individuals factor map (left plot) shows that the cohort had much greater spread with respect to the first principal component (the x-axis), especially

Table 3.5: Percentage variation explained by each principal component

	PC1	PC2	PC3	PC4	PC5	PC6
% Variation Explained	0.57	0.16	0.09	0.08	0.06	0.05
Cumulative % Variation Explained	0.57	0.72	0.81	0.89	0.95	1.00

Table 3.6: Loadings of the original variables on each principal component

	PC1	PC2	PC3	PC4	PC5	PC6
Regression Assignment	0.41	0.30	0.65	-0.01	0.57	-0.02
Binomial Assignment	0.41	0.43	0.20	0.18	-0.75	-0.06
Hypothesis Test Assignment	0.37	0.51	-0.68	-0.22	0.25	0.17
Regression Exam	0.39	-0.42	0.11	-0.78	-0.19	0.13
Binomial Exam	0.42	-0.42	-0.09	0.52	0.04	0.61
Hypothesis Test Exam	0.44	-0.33	-0.24	0.22	0.10	-0.76

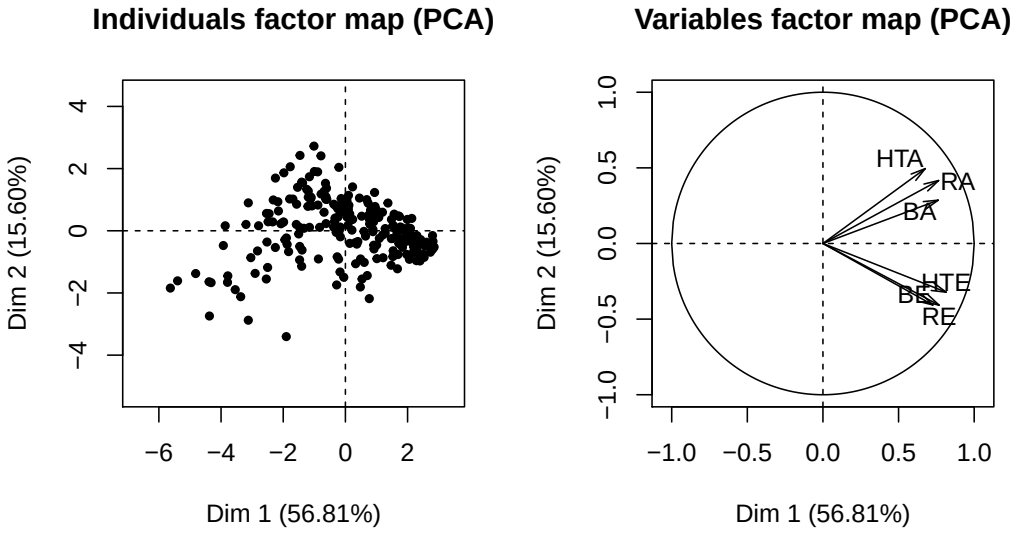

Figure 3.9: Individuals (left) and Variables (right) factor maps displaying both the cases and the variable vectors against the principal components. Vectors represent the three topics (R, B and HT) with labels ending in either A (Assignments) or E (Exams)

for cases in the bottom left quadrant. There was also a reasonable amount of spread with respect to the second principal component, with students in the bottom left quadrant again displaying the most variation. The variables factor map (right plot) in Figure 3.9 visualises the variable loadings on the principal components from Table 3.6 and confirms that all continuous variables are weighted in the same direction with respect to the first principal component; this principal component can therefore be interpreted as an "overall achievement" component. The second principal component differentiates to some extent between achievement in the assignment and achievement in the exam. The vectors on the variables factor map, in conjunction with the individuals factor map, indicates which variables were separating cases at different extremes of the map from the rest of the group.

Students on the left hand side of the x-axis (Figure 3.9) performed poorly overall in the assessment items, according to the first principal component. Extending the assignment vectors back into the bottom left quadrant, it can be seen that the students in this quadrant stand out from the group due to their underperformance in the assignment questions. Likewise, students who performed poorly in the exam questions are located in the top left quadrant. Comparing the top left quadrant to the bottom left quadrant, it can be seen that the majority of cases in the top left quadrant are closer to the centre of the cohort than those in the bottom left, indicating that it is more commonplace within the whole cohort to perform badly in the exam questions than in the assignment questions. It is also evident that many students in the bottom left quadrant are further to the left on the x-axis than students in the top left quadrant, highlighting that students who underperform on the assignment questions tend to perform worse in the course overall, with respect to the first principal component.

The majority of students who performed extremely well (furthest to the right on the x-axis) are located mostly in the bottom right quadrant. This indicates

that the factor separating these students from the group is their outstanding performance on the exam questions. Students who performed well on the assignment questions (located in the top right quadrant) are very close to the centre of the group, which shows that high performance on the assignments is not only a common occurrence within the cohort, but also does not distinguish student performance from the cohort average compared to high achievement in the exam questions.

The weightings and loadings of principal components are extremely similar between Semester 3 and the other semesters, indicating that the principal components have identified robust trends in the data. Refer to Tables B.5, B.6, C.5 and C.6, as well as Figures B.9 and C.9 for the comparison. For example, across all three semesters the cumulative percentage of total variation explained by the first two principal components is between 70% and 74%, and the first two principal components have similar weightings (all the original variables are moderately weighted against the first principal component, and the assignment question variables are oppositely weighted against the second principal component when compared to the exam question variables). Across all three semesters, the combination of the individuals and the variables factor maps showed that low performance in the assignment made cases stand out considerably from the group. The only notable difference in the PCA for the three semesters is that the students in Semester 2 who performed badly on the exam questions were more separated from the rest of the group than in Semesters 1 and 3 (see Figures B.9 and C.9).

The supplementary variables were added to the individuals factor map in Figure 3.10, with the individual cases removed to aid interpretation. Figure 3.10 shows the average positioning of each categorical variable with respect to the principal components. With respect to the first principal component, external students performed better on average than on-campus students (given that the external students were further to the right of the zero line on the plot),

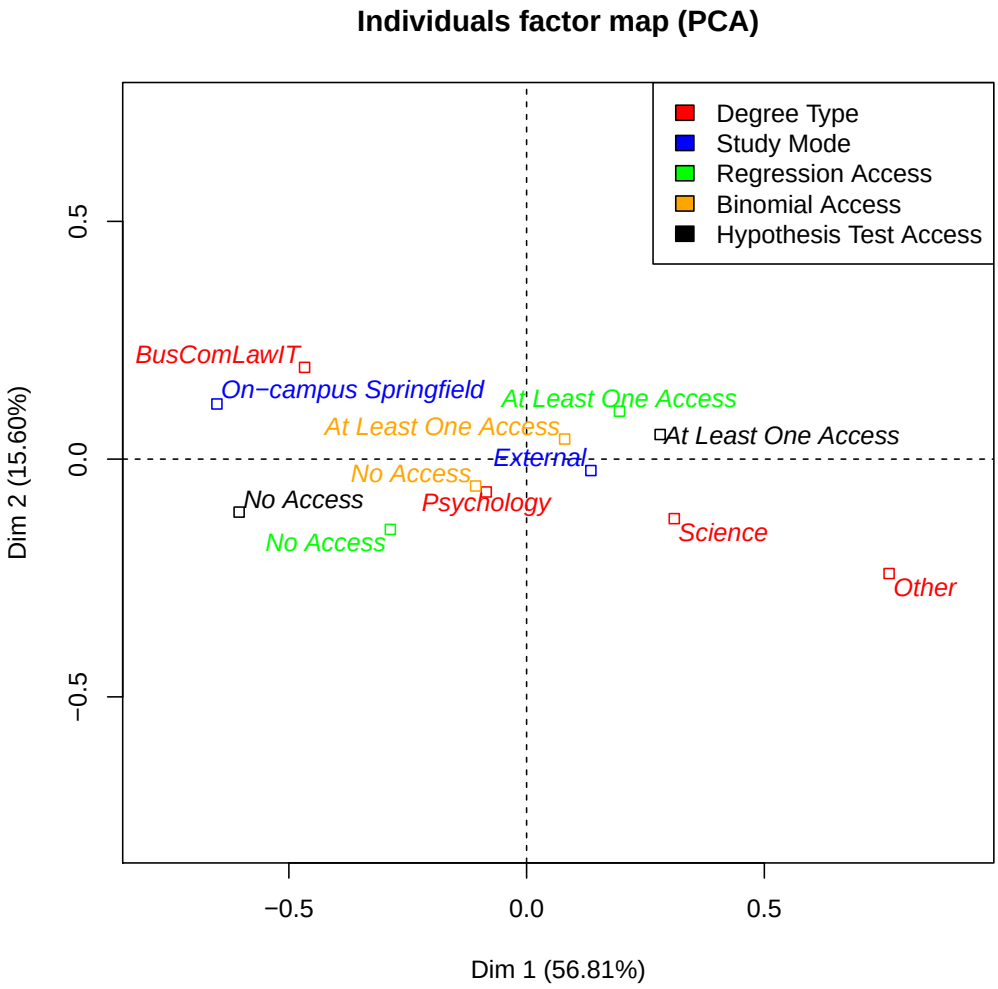

Figure 3.10: Coloured PCA Plot with Degree Type included as a qualitative supplementary variable. Note there was no On-campus Toowoomba enrolment option in Semester 3

students in the degree types 'Science' and 'Other' performed better than 'BusComLawIT' and 'Psychology' students, and students who access the OLS resources performed better than students who do not. With respect to the second principal component, it can be seen that students in the 'Other' and 'Science' degree type categories performed well on exams compared to the other degree types.

The qualitative supplementary variables in both Semester 1 and 2 also show that students who do not access the OLS resources for each topic consistently performed worse with respect to the overall achievement principal component (PC1) than students who accessed them at least once. Students in the 'Science' degree type achieved highly on exams just as they did in Semester 3, however students in the 'Other' category varied in their performance (average in Semester 1 and poor in Semester 2).

It should be noted that in Figures 3.10, B.10 and C.10, the scales on both the x- and y-axes are quite small, compared to the overall scales on the original plots. This indicates that the distances between the average positions of the categorical variables are not very large and so some caution is required in interpretation.

Chapter 4

Discussion and Conclusions

This chapter discusses the major findings in the cohort identified by the summary statistics, then discusses the results and utility of both the descriptive statistics and multivariate statistical models used in this project, and finally gives an overview of the potential interventions identified from, and limitations of, the conclusions drawn from the analyses. This will be expressed within the context of the service course analysed in this research, both from the point of view of the service course itself within the university, and where appropriate also from the perspective of a wider application of the techniques and methodologies used in this project to other scenarios outside of the specific context of this research. The implications and limitations of the data collation process (including cleaning and merging of data) on data analytics in the education field will also be discussed, as these combined data are not generally readily available.

Some of the key issues that will be discussed include the identification of true engagement of students with the course/learning materials, the early identification of students who are at risk of performing poorly, the improvement of underperforming students through intervention, and approaches to learning for both on-campus and external students.

4.1 Trends and relationships in the data

There was a clear disparity between average student performance in the assignments and in the final exam (see Figures 3.1 and 3.2). The major functional differences between these two types of assessment is the time that they are due (assignments are due throughout the semester, while the final exam is at the end) and the restrictions imposed on access to materials and time (assignments impose no restriction, whereas the final exam poses both a restriction on reference materials and a time limit of two hours). It is plausible that these differences contributed to the overall differences in achievement. In the past, high levels of anxiety have been reported by students in regards to the invigilated examination, which may affect performance. In the final examination, although material is restricted, students are allowed to use a double-sided A4 sheet of their own notes to aid them, as well as statistical formulae and tables that are provided with the examination paper. The main restriction imposed by the exam, therefore, is that it is invigilated and limited by time. The analysis of assessment data in combination with both demographic and OLS data enabled an insight into some other factors that may be influencing this difference in achievement in the context of this particular undergraduate statistics course, which will be discussed further throughout this chapter.

There was a higher proportion of students enrolled externally compared to on-campus students in Semesters 1 and 3, however in Semester 2 the proportions of external and on-campus students were approximately equal. Most full-time recommended enrolment patterns for programs at the university suggest undertaking this statistics course in Semester 2, and there are two on-campus offerings in Semester 2 to facilitate this intake of students. Full-time students are more likely to be studying on-campus, as the choice to study full-time indicates that they may have fewer commitments outside of study (such as full-time work), and so are more likely to study in Semester 2 as per the enrolment pattern recommendations and the availability of two campuses for

on-campus enrolment.

The BusComLawIT degree type was the most popular category in all semesters. The most popular degrees at the university either do not require completion of this statistics course, or are degrees that fall under the BusComLawIT category, so this result was not unexpected as the proportion of students enrolled in a BusComLawIT program is likely to be higher than that of the other degree types.

The proportion of the cohort that did not access the tutorial solutions was above 25% for all topics across all three semesters and all study modes, which is substantial given that students cannot access these resources elsewhere (see Figure 3.3 and Table 3.3). It may be that students complete the tutorials and then do not bother to download the tutorial solutions – they may assume that they are correct, or feel that they do not need to check their work. This may be more prevalent among science students, as their higher average level of quantitative skills may cause them to be overconfident in their abilities, and therefore to not access the tutorial solutions as frequently on average as students in other degrees (see Table 3.3). Other suggestions for the cause of non-access by students have been put forward by Kortemeyer (2017); students may fall behind throughout the course, and so resort to techniques such as cramming, guessing or copying, which would reduce the rate of access to tutorial solutions. The rate of access to example exams was very high in contrast to the tutorial solutions for all cohorts, further indicating that students potentially fall behind and must catch up by cramming, i.e. choosing resources they perceive will provide them with the shortest route to the knowledge needed to pass the exam. Due to the second hurdle implemented in the final exam (students must achieve at least 40% on the final exam in order to pass the course), students may also be cramming to meet this requirement.

On-campus students in particular may feel that attending the on-campus tutorials provides them with the feedback they need, and so checking with the

tutorial answers is less necessary, which may explain the higher proportion of on-campus students who did not access the tutorial solutions (Figure 3.3). This trend is consistent across semesters, indicating that there may be a fundamental difference in learning behaviours that may be influenced by study mode. Moreira (2016) has stated that online learning is still inferior to on-campus learning, highlighting that educators are slow to afford the same status to online learning. In this project, however, external students are engaging more with the course content than on-campus students. In addition to this, Figure 3.4 showed that on-campus students generally perform at a lower average achievement level and with less consistency than external students in the final examination questions (this is consistent with the relationship between OLS access and achievement described earlier). These results also contrast with the findings of McGready & Brookmeyer (2013) that both on-campus and external students perform relatively equally on assessment. McGready & Brookmeyer (2013) highlighted that investing in support for external students is now a mainstream practice and will continue to be emphasised so that any disadvantages of studying externally are minimised. The results in this project indicate that online learning at this particular university has progressed to the point at which external students have the same opportunities to achieve as on-campus students, so much so that they can now perform better despite not being on-campus. The cause for this reversal of achievement trend may be attributable to the learning behaviours encouraged by study mode, and the quality of the resources that are more frequently accessed by external students due to these learning behaviours (including many not addressed in this research). Another potential factor supporting the high performance of external students is that the university that is the focus of this research is a long-standing distance education institution, with an extensive history of educational support to external students.

All cohorts generally performed poorer on the final examination than on the assignments, as mentioned previously (see Figures 3.4 and 3.5). Although there

were still students who achieved highly in the final examination questions, there was more variability in student performance in the exams compared to the assignments, indicating that many students dropped in their performance levels in the exam. Figure 3.5 in particular shows that there was greater spread in performance on all assessment questions for students who did not access the tutorial solutions. The median achievement for students who did not access was also lower than that of the students who did access, across all assessment questions. The OLS access logs, however, are limited in that they do not provide information on *how* students engaged with the material, only whether or not they downloaded it. This therefore limits the conclusions that can be drawn from this trend. Despite this, the general underperformance on invigilated assessment, as well as the association between low academic performance and disengagement with the learning materials, could still form the basis of further research at a finer scale, possibly related to retention and attrition and their relationship with accessing materials and attending classes.

Importantly, there were many cases of underperformance in the assignment questions that were identified as outliers (shown in both Figures 3.4 and 3.5). From the perspective of assisting underperforming students, it is important that these students are able to be identified at an early stage in the teaching period. Although the contributing factors to this underperformance cannot be identified from the data in this research, it is known that all students included in the research completed all assessment items (and therefore completed the course). From this, it is clear that this particular group of students is a potential target for interventions aimed at improving academic performance.

4.2 Multivariate Methods

The OLS access variables played a major role in the generation of the clusters, with the assignment of cases to the two clusters being largely characterised by OLS access for all three semesters (see Table 3.4). Students who are in a particular cluster are more similar to each other (based on the distance matrix generated from the original variables) than they are to students in the other cluster. OLS access was shown to be linked to academic achievement based on the descriptive statistics (section 4.1), and the cluster analysis further supported this relationship (see Table 3.4). This table of results, along with the tables for Semesters 1 and 2 (Tables B.4 and C.4) showed that the mean achievement in the assessment achievement variables was consistently higher across cohorts for the clusters which mostly contained students who accessed the tutorial solutions (Cluster 2 for Semester 1, Cluster 1 for Semester 2, and Cluster 2 for Semester 3). The close proximity of the two clusters (see the ordination plot in Figure 3.7) indicates that some students who do not access the OLS resources share similar characteristics to students in the opposing cluster, and vice versa. It is possible that students who do not access the tutorial solutions are still able to achieve well by concentrating on other OLS resources or accessing resources elsewhere, just as it is possible that students who access the tutorial solutions continue to perform poorly due to a lack of true engagement with the resources. Although there is some overlap in the clusters in the ordination plot, the overall relationship between OLS access and assessment achievement is substantial, as the two clusters are still visibly distinct from each other in all three semesters (see Figures B.7 and C.7 for the Semester 1 and Semester 2 ordination plots).

The first principal component identified in the principal components analysis of all three semesters describes the overall achievement of students in all assessment questions, and explains above 50% of the total variation in the data for all semesters (see Tables 3.5, B.5 and C.5). This is consistent with

most of the variation in assessment achievement coming from students' overall performance; this could be analogous to the overall skill or effort level of the students, or their overall capacity to achieve. The second principal component distinguishes between student performance in the two different types of assessment: assignments and exams. Once overall skill has been accounted for, it is likely that the second highest source of variation comes from students who perform uncharacteristically well or poorly on a particular type of assessment, in comparison with their overall achievement pattern. Cumulatively, the first two principal components account for over 70% of the total variation across all semesters, which is a strong result given that it has occurred in all three separate datasets.

The positioning of the cases on the individuals factor map, with respect to the vectors in the variables factor map, indicate that high performance in the exam questions (bottom right quadrant) and low performance in the assignment questions (bottom left quadrant) set a subset of students apart from the group (see Figure 3.9). These students are identified as distinct from the main group of students, because with respect to the first principal component, they represent the best performing and the worst performing students overall, respectively. This aligns with the generally high achievement in assignments and lower achievement in exams shown in Figures 3.4 and 3.5. Of the students who are underperforming overall (left on the x-axis or negative on the first principal component), it is clear that those in the bottom left quadrant (those who underperform on the assignment questions) are furthest to the left on the x-axis, and are further away from the centre of the group than the students who underperform on the exam questions. These students may therefore be the best group to target when implementing strategies to improve student performance. Not only do they have the most potential to improve, but underperformance on assignment questions can be identified during the teaching period, making improvement for these students a possibility.

Both cluster analysis and principal components analysis allowed the incorporation of all types of variables (demographic, assessment and OLS) into the one model, which informed how the variables relate to each other. Cluster analysis involved the creation of a distance matrix based on all variables, and so the final ordination plot represented the distances between each case, which could not be related back to the original variables. With this in mind, principal components analysis may provide a better insight into the relationships between variables, as the principal components are linear combinations of the original variables and are therefore still relatable to them (Manly, 2005). On the other hand, the principal components analysis method used in this research restricts the usable variables in the formation of principal components to continuous variables, and any categorical variables must be added as supplementary variables afterwards (Lê *et al.*, 2008). However, using Gower distance in cluster analysis allows all types of variables to be incorporated into the distance matrix, so the categorical variables play a role in determining the distances between cases (Kalisch, 2012).

The use of these multivariate statistical methods within the context of learning analytics has shown to be informative and has added to the insights gained regarding the relationships between different sources of data. As such, there is potential for these methods to be used in future learning analytics research. The application of these methods, however, was complex, and may be difficult to implement and interpret for researchers or educators who do not have a statistical background (Greller & Drachsler, 2012). The principal components analysis, in particular, provided a detailed insight into the nature of assessment achievement through the variables factor map, and also facilitated interpretation of the relationship between categorical variables and the principal components. The use of multivariate statistical methods requires the ability to both run and interpret the analyses. As such, it may not be practical for multivariate statistical methods to become more widely used in learning analytics research (Greller & Drachsler, 2012). However, if the aim is to extend

learning analytics to incorporate multiple sources and types of data, it would be difficult to jointly consider these without implementation of some form of multivariate analysis.

4.3 Potential interventions

As evidenced by both the summary statistics and the multivariate analyses, it is beneficial in terms of achievement for students to access the course resources. It is therefore recommended, if possible, that course facilitators monitor the access logs of their cohort, to check whether there are students who are consistently not accessing the course material. This research has shown that even a simple binary indicator of access was enough to draw a connection between disengagement with course material and the group of students who are more likely to perform poorly on assessments. As such, the process of monitoring student OLS access for educators using the same OLS as in this project would require only to match the list of names in the access logs against the list of names in the demographic database.

Another potential intervention is related to student performance on the final examination – it is important that students are encouraged to maintain engagement up to the end of the teaching period. It is vital that invigilated assessment is maintained in the course despite the reduction in student performance when compared with the assignments – an invigilated test of knowledge on exit from the course is necessary to maintain standards of academic integrity and ensure that students are exiting with the essential skills they are expected to acquire as part of the course. It may also be appropriate to incorporate another invigilated assessment during the semester, firstly as an indicator of students who may be struggling with the course, and secondly to offer familiarity with the process for students who may be feeling anxious about invigilated assessment, allowing them to practise the process and stay on track with their

studies.

The results of the principal components analysis showed that overall performance accounts for the highest amount of variation in the datasets, and that poor performance in the assignment questions often resulted in the worst overall performance. This suggests that a potential valuable intervention may involve (if possible) course facilitators firstly checking to see if students who perform poorly on early assessment items are engaging with course material properly, and secondly attempting to make contact with these students to encourage them to seek help and improve. Students who are underperforming in the middle of the teaching period still have the opportunity to learn and improve for later assessment items, but if they are not accessing OLS resources, this may indicate that they have disconnected from the learning journey (Khalil & Ebner, 2017). Making contact with these students could provide them with an opportunity to re-engage with the course material and seek help. However, for this particular course this approach has been implemented in previous years (not during the teaching periods analysed in this research), and the rate of student response was quite low (pers.comm. several previous and current examiners of the course). This may suggest that the poor performance of these students was not course related, as they did not respond to offers of additional assistance with their studies. It is acknowledged that lack of engagement and underachievement may be due to factors in the students' lives that are completely unrelated to this course and its content, or their usual approach to learning.

4.4 Limitations and future research

There are many limitations that need to be taken into consideration when drawing conclusions from the analyses in this project. The scale of this project restricted the amount of data used to a small fraction of the total data avail-

able through the OLS, and also necessitated the simplification of this data (for example, the OLS access logs provide the exact times and dates that a resource was accessed by students throughout the entire semester, however this information was reduced to a binary variable for each student indicating whether or not they accessed the resource). The need to track student learning behaviours and achievement over the entirety of the teaching period excluded the use of cases where not all assessment items had been completed, which was a significant number of students.

Another limitation with learning analytics data in general is that access logs will never be able to provide a true representation of the engagement of students with the course material. A record of a student having downloaded a particular resource does not give information on *how* that student used the resource; they could have read through it extensively and repeatedly, or simply glanced at it for ten seconds before discarding it. These logs also cannot provide a true indication of the usefulness of these resources – there may be several reasons why students download or do not download a resource, independent of its quality. It is also possible for students to use other resources outside of the OLS to supplement or even replace the content available to them through the OLS, necessitating the acquisition of access logs of all resources to obtain a complete picture of the learning journeys of students (Zacharis, 2015). As such, when using this data, researchers must be very careful when interpreting trends involving OLS access.

The acquisition of assessment data for this project was very resource-intensive, requiring extensive time to be dedicated to data entry and cleaning. Educators attempting to implement learning analytics while also continuing to teach may find that it is not feasible to take on this extra workload, especially if the assessment data needs to be manually collated (Dyckhoff *et al.*, 2012). As a result of this project, changes have been made to the marking routine for the undergraduate course in statistics that was analysed; markers are now

required to store assessment data in a Microsoft Excel spreadsheet as they mark, to mitigate the occurrence of arithmetic errors. This development would significantly reduce the amount of time required for implementation of the methods used in this project, and so further research on this course and in situations where assessment data is readily available would be much easier to pursue.

The comparison between topics across the teaching period in this project was limited by the change in assessment types. For this course, a particular topic would only be assessed twice; once in an assignment, and once in the final exam. Due to the change in assessment types, a true judgement of the progression of students throughout the teaching period is hindered, as the change in assessment types could contribute to differences in achievement in these topics. For a true comparison to be made, the assessment type should be held constant.

Due to these limitations, future research could be performed on additional data from the OLS – for the OLS analysed in this project, access is available to citizenship demographic data, as well as access logs for any other OLS resource of interest. There is also the possibility of using more detailed assessment data (part marks for questions) to investigate the nuances of students' understanding of particular topics, and which parts of a topic they may succeed or perform poorly on. Now that the analytical methods used in this project have been applied and trailled for this specific learning analytics context, the formulation of more specific research questions and consequent downloading of data relevant to these could lead to potential future research.

Future research could also be conducted on the role of the OLS in the quality of data available for analysis. Since so much of the data available is automatically downloaded by (and therefore predetermined by) the OLS used for course delivery, it is worth either investigating the data collected by OLSs that are already available for use, or designing and developing an OLS so that the data

collected can be selected. Of course, the latter would be difficult in contexts where the OLS used is standardised across the educational facility, however this approach could potentially be implemented in less rigorous contexts.

Finally, the difference in academic performance between external and on-campus students identified in this project warrants future research into the learning behaviours of on-campus students, specifically how rate of attendance at the lectures and tutorials affects performance. The factors surrounding on-campus underperformance could be better understood by gathering data on how often students interact with the on-campus learning sessions, and how this compares with the levels of interaction shown by external students through the OLS.

4.5 Conclusions

This research project investigated the use of learning analytics techniques to explore the learning journeys of students in an undergraduate service course in statistics at an Australian university. Data from both the online learning system (OLS) and assessment items were collated and merged for the analysis, which involved both manual data entry and coding to combine the databases. Trends between achievement and both demographic and OLS access data were identified through descriptive statistical analyses, and multivariate statistical methods were used to enable the combination of these different types of data into single analyses. The major findings of this project include that, on average, students who are studying externally, students who are doing a degree under the "Science" degree type, and students who regularly access the course tutorial solutions are less likely to underperform on assessment items. These results not only provide an insight into the effects of the learning behaviours of students on their academic performance, but also may form the basis of future research on the underlying factors that may be influencing academic

underperformance.

References

Bainbridge, J., Melitski, J., Zahradnik, A., Laura, E. J. M., Jayaprakash, S., & Baron, J. (2015). Using learning analytics to predict at-risk students in online graduate public affairs and administration education. *Journal of Public Affairs Education* **21**, 247 – 262.

Buttner, E. H. & Black, A. N. (2014). Assessment of the effectiveness of an online learning system in improving student test performance. *Journal of Education for Business* **89**, 248 – 256.

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education* **18**, 683–695.

Cocea, M. & Weibelzahl, S. (2011). Disengagement detection in online learning: Validation studies and perspectives. *IEEE Transactions on Learning Technologies* **4**, 114–124.

de Freitas, S., Gibson, D., Du Plessis, C., Halloran, P., Williams, E., Ambrose, M., Dunwell, I., & Arnab, S. (2015). Foundations of dynamic learning analytics: Using university student data to increase retention. *British Journal of Educational Technology* **46**, 1175–1188.

Dringus, L. P. (2012). Learning analytics considered harmful. *Journal of Asynchronous Learning Networks* **16**, 87 – 100.

Dyckhoff, A. L., Zielke, D., Bltmann, M., Chatti, M. A., & Schroeder, U. (2012). Design and implementation of a learning ana-

lytics toolkit for teachers. *Journal of Educational Technology Society* **15**, 58 – 76.

Ellis, C. (2013). Broadening the scope and increasing the usefulness of learning analytics: The case for assessment analytics. *British Journal of Educational Technology* **44**, 662–664.

Ferguson, R. (2012). The state of learning analytics in 2012: A review and future challenges. *Knowledge Media Institute, Technical Report KMI-2012-01* .

Gasevic, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education* **28**, 68 – 84.

Giannakos, M. N., Sampson, D. G., & Kidziński, Ł. (2016). Introduction to smart learning analytics: foundations and developments in video-based learning. *Smart Learning Environments* **3**, 12.

Gibson, D. & de Freitas, S. (2016). Exploratory analysis in learning analytics. *Technology, Knowledge and Learning* **21**, 5–19.

Greller, W. & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Journal of Educational Technology Society* **15**, 42 – 57.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: concepts and techniques*. Elsevier/Morgan Kaufmann, Amsterdam;Boston;, 3rd edition.

Hernández-García, A. & Conde, M. A. (2014). Dealing with complexity: Educational data and tools for learning analytics. In *Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM '14, pages 263–268, New York, NY, USA. ACM.

Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014). Developing early warning systems to predict students online learning performance. *Computers in Human Behavior* **36**, 469 – 478.

Kalisch, M. (2012). Lecture slides in measuring distance - multidimensional scaling.

Khalil, M. & Ebner, M. (2017). Clustering patterns of engagement in massive open online courses (moocs): the use of learning analytics to reveal student categories. *Journal of Computing in Higher Education* **29**, 114–132.

Kortemeyer, G. (2017). The spectrum of learning analytics. *eleed* **12**.

Krijthe, J. H. (2015). *Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*. R package version 0.13.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software* **25**, 1–18.

Linting, M. & van der Kooij, A. (2012). Nonlinear principal components analysis with catpca: A tutorial. *Journal of Personality Assessment* **94**, 12–25. PMID: 22176263.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2017). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.6 — For new features, see the 'Changelog' file (in the package source).

Manly, B. F. J. (2005). *Multivariate statistical methods: a primer*. Chapman Hall/CRC Press, Boca Raton, FL, 3rd edition.

Martin, F. & Whitmer, J. C. (2016). Applying learning analytics to investigate timed release in online learning. *Technology, Knowledge and Learning* **21**, 59–74.

McGready, J. & Brookmeyer, R. (2013). Evaluation of student outcomes in online vs. campus biostatistics education in a graduate school of public health. *Preventive Medicine* **56**, 142 – 144.

Moreira, D. (2016). From on-campus to online: A trajectory of innovation, internationalization and inclusion. *International Review of Research in Open Distance Learning* **17**, 186 – 199.

Nyland, R., Davies, R. S., Chapman, J., & Allen, G. (2016). Transaction-level learning analytics in online authentic assessments. *Journal of Computing in Higher Education* pages 1–17.

Papamitsiou, Z. & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology Society* **17**, 49 – 64.

Pardo, A. & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology* **45**, 438 – 450.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Romero-Zaldivar, V.-A., Pardo, A., Burgos, D., & Kloos, C. D. (2012). Monitoring student progress using virtual appliances: A case study. *Computers Education* **58**, 1058 – 1067.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53 – 65.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Serrano-Laguna, A., Torrente, J., Moreno-Ger, P., & Fernndez-Manjn, B. (2014). Application of learning analytics in educational videogames. *Entertainment Computing* **5**, 313 – 322.

Siemens, G. (2012). Learning analytics: Envisioning a research discipline and a domain of practice. In *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 4–8, New York, NY, USA. ACM.

Siemens, G. (2013). Learning analytics. *American Behavioral Scientist* **57**, 1380–1400.

Slade, S. & Prinsloo, P. (2013). Learning analytics. *American Behavioral Scientist* **57**, 1510–1529.

Soby, M. (2014). Learning analytics. *Nordic Journal of Digital Literacy* pages 89–91.

Spector, P. (2011). Class notes on cluster analysis and R.

Strang, K. D. (2016). Do the critical success factors from learning analytics predict student outcomes?. *Journal of Educational Technology Systems* **44**, 273 – 299.

Strang, K. D. (2017). Beyond engagement analytics: which online mixed-data factors predict student learning outcomes? *Education and Information Technologies* **22**, 917–937.

Sutton, S. C. & Nora, A. (2008). An exploration of college persistence for students enrolled in web-enhanced courses: A multivariate analytic approach. *Journal of College Student Retention: Research, Theory & Practice* **10**, 21–37.

West, D., Huijser, H., & Heath, D. (2016). Putting an ethical lens on learning analytics. *Educational Technology Research and Development* **64**, 903–922.

Wickham, H., Francois, R., Henry, L., & Mller, K. (2017). *dplyr: A Grammar of Data Manipulation*. R package version 0.7.2.

Xie, Y. (2017). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.17.

Yassine, S., Kadry, S., & Sicilia, M. A. (2016). A framework for learning analytics in moodle for assessing course outcomes. In *2016 IEEE Global Engineering Education Conference (EDUCON)*, pages 261–266.

You, J. W. (2016). Identifying significant indicators using {LMS} data to predict course achievement in online learning. *The Internet and Higher Education* **29**, 23 – 30.

Zacharis, N. Z. (2015). A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *The Internet and Higher Education* **27**, 44 – 53.

Appendix A

Code for running analyses

Example code for histograms:

```
A2Hist <- ggplot(S3SumData, aes(x = S3SumData$A2_TOT)) +
   geom_histogram(aes(y = ..count../sum(..count..)),
    binwidth = 0.1, col = "black", fill = "white") + labs(x
       = "First Assignment",
    y = "Percentage") + scale_x_continuous(limits = c(0, 1),
    breaks = seq(0, 1, 0.25), labels = percent) +
       coord_cartesian(xlim = c(0,
    1)) + scale_y_continuous(limits = c(0, 0.4), breaks =
       seq(0,
    0.4, 0.1), labels = percent) + theme_bw() + theme(panel.
       border = element_blank(),
    panel.grid.major = element_blank(), panel.grid.minor =
       element_blank(),
    axis.line = element_line(colour = "black"))
```

Example code for t-tests:

```
# Parametric t-test

HTDiffPara <- t.test(S3SumData$A3_HTQPerc, +
   S3SumData$E_HTQPerc,
    paired = TRUE)

# Non-parametric t-test

HTDiffNonPara <- wilcox.test(S3SumData$A3_HTQPerc, +
   S3SumData$E_HTQPerc,
    paired = TRUE)
```

Example code for bar charts:

```
RegBarChartDB <- as.data.frame(c(levels(S3SumData$RegAccess)
   , + levels(S3SumData$RegAccess)))

names(RegBarChartDB)[1] <- "Access_Rate"

RegBarChartDB$Access_Rate <- factor(
   RegBarChartDB$Access_Rate, + levels=c("No Access","At
   Least One Access"))

RegBarChartDB$StudyMode <- c(rep("External", 2),rep("On-
   campus Springfield", 2))

RegBarChartDB$AccessPerc <- NULL

RegBarChartDB$AccessPerc[1] <- sum(S3SumData$RegAccess=="No
   Access" & S3SumData$StudyMode=="External")/sum(
   S3SumData$StudyMode=="External")

RegBarChartDB$AccessPerc[2] <- sum(S3SumData$RegAccess=="At
   Least One Access" & S3SumData$StudyMode=="External")/sum(
   S3SumData$StudyMode=="External")

RegBarChartDB$AccessPerc[3] <- sum(S3SumData$RegAccess=="No
   Access" & S3SumData$StudyMode=="On-campus Springfield")/
   sum(S3SumData$StudyMode=="On-campus Springfield")

RegBarChartDB$AccessPerc[4] <- sum(S3SumData$RegAccess=="At
   Least One Access" & S3SumData$StudyMode=="On-campus
   Springfield")/sum(S3SumData$StudyMode=="On-campus
   Springfield")


RegBar <- ggplot(RegBarChartDB, aes(Access_Rate, y=
   AccessPerc, fill = StudyMode))

RegBar <- RegBar + geom_bar(position="dodge", stat="identity
   ") + labs(x="Regression OLS Access", y="Percentage", fill
   ="Study Mode")

RegBar <- RegBar + scale_y_continuous(limits = c(0,1),
   breaks = seq(0, 1, .1), labels = percent) +
   scale_fill_discrete(labels=c("External (n=188)", "On-
   campus Springfield (n=39)"))

RegBar <- RegBar + theme_bw() + theme(panel.border =
   element_blank(), panel.grid.major = element_blank(),
   panel.grid.minor = element_blank(), axis.line =
   element_line(colour = "black"), legend.justification=c
   (1,1), legend.position=c(1,1))
```

Example code for boxplots:

```
RegABoxplot <- ggplot(S3SumData, aes(StudyMode, A2_RegQPerc)
   )

RegABoxplot <- RegABoxplot + stat_boxplot(geom = "errorbar",
    width = 0.5) + geom_boxplot(width = 0.5) +
       scale_y_continuous(limits = c(0,
    1), breaks = seq(0, 1, 0.2), labels = percent)

RegABoxplot <- RegABoxplot + labs(x = " ", y = "Regression
   Achievement")

RegABoxplot <- RegABoxplot + theme_bw() + theme(panel.border
    = element_blank(),
    panel.grid.major = element_blank(), panel.grid.minor =
       element_blank(),
    axis.line = element_line(colour = "black"))

RegABoxplot <- RegABoxplot + scale_x_discrete(labels = c(
   External = "External \n (n=188)",
    `On-campus Springfield` = "On-campus \n Springfield \n (
       n=39)"))
```

Example code for cluster analysis:

```
library(dplyr)  # for data cleaning

library(cluster)  # for gower similarity and pam

library(Rtsne)  # for t-SNE plot


set.seed(1680)  # for reproducibility

gower_dist <- daisy(S3Cluster, metric = "gower")

gower_mat <- as.matrix(gower_dist)


# use PAM (partitioning around medoids) cluster algorithm on

# distance matrix Calculate silhouette width for many k
   using

# PAM

sil_width <- c(NA)

for (i in 2:10) {
    pam_fit <- pam(gower_dist, diss = TRUE, k = i)
    sil_width[i] <- pam_fit$silinfo$avg.width
}

sil_data <- as.data.frame(cbind(c(2:10), sil_width[2:10]))
```

```
# Silhouette width plot code

SilWidthPlot <- ggplot(data = sil_data, aes(x = sil_data[,
   1],
    y = sil_data[, 2])) + geom_line() + geom_point()

SilWidthPlot <- SilWidthPlot + labs(x = "Number of Clusters
   ",
    y = "Silhouette Width") + theme_bw() + theme(panel.
       border = element_blank(),
    panel.grid.major = element_blank(), panel.grid.minor =
       element_blank(),
    axis.line = element_line(colour = "black"))


pam_fit <- pam(gower_dist, diss = TRUE, k = 2)

pam_results <- S3Cluster %>% # dplyr::select() %>%

mutate(cluster = pam_fit$clustering) %>% group_by(cluster)
   %>%
    do(the_summary = summary(.))


# Plot the cluster analysis

tsne_obj <- Rtsne(gower_dist, is_distance = TRUE)

tsne_data <- tsne_obj$Y %>% data.frame() %>% setNames(c("X",
    "Y")) %>% mutate(cluster = factor(pam_fit$clustering),
       name = S3Cluster$DegreeType)

ClusterPlot <- ggplot(aes(x = X, y = Y), data = tsne_data) +
    geom_point(aes(color = cluster)) + theme_bw() + theme(
       panel.border = element_blank(),
    panel.grid.major = element_blank(), panel.grid.minor =
       element_blank(),
    axis.line = element_line(colour = "black"))
```

Example code for PCA:

```
library(FactoMineR)

colnames(S3Cluster)[6:11] <- c("RA", "BA", "HTA", "RE", "BE
   ",
    "HTE")  #Rename variables for plotting

# Perform PCA

res.pca2 <- PCA(S3Cluster[, c(1:5, 6:11)], scale.unit = TRUE
   ,
    ncp = 3, quali.sup = c(1:5), graph = F)

par(mfrow = c(1, 2))
```

```
PCAInd <- plot.PCA(res.pca2, axes = c(1, 2), choix = "ind",
   label = c("none"),
    invisible = "quali")

PCAVec <- plot.PCA(res.pca2, axes = c(1, 2), choix = "var")

par(mfrow = c(1, 1))
```

Appendix B

Semester 1 Results

B.1 Assessment Achievement

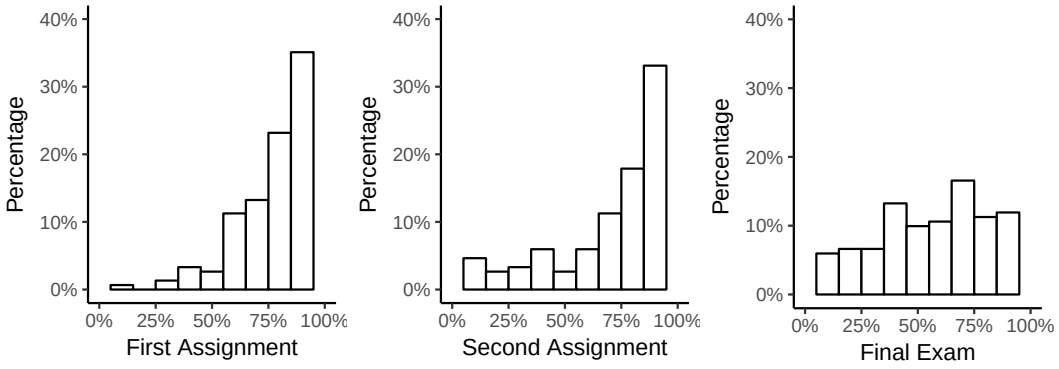

Figure B.1: Distribution of overall achievement for each assessment item

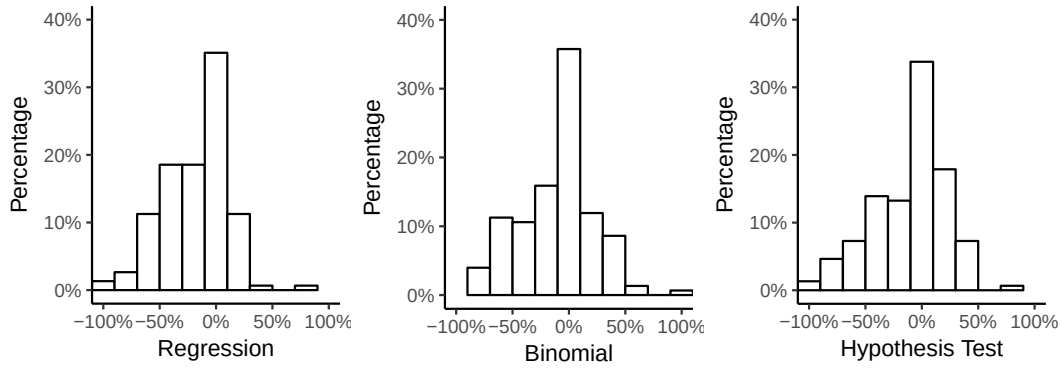

Figure B.2: Histograms of differences between assignment and exam achievement for each chosen topic (*Difference = Exam – Assignment*)

Table B.1: p-values for testing of differences between assignment and exam scores in the three topics

Assessment Topics	Parametric	Nonparametric
Regression	5.04E-12	1.40E-10
Binomial	6.57E-04	1.64E-03
Hypothesis Test	8.14E-05	1.28E-03

B.2 Relationships between data sources

Table B.2: Distribution of Degree Type by Study Mode

Degree Type	External	On-campus Toowoomba	Total
BusComLawIT	67	21	88 (58%)
Psychology	11	6	17 (11%)
Science	31	5	36 (24%)
Other	9	1	10 (7%)
Total	118 (78%)	33 (22%)	151 (100%)

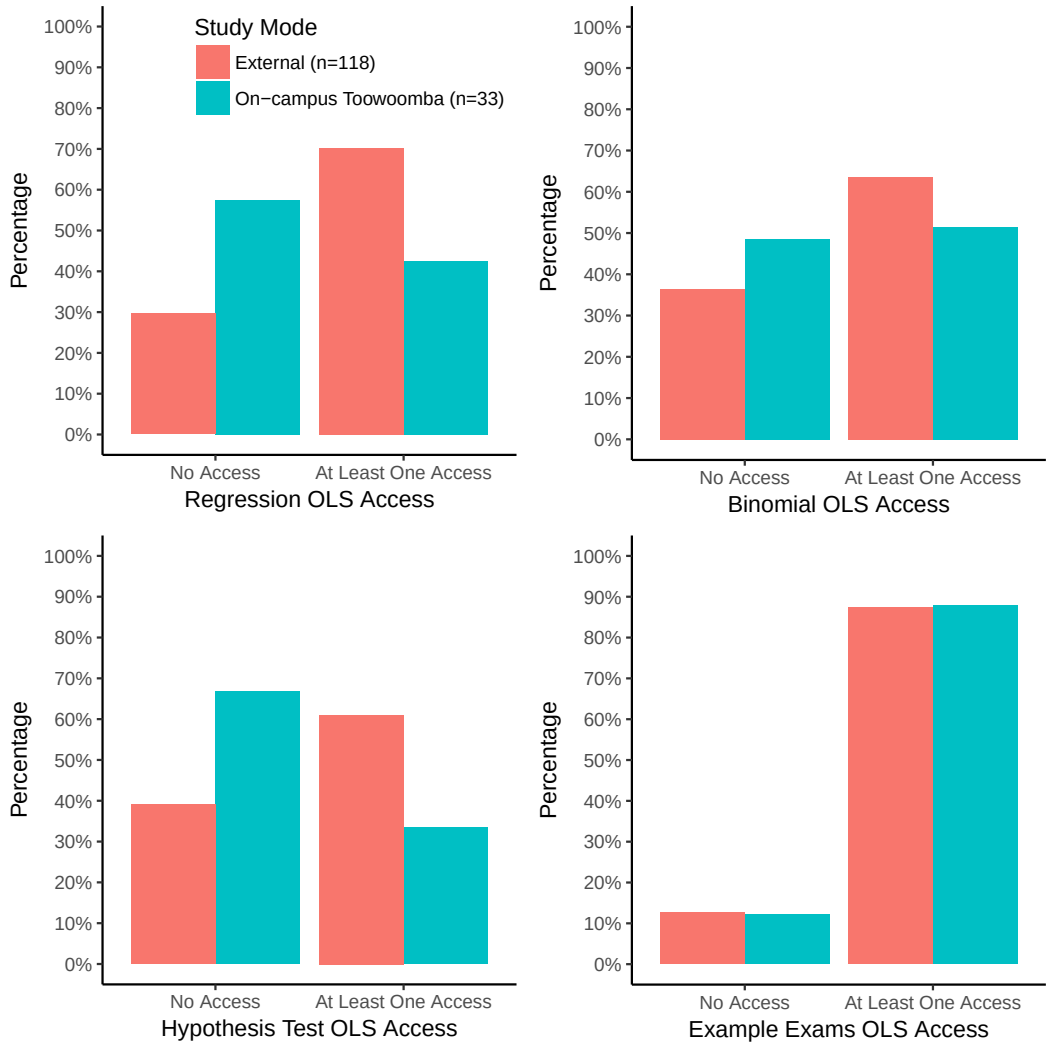

Figure B.3: Distribution of OLS access by study mode; OLS objects include tutorial solutions for the three chosen topics and the example exams

Table B.3: OLS access (percentage) of cohort for each degree type

OLS Objects	BusComLawIT (n=88)	Psychology (n=17)	Science (n=36)	Other (n=10)
Regression	63	65	69	60
Binomial	56	76	72	40
Hypothesis Test	55	41	67	40
Example Exams	83	100	89	100

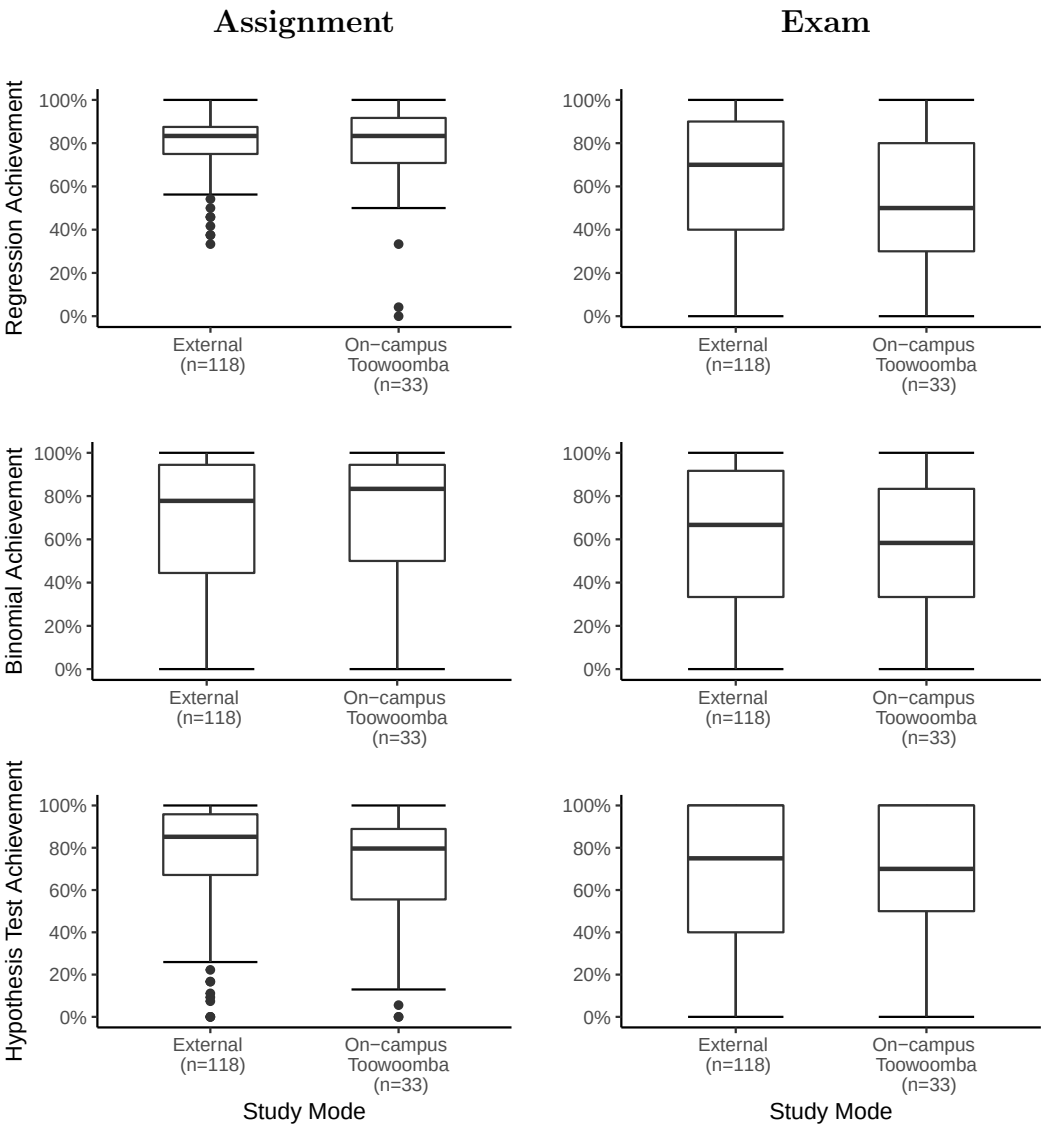

Figure B.4: Distribution of achievement on each topic for both assignment and exam questions by study mode; assignment questions on the left, exam questions on the right. Note that the example exams OLS access logs are not considered here.

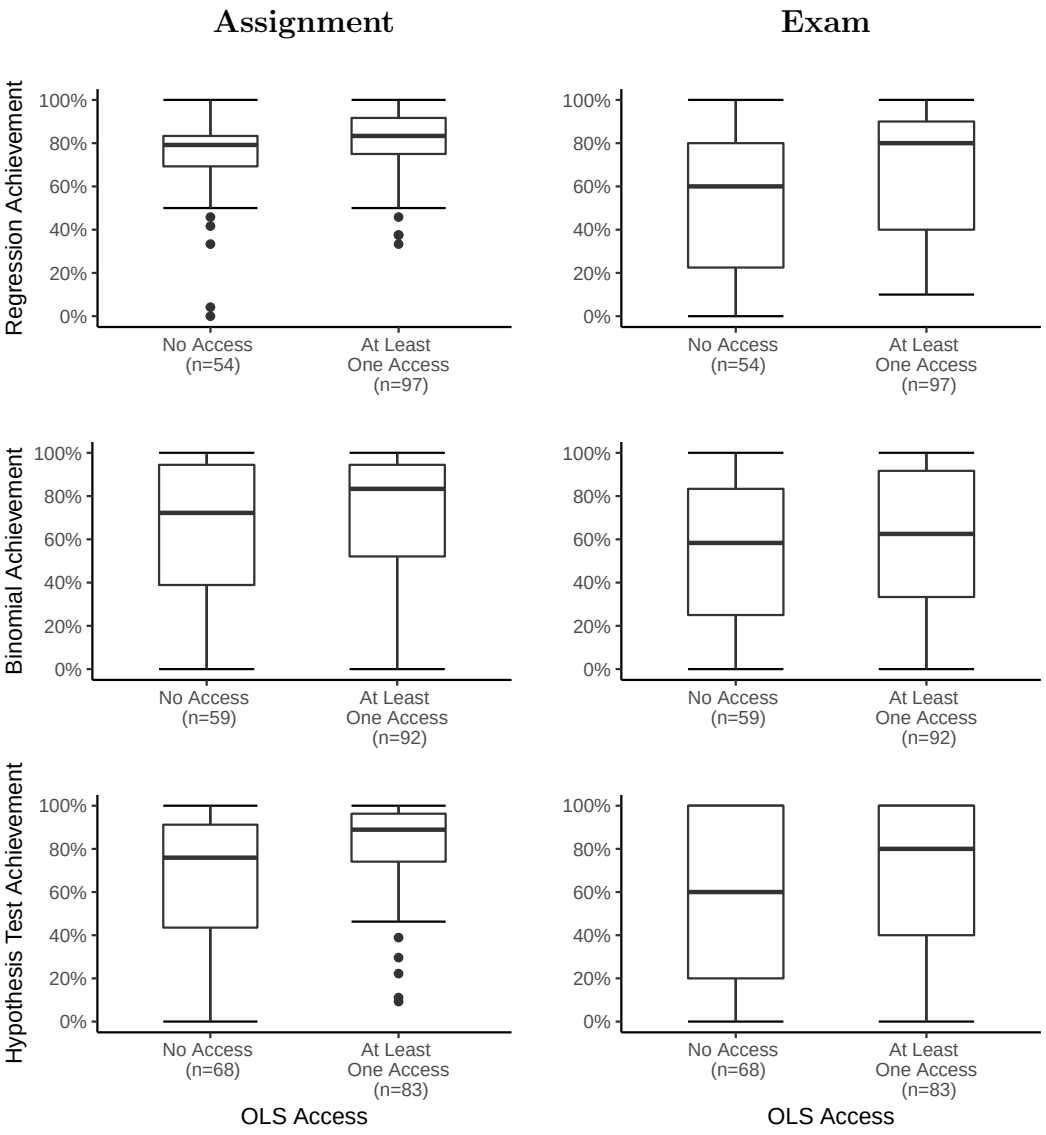

Figure B.5: Distribution of achievement on each topic for both assignment and exam questions by the frequency of access to tutorial solutions; assignment questions on the left, exam questions on the right. Note that the example exams OLS access logs are not considered here.

B.3 Multivariate Analyses

Cluster Analysis

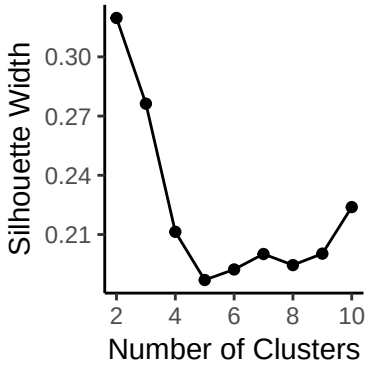

Figure B.6: Silhouette Width for different numbers of clusters

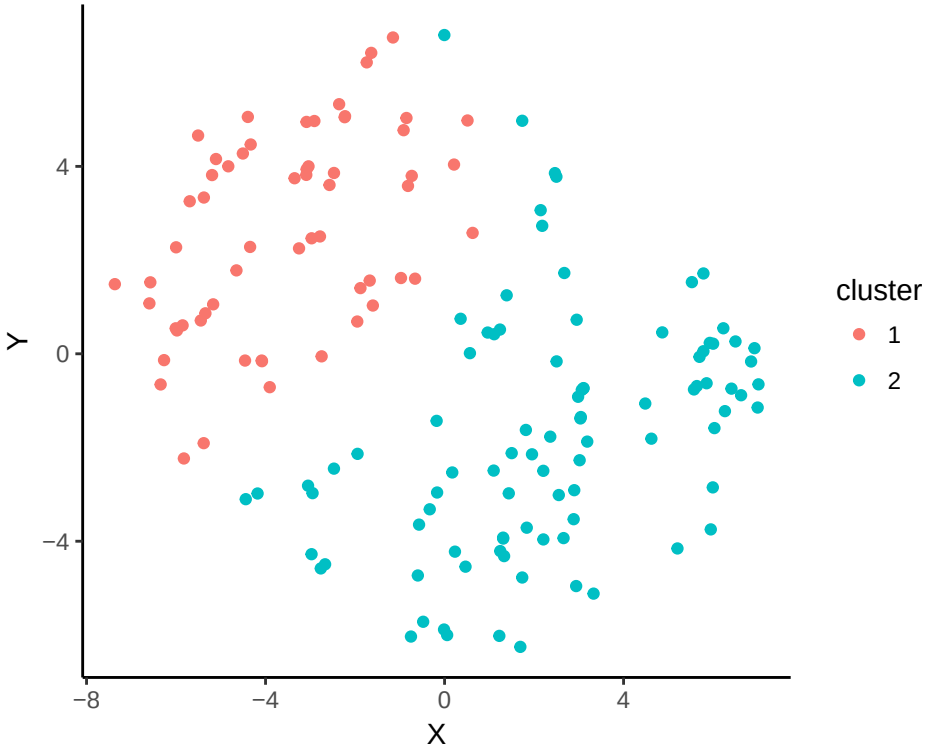

Figure B.7: 2D ordination plot of aggregate distances between cases, with cases coloured by assigned cluster

Cluster Analysis – Summary Statistics on Clusters

Table B.4: Frequencies of Degree Type, Study Mode and OLS Access, and mean and standard deviation of achievement in assessment questions, by Cluster

Variable	Labels	Cluster 1 n=60	Cluster 2 n=91	Total n=151
Degree Type	BusComLawIT	37	51	88
	Psychology	5	12	17
	Science	12	24	36
	Other	6	4	10
Study Mode	External	42	76	118
	On-campus Toowoomba	18	15	33
Regression Access	No Access	48	6	54
	At Least One Access	12	85	97
Binomial Access	No Access	51	8	59
	At Least One Access	9	83	92
Hypothesis Test Access	No Access	52	16	68
	At Least One Access	8	75	83
Assessment Questions		Mean (SD)	Mean (SD)	
	Reg Assignment	73.9% (19.6%)	82% (13.9%)	
	Bin Assignment	65% (32.5%)	70.5% (29.3%)	
	HT Assignment	66.7% (31.6%)	78.8% (23%)	
	Reg Exam	53.3% (33.5%)	67% (27.5%)	
	Bin Exam	56% (32.6%)	61.3% (32.5%)	
	HT Exam	59.3% (37.6%)	66.4% (33.6%)	

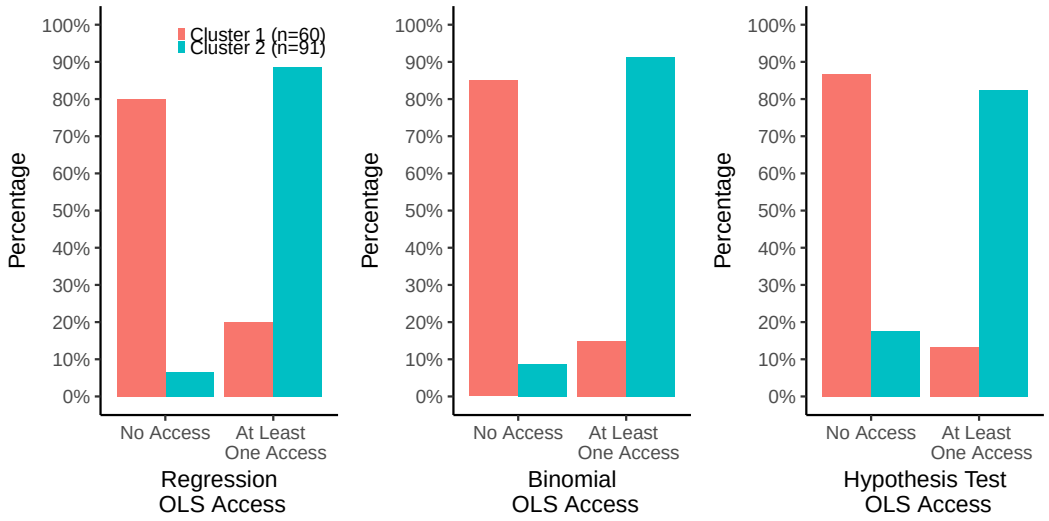

Figure B.8: OLS access by Cluster membership

Principal Components Analysis

Table B.5: Percentage variation explained by each principal component

	PC1	PC2	PC3	PC4	PC5	PC6
% Variation Explained	0.58	0.13	0.09	0.08	0.07	0.05
Cumulative % Variation Explained	0.58	0.71	0.80	0.88	0.95	1.00

Table B.6: Loadings of the original variables on each principal component

	PC1	PC2	PC3	PC4	PC5	PC6
Regression Assignment	0.39	-0.40	-0.62	0.12	-0.53	0.05
Binomial Assignment	0.40	-0.56	0.26	-0.09	0.44	0.50
Hypothesis Test Assignment	0.45	-0.18	0.33	0.31	0.08	-0.74
Regression Exam	0.42	0.26	-0.25	-0.78	0.20	-0.23
Binomial Exam	0.41	0.36	0.53	-0.07	-0.58	0.30
Hypothesis Test Exam	0.39	0.54	-0.31	0.52	0.38	0.22

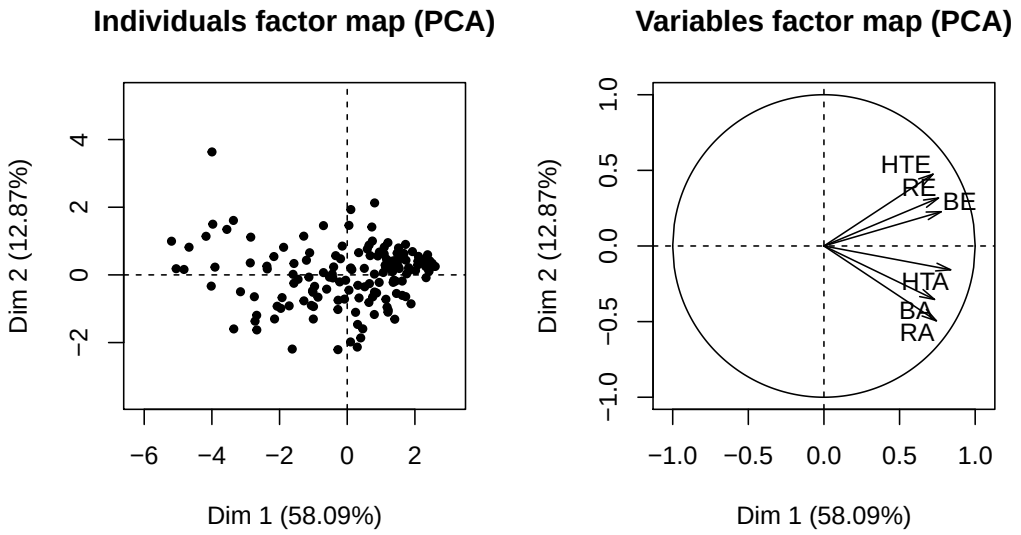

Figure B.9: Individuals (left) and Variables (right) factor maps displaying both the cases and the variable vectors against the principal components. Vectors represent the three topics (R, B and HT) with labels ending in either A (Assignments) or E (Exams)

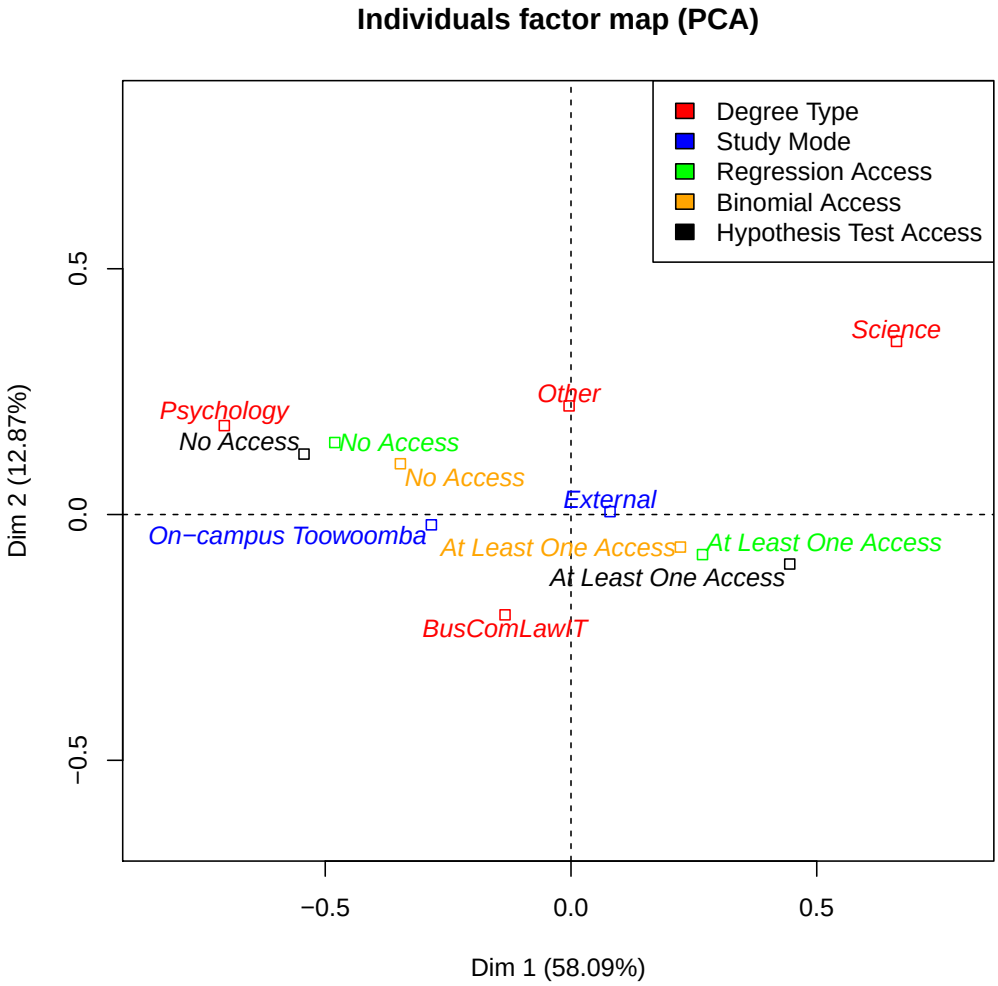

Figure B.10: Coloured PCA Plot with Degree Type included as a qualitative supplementary variable

Appendix C

Semester 2 Results

C.1 Assessment Achievement

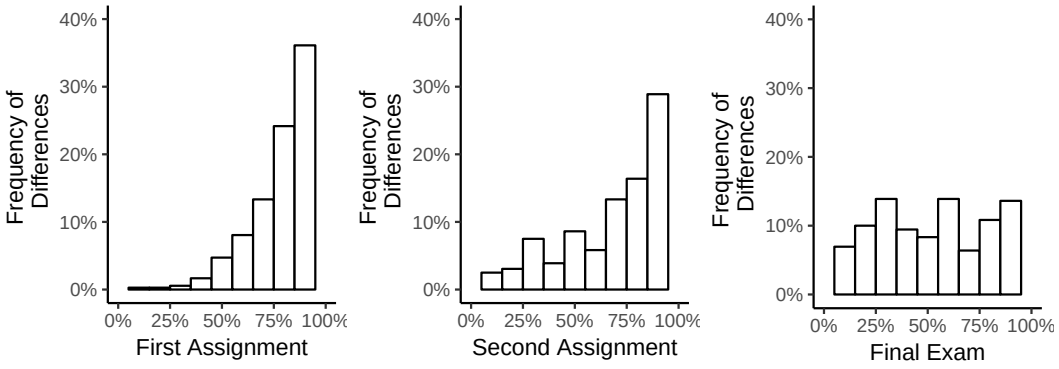

Figure C.1: Distribution of overall achievement for each assessment item

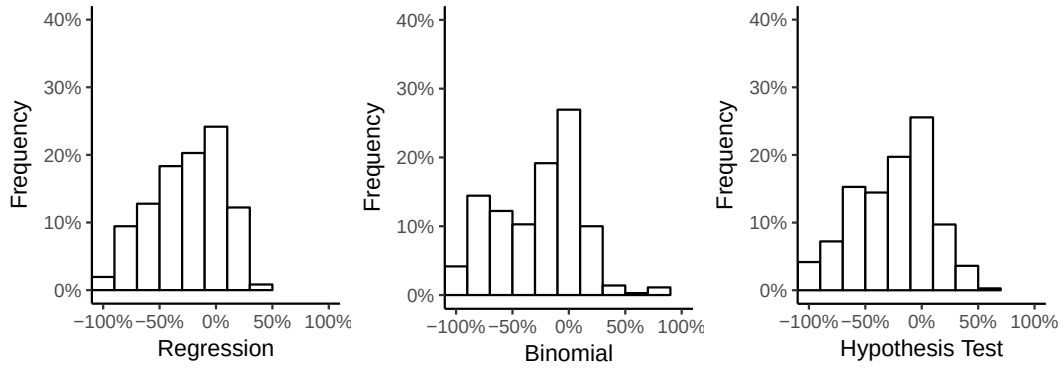

Figure C.2: Histograms of differences between assignment and exam achievement for each chosen topic (*Difference = Exam – Assignment*)

Table C.1: p-values for testing of differences between assignment and exam scores in the three topics

Assessment Topics	Parametric	Nonparametric
Regression	5.72E-40	3.44E-33
Binomial	2.21E-33	5.54E-28
Hypothesis Test	1.43E-34	1.66E-29

C.2 Relationships between data sources

Table C.2: Distribution of Degree Type by Study Mode

Degree Type	External	On-campus Springfield	On-campus Toowoomba	Total
BusComLawIT	110	36	82	228 (63%)
Psychology	20	18	11	49 (14%)
Science	37	0	29	66 (18%)
Other	7	5	5	17 (5%)
Total	174 (48%)	59 (16%)	127 (35%)	360

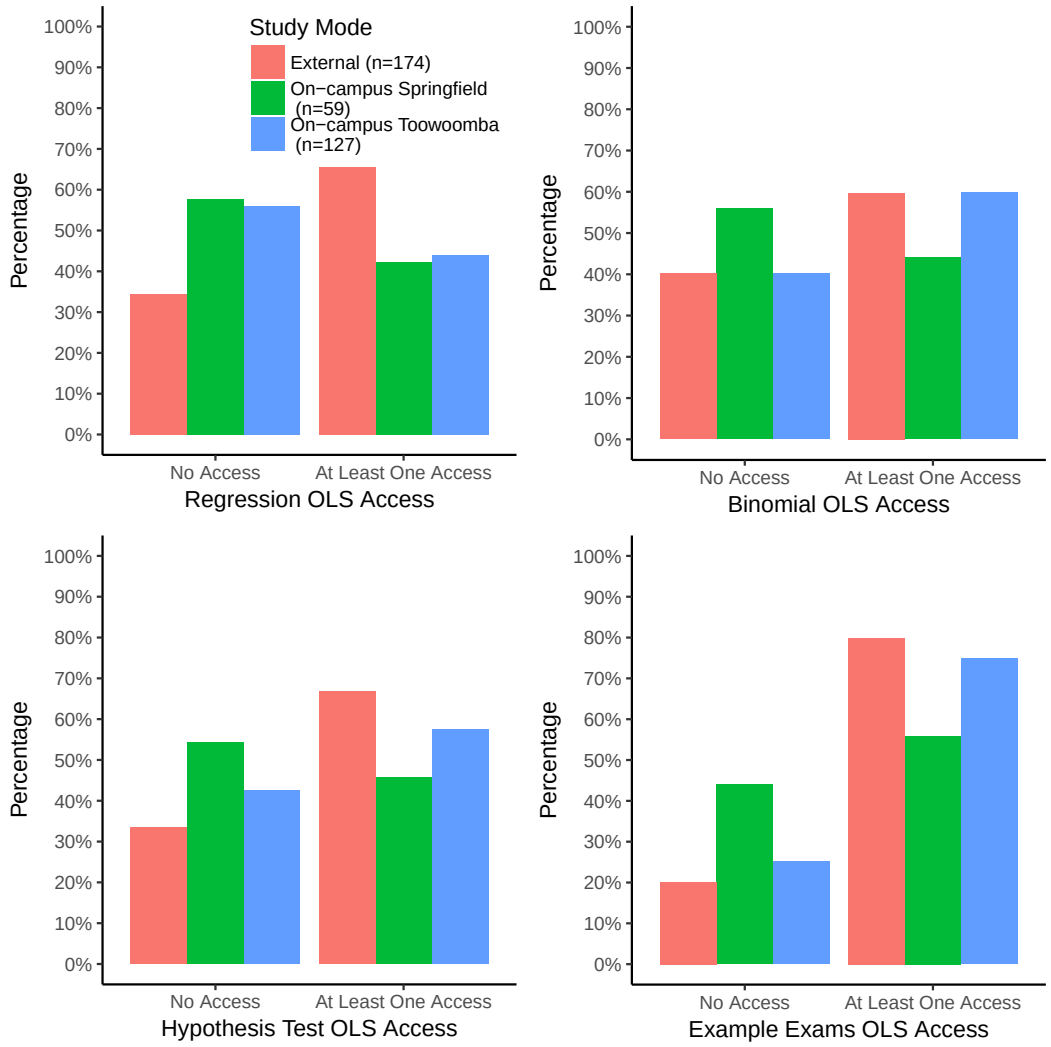

Figure C.3: Distribution of OLS access by study mode; OLS objects include tutorial solutions for the three chosen topics and the example exams

Table C.3: OLS access (percentage) of cohort for each degree type

OLS Objects	BusComLawIT (n=228)	Psychology (n=49)	Science (n=66)	Other (n=17)
Regression	53	63	53	53
Binomial	57	57	64	41
Hypothesis Test	59	57	65	59
Example Exams	74	71	77	71

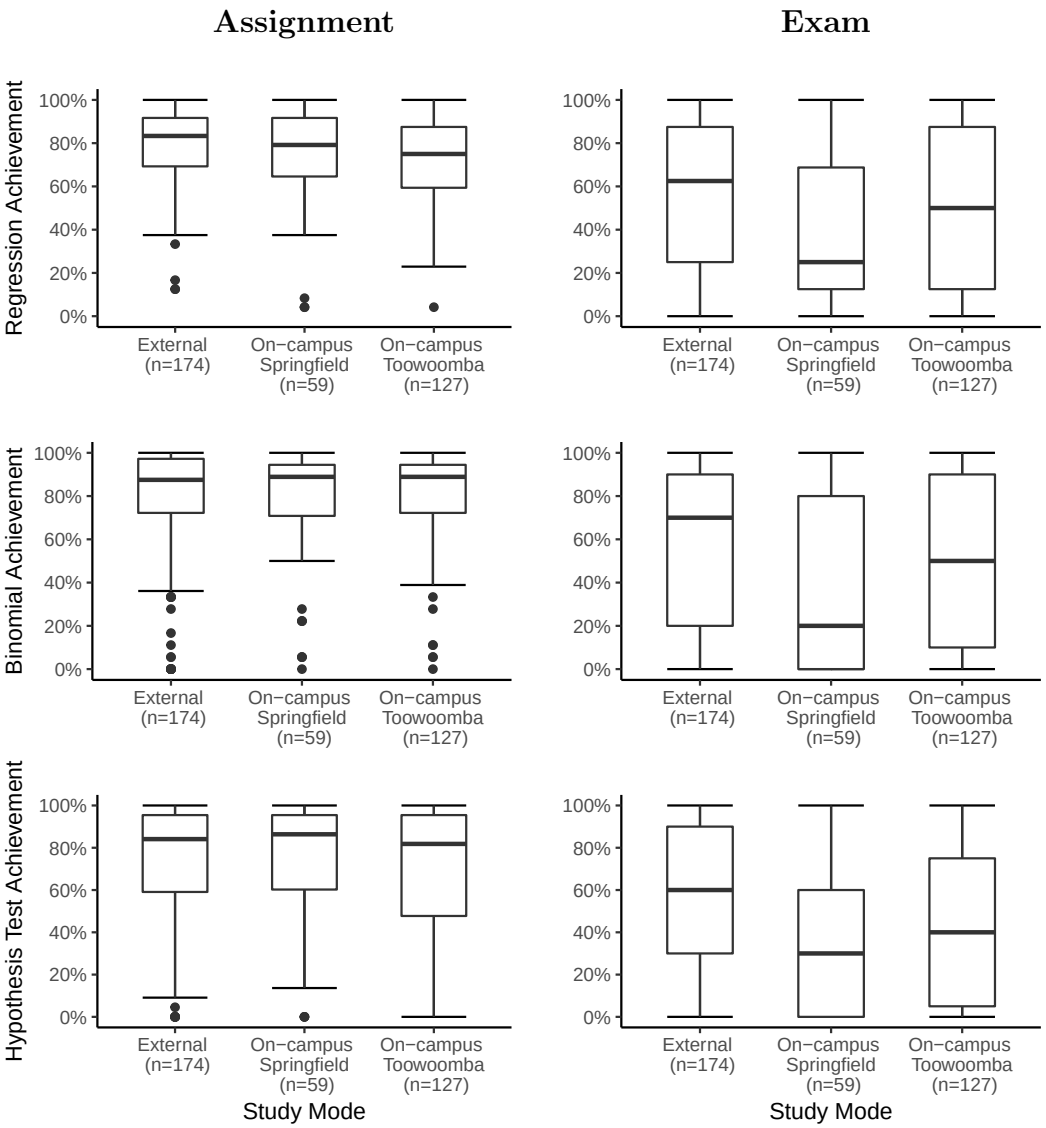

Figure C.4: Distribution of achievement on each topic for both assignment and exam questions by study mode; assignment questions on the left, exam questions on the right. Note that the example exams OLS access logs are not considered here.

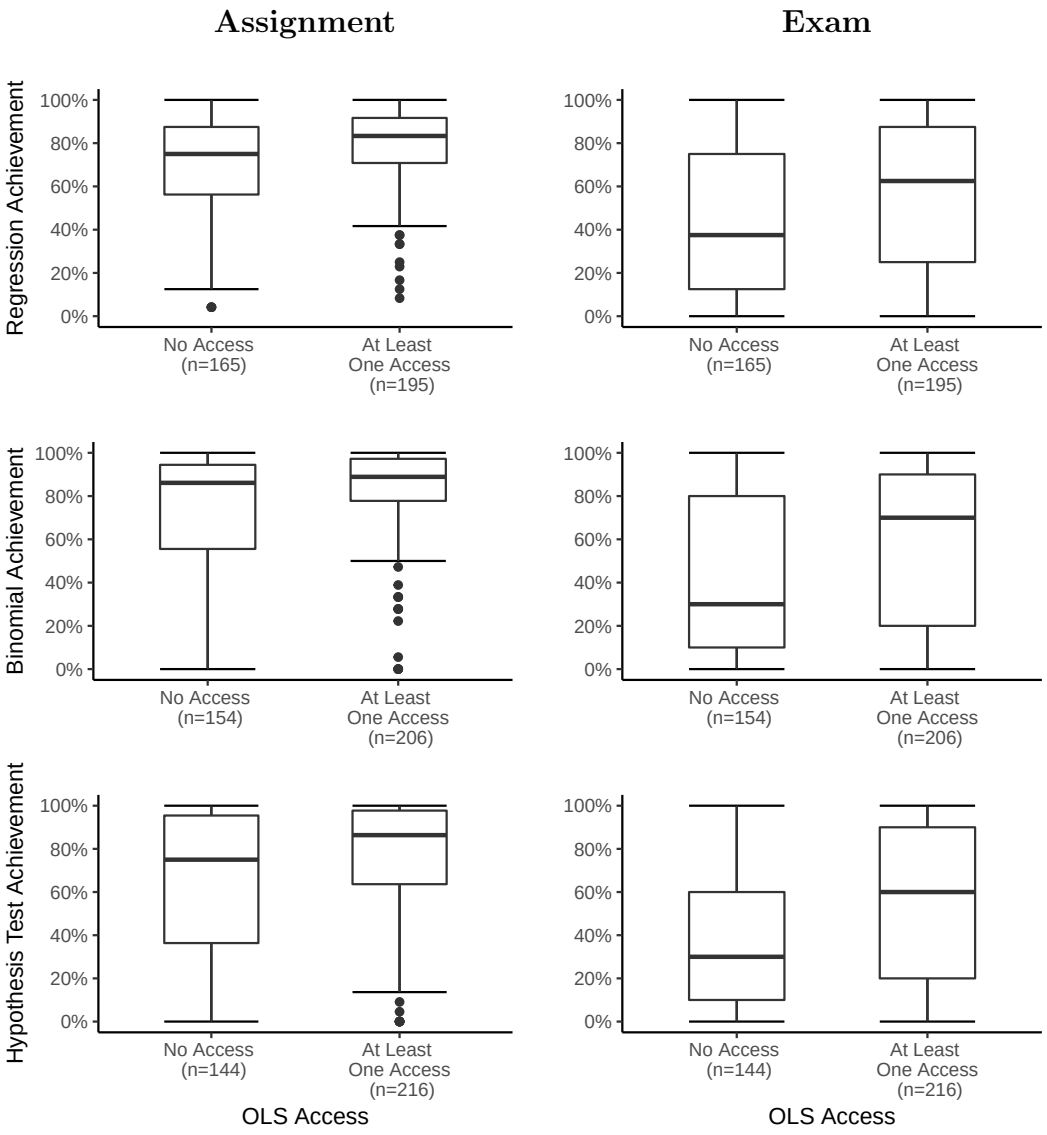

Figure C.5: Distribution of achievement on each topic for both assignment and exam questions by the frequency of access to tutorial solutions; assignment questions on the left, exam questions on the right. Note that the example exams OLS access logs are not considered here.

C.3 Multivariate Analyses

Cluster Analysis

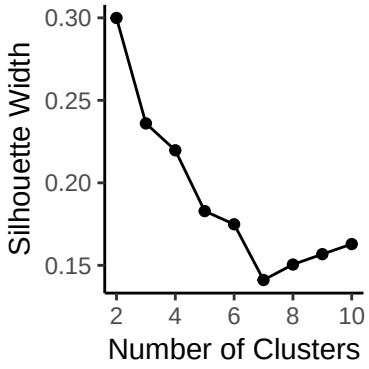

Figure C.6: Silhouette Width for different numbers of clusters

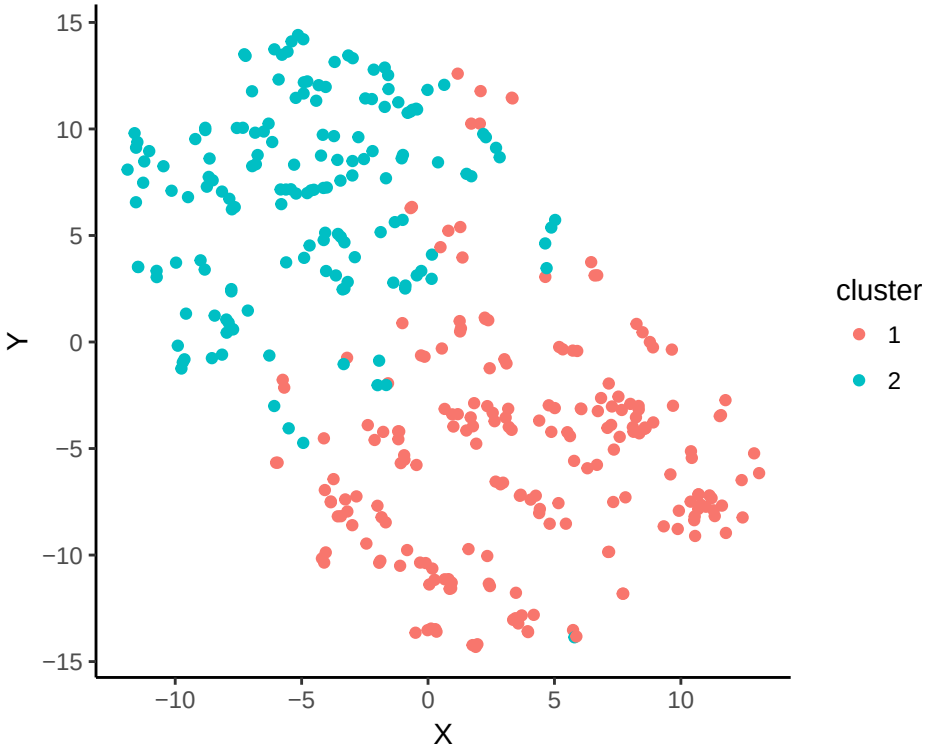

Figure C.7: 2D ordination plot of aggregate distances between cases, with cases coloured by assigned cluster

Cluster Analysis – Summary Statistics on Clusters

Table C.4: Frequencies of Degree Type, Study Mode and OLS Access, and mean and standard deviation of achievement in assessment questions, by Cluster

Variable	Labels	Cluster 1 n=203	Cluster 2 n=157	Total n=360
Degree Type	BusComLawIT	129	99	228
	Psychology	26	23	49
	Science	41	25	66
	Other	7	10	17
Study Mode	External	127	47	174
	On-campus Springfield	21	38	59
	On-campus Toowoomba	55	72	127
Regression Access	No Access	27	138	165
	At Least One Access	176	19	195
Binomial Access	No Access	29	125	154
	At Least One Access	174	32	206
Hypothesis Test Access	No Access	25	119	144
	At Least One Access	178	38	216
Assignment Questions		Mean (SD)	Mean (SD)	
	Reg Assignment	80% (18.3%)	69.7% (20.2%)	
	Bin Assignment	83.1% (22.5%)	72% (29%)	
	HT Assignment	79% (25.9%)	61.8% (32.6%)	
	Reg Exam	60.5% (35.4%)	37.3% (34.3%)	
	Bin Exam	64.5% (35.3%)	35.3% (35.2%)	
	HT Exam	57% (35.4%)	33.9% (31.2%)	

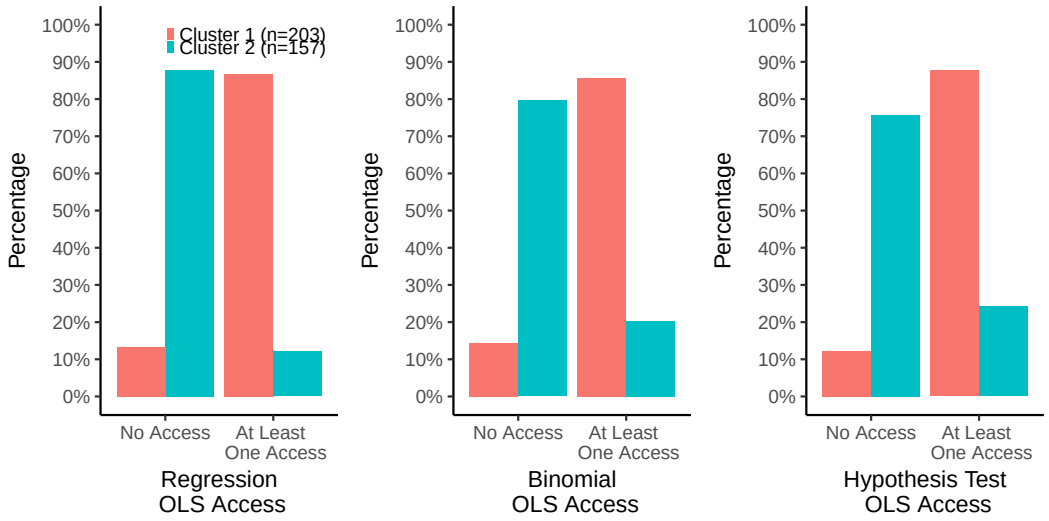

Figure C.8: OLS access by cluster membership

Principal Components Analysis

Table C.5: Percentage variation explained by each principal component

	PC1	PC2	PC3	PC4	PC5	PC6
% Variation Explained	0.57	0.16	0.09	0.07	0.06	0.05
Cumulative % Variation Explained	0.57	0.73	0.82	0.89	0.95	1.00

Table C.6: Loadings of the original variables on each principal component

	PC1	PC2	PC3	PC4	PC5	PC6
Regression Assignment	0.36	0.58	-0.46	0.34	0.36	0.27
Binomial Assignment	0.42	0.41	-0.12	-0.56	-0.38	-0.43
Hypothesis Test Assignment	0.39	0.25	0.83	0.27	-0.09	0.11
Regression Exam	0.41	-0.40	-0.23	0.58	-0.29	-0.43
Binomial Exam	0.43	-0.38	-0.14	-0.26	-0.30	0.70
Hypothesis Test Exam	0.42	-0.35	0.11	-0.29	0.74	-0.22

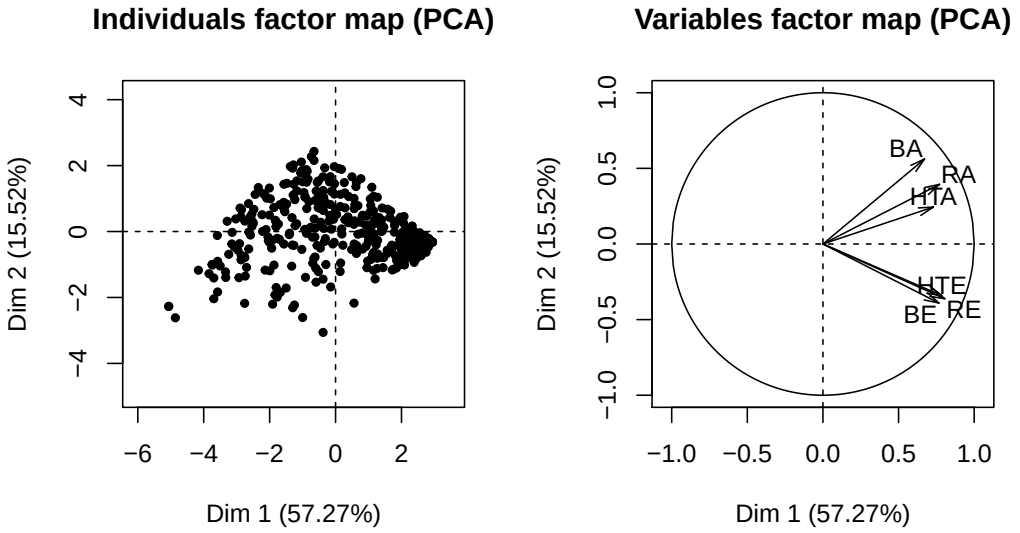

Figure C.9: Individuals (left) and Variables (right) factor maps displaying both the cases and the variable vectors against the principal components. Vectors represent the three topics (R, B and HT) with labels ending in either A (Assignments) or E (Exams)

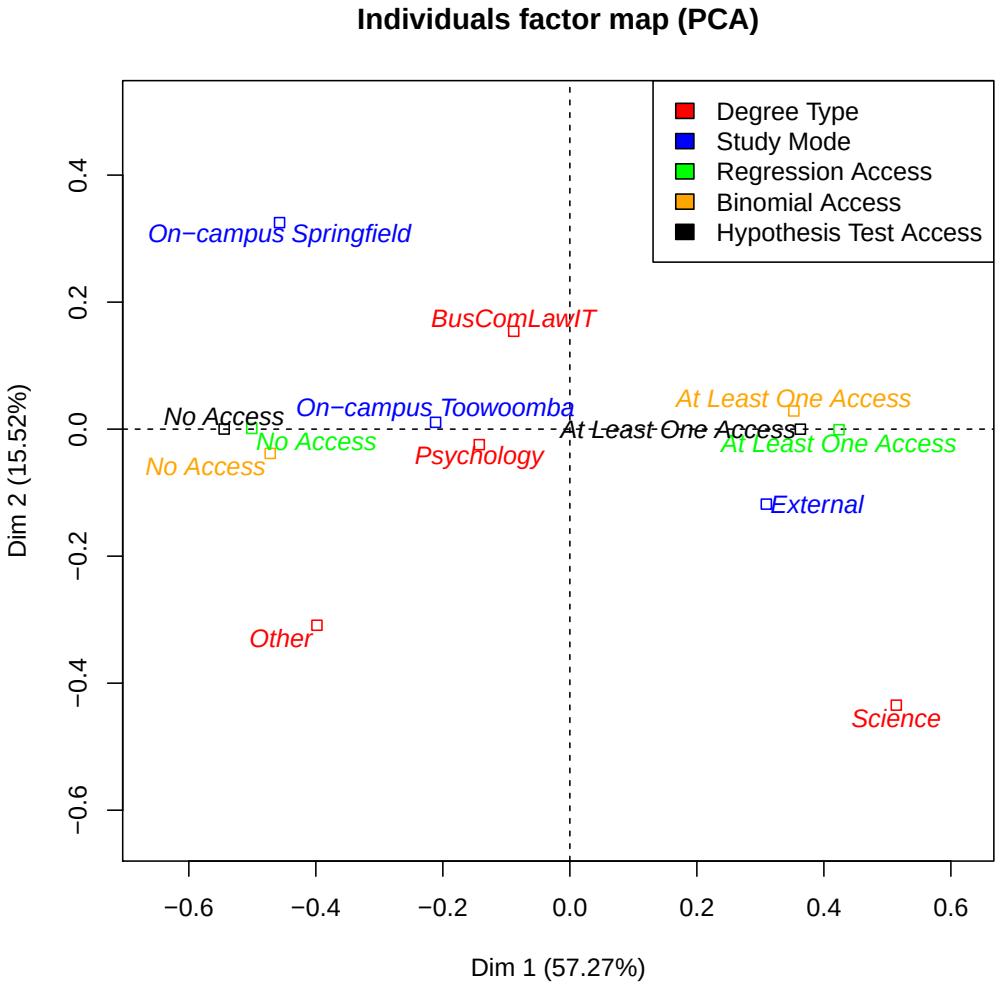

Figure C.10: Coloured PCA Plot with Degree Type included as a qualitative supplementary variable