Measures of engagement in the first three weeks of Higher Education predict subsequent activity and attainment in first year undergraduate students: a UK case study

Robert J. Summers[a]*, Helen E. Higson[b] and Elisabeth Moores[a]

[a]College of Health and Life Sciences, Aston University, Birmingham, UK; [b]Aston Business School, Aston University, Birmingham, UK

**Corresponding author:* R.J.Summers@aston.ac.uk

Running Title: Predicting activity and attainment in undergraduates

Version Date: 27 September 2020

Published in: Assessment and Evaluation in Higher Education,

DOI: 10.1080/02602938.2020.1822282

Abstract

Effective use of learning analytics systems has been purported to confer various benefits to learners in terms of both attainment and retention. There is, however, little agreement on which data are meaningful or useful. Whilst measures of engagement might correlate with outcomes, thereby retrospectively 'predicting' them, there are fewer studies which attempt to predict using 'live' system data in a face-to-face teaching environment. This study reports an analysis of week by week data from a learning analytics system which monitored 1,602 first year UK undergraduates. Uniquely, although students could view their own data, no formal interventions took place. Results showed that students who obtained the highest end-of-year marks were more likely to be in a higher engagement quintile as early as the first 3-4 weeks, and that early engagement was highly predictive of future engagement. Students who started in a higher engagement quintile, but their engagement decreased, were more likely to have higher marks than those that started in a lower quintile and then increased their engagement. Early measures of engagement are therefore predictive of future behaviour and of future outcomes, a finding which has important implications for universities wishing to improve student outcomes.

Keywords: learning analytics, student engagement, higher education, attainment

Introduction

Learning Analytics (LA) is a relatively new field concerned with measuring, collecting, analysing and reporting data about students and their learning environment, with the goal of understanding and improving learning (Ferguson, 2012). The types of data utilised in learning analytics are varied, dependent on the system, and on decisions made in the design of the system (Gašević, Tsai, Dawson & Pardo, 2019). Information from the log files of popular Virtual Learning Environments (VLEs) that have been used in learning analytics systems include VLE Logins (e.g., Jayaprakash, Moody, Lauría, Regan & Baron, 2014), course material access (e.g. Lu, Huang, Huang, & Yang, 2017), online quiz attempts (e.g., Macfadyen & Dawson, 2010), discussion forum interactions (e.g., Beer, Jones & Clark, 2009) and the results of ongoing assessment (e.g., Arnold & Pistilli, 2012). Often, the total number of *hits* is the data-point of interest (e.g. number of logins, forum posts, page views, *etc*.), but some systems use the time spent on tasks (e.g., Santos, Verbert, Govaerts & Duval, 2013). In addition to students' interactions with the VLE, there has been some work incorporating data generated by the student, such as self-report data about their own learning (Ellis, Han & Pardo, 2017) or emotional well-being (Villano, Harrison, Lynch & Chen, 2018). This *digital footprint*, both individually and collectively, contains potentially valuable information about learners, learning, courses, and the university itself (Gašević, Dawson & Siemens, 2015). The footprint is often combined with demographic data such as age, sex and ethnicity, as well as prior academic history (such as grade-point average) and background (Arnold & Pistilli, 2012; Jayaprakash *et al.*, 2014). A recent report (Sclater, Peasgood & Mullan, 2016) identified four potential key benefits of using learning analytics for UK Higher Education Institutions. They suggest that, used effectively, learning analytics can help to (i) ensure and improve quality of teaching, (ii) improve student retention, (iii) narrow attainment gaps between sub-populations of students, and (iv) enable the delivery of personalised learning.

Unfortunately, there is little agreement on which data are meaningful or even useful in learning analytics systems (Agudo-Peregrina, Iglesias-Pradas, Conde-González & Hernández-García, 2014), but some reasonably consistent patterns have been identified in the literature. The most basic level of VLE activity is the number of interactions with it (such as logins, or content pages accessed). A large amount of research has found correlations between this basic level of VLE activity and academic performance (Macfadyen & Dawson, 2010; Mogus, Djurdjevic & Suvak, 2012; You, 2016; Waheed *et al.* 2020), and that such activity can account for between 8% and 36% of the variance in end-of-year mark in online courses (Morris, Finnegan & Wu, 2005; Ramos & Yudko, 2008; Macfadyen & Dawson, 2010; Agudo-Peregrina *et al*., 2014), but not in a more traditional setting (Agudo-Peregrina *et al.*, 2014). Boulton *et al*. (2018) found a relationship between the amount of time spent on the VLE and course performance in a traditional 'bricks and mortar' setting, but the effect was course-dependent and revealed substantial variation in behaviour (e.g. students with low VLE activity obtaining high marks, and vice versa).

Arguably, the accessing and viewing of pages on the VLE can be classified as a passive activity (Agudo-Peregrina *et al*., 2014) and so there is increasing interest in VLE data that reflects more active areas of learning, such as participation in online discussion forums or online quizzes, and its relationship to academic performance. Macfadyen and Dawson (2010) found that the number of discussion messages posted accounted for 27% of the variance of the course mark, whereas reading posts did not correlate significantly with mark. Morris *et al.* (2005), however, found that both the number of discussion posts viewed and the time spent viewing them accounted for 18% and 14% of the variability in course marks respectively, yet Ramos and Yudko (2008) found no such relationship. These contradictory results may be explained by Beer, Jones and Clark (2009), who found a positive relationship between course grade and the

number of discussion posts created, but only when course instructors were more active in the discussion forums.

The link between lecture attendance and academic performance is now well established (Newman-Ford, Fitzgibbon, Lloyd, & Thomas, 2008), and has been found to account for around 14% of the variation in course marks, although the causal relationship remains the subject of some debate (Credé, Roch, & Kieszczynka, 2010). The introduction of lecture capture further complicates the relationship between attendance and course success as it has been argued that it influences attendance (see e.g. Moores, Birdi & Higson, 2019 for a review). There is evidence that students make more use of pre-recorded lectures during assessment and revision periods (e.g., Brady, Wong & Newton, 2013), but the relationship between academic outcomes and lecture capture is equivocal. While some studies find evidence for a positive relationship between the introduction of lecture capture and academic outcomes (Bollmeier, Wenger & Forinash, 2010; Wiese & Newton, 2013) the impact may be moderated by an overall reduction in lecture attendance (Mallinson & Baumann, 2015; Edwards & Clinton, 2019).

A number of studies have revealed relationships between library use and attainment, though the correlations are generally quite low. For example, checkouts of physical resources (Allison, 2015; Renaud *et al.*, 2015) accounted for only 1% of the variance in student marks. Even when the measure of library use incorporated access to electronic resources, academic skills instruction and use of computer workstations (Soria, Fransen & Nackerud, 2013; Thorpe, Lukes, Bever & He, 2016) the results were little better. Thorpe *et al.*'s study accounted for around 10% of the variance in students' grades, albeit with a very small sample (n=57), at a university that caters solely to distance learning students. Indeed, a meta-analysis by Robertshaw and Asher (2019) concluded that there was little evidence to demonstrate the

positive effect of libraries on university academic achievement, but acknowledged that the correlation that exists may be driven by the tendency of high-achieving students to make more use of the library.

Although many learning analytics systems tend to process the digital footprint in near real-time, most of the work that looks at the pattern of VLE activity over the course of the academic year deals with techniques to process the data (Młynarska, Greene & Cunningham, 2016; Hassan *et al.*, 2019; Peach, Yaliraki, Lefevre & Barahona, 2019; Waheed *et al.*, 2020) rather than reporting what the patterns are. Macfadyen and Dawson (2010; p593) found a consistent difference between high- and low-achieving students in their time spent online as term progressed but there was no systematic relationship with term progression; the amount of time spent online varied substantially from week to week. However, this and other similar data (Kuzilek, Hlosta & Zdrahal, 2017) are from online courses and may not reflect behaviour in a face-to-face teaching environment. Here, we present an analysis of data from a cohort of 1,602 first-year undergraduates from a 'bricks and mortar' university in the UK provided by an extant learning analytics system in order to (i) examine the relationship between academic performance and the digital footprint over the course of the academic year, (ii) identify engagement activity that provides early predictive power of academic performance, and (iii) identify correlates of academic performance at early vs. overall stages. The data are particularly interesting, because whilst students in this cohort could see their own learning analytics data, the University were in a trial year and so no policies or interventions were yet in place to intervene if a student's data suggested that intervention might be necessary. Thus, with the exception of any potential effects of merely measuring, recording and displaying the data, we can observe something approximating behaviour without interventions.

Materials and Methods

Sample data/participants

Aston University is a research-active, medium-sized UK university with an ethnically diverse population relative to other UK institutions. Approximately 51% of the sample read STEM (Science, Technology, Engineering, Maths) subjects, 42% were in the Business School, and the remainder (7%) read Languages and Social Sciences (usually in combination with a subject from one of the other Schools).

Undergraduate records – demographics and end-of-year performance – were obtained from the University's electronic records systems for full-time, first-year, home undergraduate students who first enrolled for the 2018/9 academic year and were still listed as current at the beginning of the second semester of the 2019/20 academic year. The initial sample comprised 1,823 student records. After removing students on courses with fewer than ten students (so normalization of VLE activity by course could take place, see below) the sample reduced to 1,602 students. Demographic data comprised sex (male/female), ethnicity (declared by the students themselves using the 18 categories used for United Kingdom census data, but later grouped into the superordinate categories of 'Asian or Asian British,' 'Black/African/Caribbean/Black British', 'White', 'Other, including Mixed/Multiple ethnic groups' and 'Unknown', socio-economic class (Socio-economic classification (NS-SEC) analytic classes, 1 = higher managerial, administrative and professional occupations and 8 = never worked and long-term unemployed, 9 = not classified), and POLAR4 Quintile (measure of participation of young people in UK Higher Education by geographic region, 1 = areas of lowest participation, 5 = areas of highest participation). The end-of-year performance (%) is the mean performance over the academic year. Of the 1,602 students 846 (53%) were female, 923 (58%) were Asian, 232 (14%) were black and 327 (20%) were white.

Measures

All undergraduate modules at Aston have a presence on the University VLE, where University announcements, timetables and course materials can be accessed. Since 2018, attendance at lectures and seminars has been electronically recorded by students swiping their identity card; though neither attendance nor the act of recording attendance is compulsory for home students. Additionally, all lectures are recorded and available through the VLE via a lecture capture system (LCS). Aston's learning analytics system, provided by *Solutionpath*, aggregates the log data from the VLE, attendance recording system, and lecture recordings on a daily basis. Six data feeds comprised the digital footprint: (i) ***VLE logins*****:** number of logins the student made to the VLE system, (ii) ***VLE course accesses*****:** number of times the student accessed course materials**,** (iii) ***Attendance:*** number of lectures that the student attended (the number of lectures students were expected to attend is also included which allows the percentage attendance to be computed) (iv) ***LCS*****:** number of times the student viewed recorded lectures, (v) ***VLE assessment accesses*****:** number of times the student attempted online quizzes and (vi) ***Library*****:** number of printed materials checked out of the library by the student. Unfortunately there were some system outages during the year resulting in incomplete data for week 5 of semester 1 for the attendance system, and weeks 7, 8 and 11 of semester 1 and week 6 of semester 2 for the LCS.

Analyses

For each student, the daily data for these six feeds was aggregated on a weekly basis for the 21 teaching weeks of the 2018/19 academic year (other weeks being for revision or assessment only). There was no university-wide strategy for learning interventions based on the data during the 2018/19 academic year.

The cohort was rank-ordered by end-of-year marks and divided into mark quintiles, see Table 1. The difference in mean marks between the top and bottom quintiles was 25.3 percentage points. 60% of students obtained a first year mark between 56% and 71%.

All statistical analysis was computed in R version 3.6.3 (R Core Team, 2020). Multiple linear regression was performed using *lm* in R. The relative importance of regressors was assessed using the *relaimpo* package (Grömping, 2006) for its implementation of the algorithm proposed by Lindeman, Merenda and Gold (1980) that evaluates all possible factor permutations to estimate each factor's variance contribution.

[INSERT TABLE 1 ABOUT HERE]

Results

Relationship between academic attainment and the digital footprint over the course of the academic year

For each mark quintile, the distribution of each component of the digital footprint for the academic year was computed and is plotted in Figure 1. For nearly all of the components, there is a positive relationship between mark quintile and the mean of the metric (shown with a black diamond); higher mark quintiles have a higher average engagement. There is, however, substantial overlap of the measures across mark quintile and the positive relationship is stronger for some metrics (e.g. VLE course access and attendance) than others (e.g. LCS). Notably, there is no systematic relationship between mark quintile and library use; however this may be due to the limitation that the feed does not track the use of electronic resources or use of books not removed from the library. It is also worth noting that the distribution of percentage attendance in mark quintile 1 is inverted in comparison to mark quintile 5, suggesting that there are more students with a lower percentage of attendance in mark quintile 1.

[INSERT FIGURE 1 ABOUT HERE]

Figure 2 shows the time-series of the data feeds stratified by mark quintile for the teaching weeks of each semester of the 2018/9 academic year. For four of the data feeds – logins, course accesses, attendance and LCS – it is clear from the first week in the academic year that students in the highest mark quintile are already engaging more with their studies and the VLE. From week 1 - and remarkably consistently throughout the academic year - students from the highest mark quintile attend lectures, login to the VLE, access online course materials, and review recorded lectures at greater levels than those students from the lower mark quintiles. The same pattern is mostly replicated in assessment accesses, though there is a lot more variation as a function of mark quintile in the first semester. Similar to Figure 1f, there is no systematic relationship between mark quintile and the library data feed (Figures 2k and 2l), although for the first six weeks of the academic year mark quintile 1 is a clear outlier.

[INSERT FIGURE 2 ABOUT HERE]

The patterns for logins and course access (Figure 2a-d) show an initial rise and then remain relatively stable for the rest of the academic year. In comparison with students from mark quintile 1, students from mark quintile 5 access the VLE 1.8 times as often and access course materials 2.1 times more. Unsurprisingly, the correlation between logins and course accesses is high ($r=0.57$). The drop in course accesses over weeks 7-10 of semester 2 that is most evident in mark quintile 5 is consistent with the idea that these high-performing students engage with the course materials early.

For the highest mark quintile, attendance (Figure 2e) remains fairly stable for the first 9 weeks of semester 1 before reducing by ~20 % points by week 11 (~2 lectures/week). For the lowest mark quintile, attendance declines after week 4 or 5 reducing by ~25 % points. In semester 2 (Figure 2f) the top four mark quintiles begin at broadly the same level of attendance as they did for semester 1 whereas the attendance of the lowest mark quintile is around 10 % points

(~1 lecture/week) lower than in semester 1. Unlike semester 1, attendance falls relatively steadily over the whole semester for all mark quintiles in semester 2 (c.f. Newman-Ford *et al.*, 2008).

Despite the system issues (see above) the time-series for LCS views (Figure 2g and 2h) clearly illustrates that students from the two highest mark quintiles engage far more with recorded lectures than those from the other mark quintiles (~1.7 times as many views), particularly early in semester 1. Note that if percentage attendance is computed from the number of lectures attended and the number of LCS views then mean percentage attendance for some weeks can be as high as 90% for students from the highest mark quintile; though the data does not tell us whether students are reviewing previously attended lectures or catching up on missed lectures. Whilst it cannot be ruled out that lecture capture has depressed attendance, this is supporting evidence for a positive relationship between use of lecture capture and academic outcome (Bollmeier *et al.*, 2010; Wiese & Newton, 2013).

Notwithstanding the more variable patterns for assessment access (Figure 2i and 2j), it is clear that students in mark quintiles 4 and 5 are outliers compared with those from the remaining quintiles, particularly in semester 2.

Finally, there is no systematic relationship between the checking out of printed materials from the library (Figure 2k and 2l) and mark quintile across the academic year. With the exception of the first six weeks of the first semester, where students from mark quintile 1 make less use of the library than those of other mark quintiles, there is substantial overlap and change in rank ordering between the mark quintiles over the course of the academic year. Notably, students from mark quintile 5 check out fewer printed materials in the last three weeks of semester 2

than the other mark quintiles. As this feed is limited to the use of printed resources, and therefore does not cover access to electronic textbooks or journal articles, the usefulness of this data may be limited though it is consistent with the overall lack of relationship between library usage and marks (Robertshaw & Asher, 2019).

In summary, it is evident as early as the first 3-4 weeks of the academic year that those students who obtain the highest end-of-year marks are more likely to attend lectures and interact more with the VLE.

Engagement activity as an early indicator of academic performance

There is a clear and substantial difference in the degree to which students from the highest and lowest mark quintiles access online course materials early on in the academic year, which makes course-access suitable for use as a measure of engagement. Therefore, using the course-access data from weeks 1 to 3 students were divided into activity quintiles in the following manner. For each student, the total number of accesses to course materials, C_{TOT}, was computed. In order to account for differences in the amount of course materials provided on the VLE between courses, the z-score for C_{TOT} was computed for each student grouped within each course. Students were then ranked in order of the z-score for C_{TOT} and divided into activity quintiles.

[INSERT TABLE 2 ABOUT HERE]

Mean, minimum and maximum marks for each of the activity quintiles are shown in Table 2. There is a systematic, albeit small, relationship between the mean mark and each activity quintile. Those students in the highest activity quintile obtain end of year marks that are, on average, 7.2% points higher than those from the lowest activity quintile. Given that the largest difference in marks between the highest and lowest mark quintile is 25.3% points (see Table

1), the activity quintiles computed here, from just the first 3 weeks of a single data feed, can account for 28% of the difference in marks between the top and bottom mark quintiles. Note that (i) activity quintiles do not take into account any other variables, such as prior attainment, subjects studied or other student demographic variables and (ii) that no student in the top three activity quintiles fails their first-year course by scoring less than 40%. The relationship between activity quintile and mark quintile is illustrated with a Sankey diagram indicating the proportions of students from each activity quintile that end up in each mark quintile (Figure 3). Only 7% of students from activity quintile 1 achieve a mark that puts them in the top mark quintile (end of year mark >71%) whereas 31% of students in activity quintile 5 end up in mark quintile 5. Furthermore, 75% of students in activity quintile 5 obtain a mark that puts them in the top three mark quintiles (i.e., mark quintiles 3, 4 or 5; mark>61%) compared with 46% of students in activity quintile 1. Approximately 9% of students in activity quintile 5 finish with a mark in mark quintile 1, indicating that some of the weakest students are working hard to improve their performance (Sclater *et al.*, 2016), albeit unsuccessfully.

[INSERT FIGURE 3 ABOUT HERE]

The time-series of each of the six data feeds stratified by activity quintile is shown in Figure 4. It is noteworthy that whilst the quintiles were based on just the first three weeks of activity, with few exceptions the ranking remains consistent across the entire academic year. As before, caution should be used in the assessment of the data feeds for logins and course accesses (Figure 4a-d) as these measures are highly correlated with each other (r=0.57) and the method for computing activity quintile necessarily results in the large degree of separation between the first and fifth activity quintile for these data feeds, at least over the first 3 weeks. Nonetheless, the data feeds that do not show strong correlations with course access – attendance (r=0.22), assessment accesses (r=0.23), library use (r=0.18), and LCS (r=0.21) – also demonstrate a relationship with activity quintile.

[INSERT FIGURE 4 ABOUT HERE]

Students who begin the year in a given activity quintile may not end it in the same activity quintile. In order to determine the effect that changes in activity quintile throughout the year might have on marks we computed a late-stage activity quintile from the last three weeks of the data. The mean marks for each combination of activity quintile and late-stage activity quintile are shown in Table 3. Students who begin and end the academic year in the highest activity quintile achieve a mean grade of 69.3% vs. 61.0% for those students who begin in the lowest activity quintile but finish in the highest activity quintile.

[INSERT TABLE 3 ABOUT HERE]

Identification of correlates of academic performance at early vs. overall stages

The previous two sections have demonstrated the relationship between end-of-year marks and the digital footprint, and that early interaction with the online system is predictive of other aspects of the digital footprint throughout the academic year. In this section we use multiple linear regression to investigate the extent to which students' end-of-year mark can be accounted for by aspects of their demographics and their digital footprint; due to the high correlation between logins and course material access, logins were excluded from the model. Additionally, to account for differences in the mean mark for each course, the course identifier was included as a factor (35 levels).

The model was computed on two sets of data, one from the first three weeks of the academic year, and one from the whole academic year. Model coefficients and the proportion of variance explained by the whole model and individual factors are in Table 4.

[INSERT TABLE 4 ABOUT HERE]

The multiple linear relationships between the data and end-of-year marks can account for 29.7% and 38.9% of the early and complete data respectively. Although most of the factors contribute significantly to the overall model, only a small number contribute substantial unique variance to the total variance explained by the model. In both models differences in the mean mark by course are the single largest contributor to the total variance – 11.9% and 10.4% for the early and complete data, respectively. For the early data, course accesses (5.4%), ethnicity (4.0%) and attendance (3.7%) are the next largest contributors to the overall variance with the remaining factors contributing 1.4% or less; sex, ethnicity and library use were not significant. For the complete data, attendance (9.8%) and course accesses (8.0%) contributed most (behind course mark differences); the remaining factors contributed no more than 3.4% of the variance each. Once the difference in mean end-of-year mark by course is accounted for, the combination of demographic data and the digital footprint accounts for a small but significant proportion of the variance in end-of-year marks.

Discussion

In common with previous research (e.g., Newman-Ford *et al.*, 2008; Mogus *et al.*, 2012; Boulton *et al.*, 2018) students who obtain the highest marks attend more lectures and use the VLE more over the course of the academic year. The novel finding of this paper demonstrates that this pattern of behaviour begins early in the academic year and tends to continue throughout it. These data contrast with the limited available data on the time-series of VLE interactions; e.g. for online courses VLE interactions begin high, but decline as the term progresses (Hassan *et al.*, 2019; p1940). One aspect of VLE activity related to the accessing of course materials early in the academic year was predictive of other aspects of the digital footprint (e.g. attendance, library use) throughout the academic year and positively related to end-of-year marks. Using this dataset, we are unable to be ascertain whether this pattern reflects that students with higher attainment prior to coming to University also have higher attendance,

or whether higher attendance facilitates higher attainment; this could be a question for future research.

A linear model of the relationship between student demographics and the digital footprint with end-of-year marks found that attendance and access to course materials accounted for 12.1% of the total variance with only 3 weeks of data, rising to 23.4% of the total variance for data from the full academic year. Although Agudo-Peregrina *et al.* (2014) found a significant relationship between VLE activity and end-of-year mark for online courses they did not do so for courses with face-to-face teaching components. Nonetheless, they did find that VLE activity related to the transmission of course content accounted for around 10% of the variance in end-of-year marks, similar to that reported here for the complete year's data. Their overall linear model for data from online teaching accounted for between 23.9% and 35.6% of the variance depending on how their data were partitioned; the lower end of this range is comparable with the amount of variance accounted for in this study solely by the digital footprint (23.4%). It is worth noting that Agudo-Peregrina *et al.* (2014) found that student-teacher interactions – such as course-related messages exchanged between students and teachers – accounted for the majority of the explained variance in their model; these data were not measured by our system.

The result here, where attendance is the most significant component of the digital footprint in explaining end-of-year marks for the complete data, but not early data, complements the result of Marbouti *et al.*, (2015). Marbouti *et al.* used a logistic regression model to predict students who would succeed in an end-of-semester exam from data at weeks 2, 4 and 9 of a single-semester course. Attendance data was not a significant component of the model until week 4, and at week 9 became, albeit slightly, the largest component of the model.

Undoubtedly, improving the data comprising the digital footprint would improve the prediction accuracy of end-of-year marks. For example, data for the library only covers checked-out printed material whereas ideally, access to electronic resources would be covered as well; though see Robertshaw and Asher (2019). The LCS data feed is limited in that we do not know whether students are catching up on missed lectures or supplementing the experience of lecture attendance to improve notes. Differences between purely online and blended teaching environments mean that instructor-student or student-student interaction is more readily measured, e.g. discussion forum posts (Beer *et al.*, 2009), than with classroom-based discussion groups. One aspect missing from our data are the results of ongoing assessment (e.g. regular assignments, coursework, end of semester marks) which, while the amount and type are likely to be course dependent, are significant predictors of overall course success (Jayaprakash *et al.*, 2014).

LA systems that are designed to predict 'at-risk' students early in the academic year usually use complex machine-learning models built from training data prior to the system's introduction (Arnold & Pistilli, 2012; Jayaprakash *et al.*, 2014; Milliron & Malcolm, 2014). As the outcomes for these students are known (academic performance, withdrawal status) the performance of these predictive models on the training data can be reported (Jayaprakash *et al.*, 2014; Milliron & Malcolm, 2014), but the usefulness of the system is perhaps better inferred from the proportion of students identified as at-risk in the live system. For example, Milliron and Malcolm (2014) reported a system that reduced the number of students failing or withdrawing by approximately 3% (210 students), but required interventions in half of the student population, despite a claimed accuracy of 87%. The risk of intervening too often is underscored by Jayaprakash *et al.* (2014) who found that the rate of withdrawal for students enrolled on courses in the learning analytics system was 23.3% compared with 13.5% in a control group.

As an alternative to complex machine learning models to identify 'at-risk' students Foster & Siddle (2019) describe a learning analytics system that issues 'no engagement' alerts to students who fail to register any activity in their digital footprint – e.g. VLE activity, library book checkouts, etc. – for a period of 14 days. Overall the system issued alerts to approximately 9% of students and was 43% more likely to issue alerts for students from disadvantaged backgrounds even though demographic data did not contribute to the system's alerting decision. Assuming the no-engagement alerts work in improving outcomes, Foster & Siddle (2019) have demonstrated that a system can be used to reduce disparities in attainment between different populations without using demographic data, which risks generating false positives. In our data too, measures of attendance and course access explained a greater proportion of variance in attainment than did demographic measures of disadvantage (POLAR4 and socio-economic class), although ethnicity did account for a greater proportion than attendance alone in the early weeks (4.0% vs. 3.7%). Overall, these results indicate that targeting interventions based on behaviour rather than demographics should be a successful strategy.

Merely making students aware that they are 'at-risk' of not succeeding is often sufficient to improve academic outcomes (Arnold & Pistilli, 2012; Jayaprakash *et al.*, 2014; Milliron & Malcolm, 2014). Unfortunately, the evidence can be difficult to interpret because insufficient detail is given about control and experimental groups. When one system, Course Signals (Arnold & Pistilli, 2012), was re-examined (Main & Griffith, 2019) the reported gains in academic outcome were much more modest and difficult to interpret due to significant differences in the demographics and prior attainment of those students enrolled in modules using Course Signals vs. those who were not. More complex interventions, such as peer mentoring, may confer no additional benefit to 'at-risk' students. For example, compared with a control group, where no interventions took place, Jayaprakash *et al.* (2014) found a similar improvement in grades for two groups of students regardless of whether any interventions that

took place were simply informative or also included access to further resources and mentoring. Unfortunately, no data were reported to indicate whether or not students in the latter group used the extra resources available.

Many of the interventions identified in a meta-analysis by Sønderlund *et al.* (2019, p2612) 'put the onus primarily on the student to change behaviour when prompted' and that 'optimal intervention effectiveness may be better achieved if both student and institution (including teachers) are expected to react to negative student forecasts'. Furthermore, the interventions are almost universally academic interventions, which given the context of UK higher education where many students struggle to balance aspects of study such as, for example lecture attendance, with work or caring responsibilities may require different kinds of support (see e.g. Moores, Birdi, & Higson, 2019 for a review).

The importance of early intervention on academic outcomes is crucial (Table 3) since even students who begin the year in the lowest activity quintile but, for the last 3 teaching weeks, are in the highest activity quintile achieve marks at least 3.4 % points lower than other students who end the year in the highest activity quintile. With this in mind, encouraging all students, at the beginning of the academic year, to engage with instructors to identify the skills they need to develop in order to become successful undergraduates and making the development of these skills a core part of the course itself may prove fruitful. Learning analytics systems can supplement this work by ensuring that appropriate data are collected to help test the success of these interventions, e.g. results of intermediate assessments, and identify struggling students through no-engagement alerts (Foster & Siddle, 2019).

Funding details

This work was supported by the Centre for Innovation in Learning and Education (CILE), a Catalyst OfS funded project. The joint Aston/Cranfield virtual centre aims to develop new knowledge in innovative education, business-engaged educational design and innovative delivery modes in undergraduate provision within UK Higher Education.

Disclosure statement

We have no conflicts of interest or financial interests relating to this work.

Acknowledgements

We gratefully acknowledge the assistance of Solutionpath Ltd and Tai Luong with the data output.

Data availability statement

Due to difficulties in properly anonymising the dataset we are unable to share the data associated with this article.

Notes on Contributors

Robert Summers is postdoctoral research fellow at Aston University. He has a background in psychology, with particular focus on visual and auditory perception.

Helen Higson is Professor of Higher Education Learning and Management at Aston University. She has undertaken extensive research in the areas of employability and work-based learning, intercultural competences and strategic issues in HE management.

Elisabeth Moores is Associate Pro-Vice Chancellor at Aston University and has research interests in dyslexia, widening participation and evaluation of education.

ORCID

Robert Summers ORCID: 0000-0003-4857-7354

Helen Higson ORCID: 0000-0003-3433-2823

Elisabeth Moores ORCID: 0000-0003-3997-0832

References

Agudo-Peregrina, Á. F., Iglesias-Pradas, S., Conde-González, M. Á., & Hernández-García, Á. (2014). Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in Human Behavior*, *31*(1), 542–550. https://doi.org/10.1016/j.chb.2013.05.031

Allison, D. (2015). Measuring the Academic Impact of Libraries. *Portal: Libraries and the Academy*, *15*(1), 29–40. https://doi.org/10.1353/pla.2015.0001

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12*, 267. https://doi.org/10.1145/2330601.2330666

Beer, C., Jones, D., & Clark, K. (2009). The indicators project identifying effective learning: Adoption, activity, grades and external factors. *ASCILITE 2009 - The Australasian Society for Computers in Learning in Tertiary Education*, 60–70.

Bollmeier, S. G., Wenger, P. J., & Forinash, A. B. (2010). Impact of online lecture-capture on student outcomes in a therapeutics course. *American Journal of Pharmaceutical Education*, *74*(7), 1–6. https://doi.org/10.5688/aj7407127

Boulton, C. A., Kent, C., & Williams, H. T. P. (2018). Virtual learning environment engagement and learning outcomes at a 'bricks-and-mortar' university. *Computers and Education*, *126*, 129–142. https://doi.org/10.1016/j.compedu.2018.06.031

Brady, M., Wong, R., & Newton, G. (2013). Characterization of Catch-Up Behavior: Accession of Lecture Capture Videos Following Student Absenteeism. *Education Sciences*, *3*(3), 344–358. https://doi.org/10.3390/educsci3030344

Credé, M., Roch, S. G., & Kieszczynka, U. M. (2010). Class Attendance in College. *Review of Educational Research*, *80*(2), 272–295. https://doi.org/10.3102/0034654310362998

Edwards, M. R., & Clinton, M. E. (2019). A study exploring the impact of lecture capture availability and lecture capture usage on student attendance and attainment. *Higher Education*, *77*(3), 403–421. https://doi.org/10.1007/s10734-018-0275-9

Ellis, R. A., Han, F., & Pardo, A. (2017). Improving learning analytics - Combining observational and self-report data on student learning. *Educational Technology and Society*, *20*(3), 158–169.

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5–6), 304–317. https://doi.org/10.1504/IJTEL.2012.051816

Foster, E., & Siddle, R. (2019). The effectiveness of learning analytics for identifying at-risk students in higher education. *Assessment & Evaluation in Higher Education*, 1–13. https://doi.org/10.1080/02602938.2019.1682118

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, *59*(1), 64–71. https://doi.org/10.1007/s11528-014-0822-x

Gašević, D., Tsai, Y.-S., Dawson, S., & Pardo, A. (2019). How do we start? An approach to learning analytics adoption in higher education. *The International Journal of Information and Learning Technology*, *36*(4), 342–353. https://doi.org/10.1108/IJILT-02-2019-0024

Grömping, U. (2006). Relative Importance for Linear Regression in R: The Package relaimpo. *Journal of Statistical Software*, *17*(1), 1–27.

Hassan, S., Waheed, H., Aljohani, N. R., Ali, M., Ventura, S., & Herrera, F. (2019). Virtual

learning environment to predict withdrawal by leveraging deep learning. *International Journal of Intelligent Systems*, *34*(8), 1935–1952. https://doi.org/10.1002/int.22129

Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics*, *1*(1), 6–47. https://doi.org/10.18608/jla.2014.11.3

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Data Descriptor: Open University Learning Analytics dataset. *Scientific Data*, *4*(1), 1–8. https://doi.org/10.1038/sdata.2017.171

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to Bivariate and Multivariate Analysis.* Glenview, IL: Scott Foresman.

Lu, O. H. T., Huang, J. C. H., Huang, A. Y. Q., & Yang, S. J. H. (2017). Applying learning analytics for improving students engagement and learning outcomes in an MOOCs enabled collaborative programming course. *Interactive Learning Environments*, *25*(2), 220–234. https://doi.org/10.1080/10494820.2016.1278391

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers and Education*, *54*(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

Main, J. B., & Griffith, A. L. (2019). From SIGNALS to success? The effects of an online advising system on course grades. *Education Economics*, *27*(6), 608–623. https://doi.org/10.1080/09645292.2019.1674248

Mallinson, D. J., & Baumann, Z. D. (2015, June 19). Lights, Camera, Learn: Understanding the Role of Lecture Capture in Undergraduate Education. *PS - Political Science and Politics*, Vol. 48, pp. 478–482. https://doi.org/10.1017/S1049096515000281

Marbouti, F., Diefes-Dux, H. A., & Strobel, J. (2015). Building course-specific regression-

based models to identify at-risk students. *Proceedings of the 122nd ASEE Annual Conference & Exposition*. Seattle, WA: The American Society for Engineering Educators.

Milliron, M. D., & Malcolm, L. (2014). Insight and Action Analytics: Three Case Studies to Consider. *Research & Practice in Assessment*, (9), 70–89.

Młynarska, E., Greene, D., & Cunningham, P. (2016). Time series analysis of VLE activity data. *Proceedings of the 9th International Conference on Educational Data Mining, EDM 2016*, 613–614.

Mogus, A. M., Djurdjevic, I., & Suvak, N. (2012). The impact of student activity in a virtual learning environment on their final mark. *Active Learning in Higher Education*, *13*(3), 177–189. https://doi.org/10.1177/1469787412452985

Moores, E., Birdi, G. K., & Higson, H. E. (2019). Determinants of university students' attendance. *Educational Research*, *61*(4), 371–387. https://doi.org/10.1080/00131881.2019.1660587

Morris, L. V., Finnegan, C., & Wu, S. S. (2005). Tracking student behavior, persistence, and achievement in online courses. *Internet and Higher Education*, *8*(3), 221–231. https://doi.org/10.1016/j.iheduc.2005.06.009

Newman-Ford, L., Fitzgibbon, K., Lloyd, S., & Thomas, S. (2008). A large-scale investigation into the relationship between attendance and attainment: a study using an innovative, electronic attendance monitoring system. *Studies in Higher Education*, *33*(6), 699–717. https://doi.org/10.1080/03075070802457066

Peach, R. L., Yaliraki, S. N., Lefevre, D., & Barahona, M. (2019). Data-driven unsupervised clustering of online learner behaviour. *Npj Science of Learning*, *4*(1), 14.

https://doi.org/10.1038/s41539-019-0054-0

R Core Team. (2020). A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, p. https://www.R-project.org. Retrieved from http://www.r-project.org

Ramos, C., & Yudko, E. (2008). "Hits" (not "Discussion Posts") predict student success in online courses: A double cross-validation study. *Computers and Education*, *50*(4), 1174–1182. https://doi.org/10.1016/j.compedu.2006.11.003

Renaud, J., Britton, S., Wang, D., Ogihara, M., Mader, C., Maristany, L., & Zysman, J. (2015). Mining library and university data to understand library use patterns. *The Electronic Library*, *33*(3), 355–372. https://doi.org/10.1108/EL-07-2013-0136

Robertshaw, M. B., & Asher, A. (2019). Unethical numbers? A meta-analysis of library learning analytics studies. *Library Trends*, *68*(1), 76–101. https://doi.org/10.1353/lib.2019.0031

Santos, J. L., Verbert, K., Govaerts, S., & Duval, E. (2013). Addressing learner issues with StepUp! *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK '13*, 14. https://doi.org/10.1145/2460296.2460301

Sclater, N., Peasgood, A., & Mullan, J. (2016). *Learning Analytics in Higher Education A review of UK and international practice Full report*. Retrieved from https://www.jisc.ac.uk/sites/default/files/learning-analytics-in-he-v3.pdf

Sønderlund, A., Hughes, E., & Smith, J. (2019). The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology*, *50*(5), 2594–2618. https://doi.org/10.1111/bjet.12720

Soria, K. M., Fransen, J., & Nackerud, S. (2013). Library use and undergraduate student

outcomes: New evidence for students' retention and academic success. *Portal*, *13*(2), 147–164. https://doi.org/10.1353/pla.2013.0010

Thorpe, A., Lukes, R., Bever, D. J., & He, Y. (2016). The impact of the academic library on student success: Connecting the dots. *Portal*, *16*(2), 373–392. https://doi.org/10.1353/pla.2016.0027

Villano, R., Harrison, S., Lynch, G., & Chen, G. (2018). *Linking early alert systems and student retention: a survival analysis approach*. https://doi.org/10.1007/s10734-018-0249-y

Waheed, H., Hassan, S., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, *104*, 106189. https://doi.org/10.1016/j.chb.2019.106189

Wiese, C., & Newton, G. (2013). Use of Lecture Capture in Undergraduate Biological Science Education. *The Canadian Journal for the Scholarship of Teaching and Learning*, *4*(2). https://doi.org/10.5206/cjsotl-rcacea.2013.2.4

You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *Internet and Higher Education*, *29*, 23–30. https://doi.org/10.1016/j.iheduc.2015.11.003

Table 1: Mean, minimum and maximum end-of-year marks for each mark quintile.

Mark quintile	Mean mark (%)	Minimum mark (%)	Maximum mark (%)
Q1	50.5	33.4	56.0
Q2	58.7	56.0	61.3
Q3	63.5	61.3	65.8
Q4	68.2	65.8	71.0
Q5	75.8	71.1	90.6

Table 2: Mean, minimum and maximum end-of-year marks for each activity quintile.

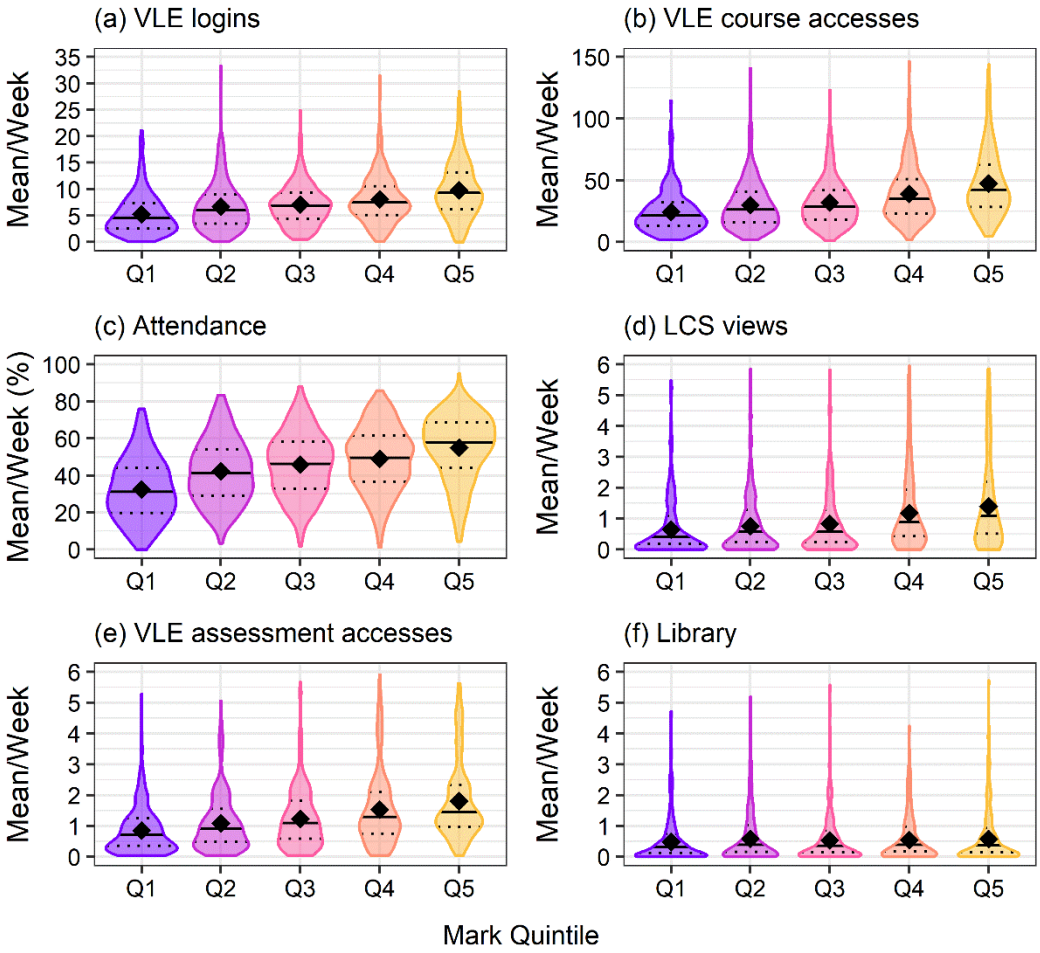
Activity quintile	Mean mark (%)	Minimum mark (%)	Maximum mark (%)
Q1	59.9	35.4	88.9
Q2	61.9	33.4	82.6
Q3	63.3	40.2	84.0
Q4	64.5	41.2	82.7
Q5	67.1	45.1	90.6

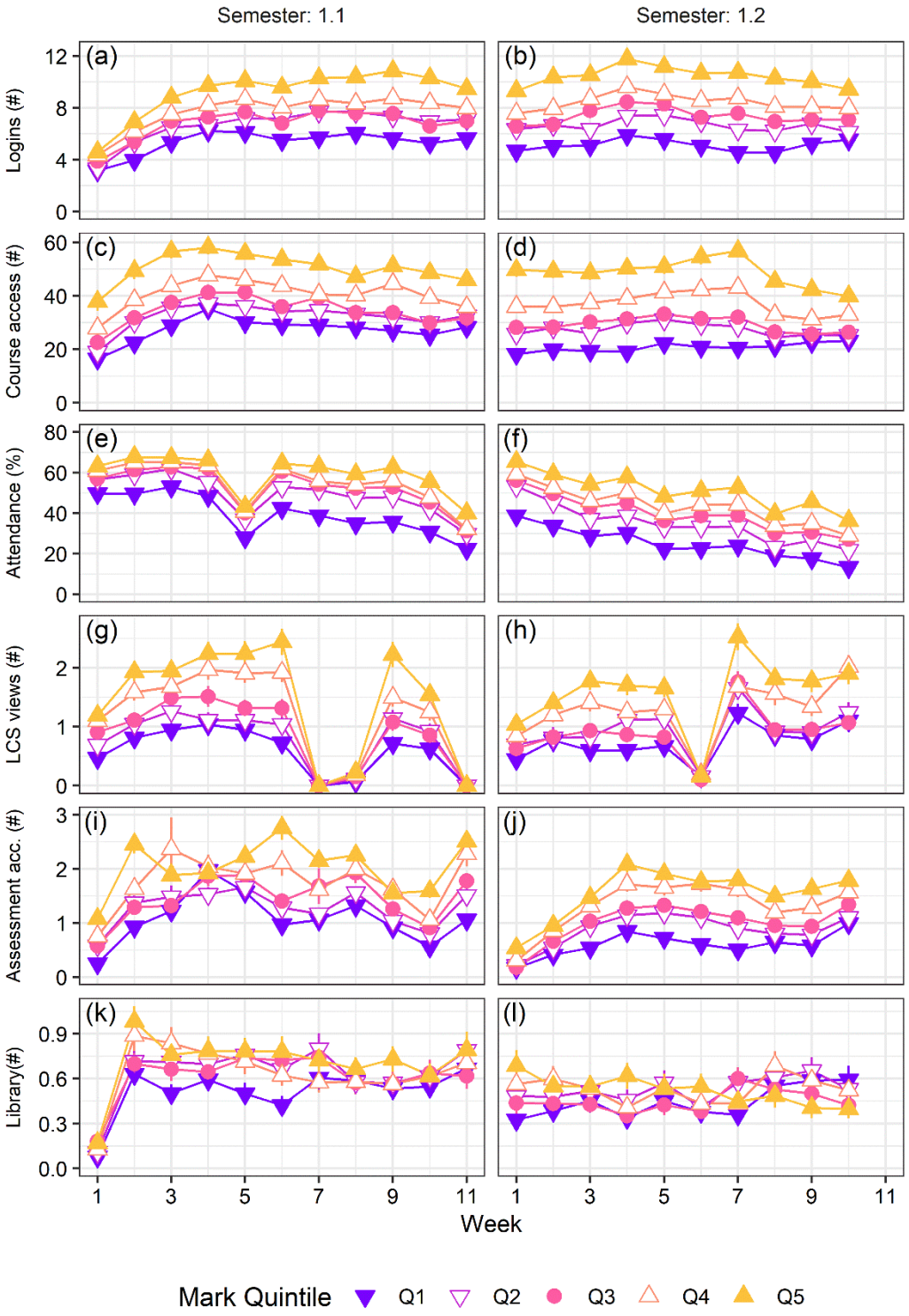
Table 3: Mean mark (%) for each activity quintile computed from the first 3 weeks of the teaching year (rows) based on the activity quintile computed from the last 3 weeks of the teaching year (columns).

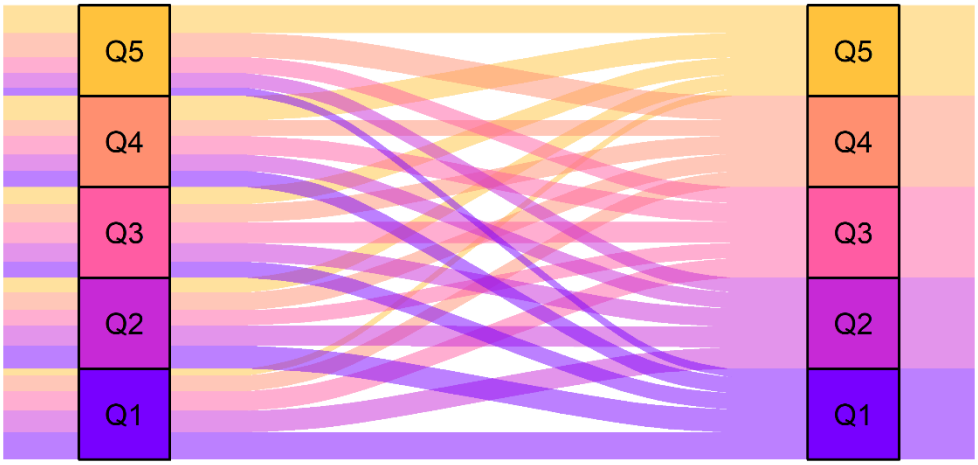
		Activity quintile computed from weeks 21-24				
		Q1	Q2	Q3	Q4	Q5
Activity quintile computed from weeks 1-3	Q1	59.3	59.3	60.3	61.9	61.0
	Q2	59.1	61.6	62.6	63.9	64.4
	Q3	61.4	62.5	63.0	63.7	66.7
	Q4	62.7	62.5	65.1	65.0	66.4
	Q5	63.0	67.1	64.0	65.3	69.3

Table 4: Coefficients of the linear regression model and their contribution to overall variance computed over data from the first three weeks of the academic year (left hand columns) and the whole academic year (right hand columns). P-values indicate the significance of t-tests on the individual model coefficients. Asterisks indicate the significance of the variance contribution of each component in an ensemble ANOVA; * = p<0.05, ** = p<0.01, *** = p<0.001.

		Data from the first three weeks				*Complete data*			
		B	β	p	r^2	B	β	p	r^2
Overall		-	-	-	0.297***	-	-	-	0.389***
Sex	Male	-0.831	-0.046	0.068	0.001	0.137	0.008	0.748	0.001
Ethnicity	Asian	-4.132	-0.227	<0.001	0.040***	-3.670	-0.202	<0.001	0.034***
	Black	-5.567	-0.218	<0.001		-4.636	-0.181	<0.001	
	Mixed or Other	-3.996	-0.106	<0.001		-3.344	-0.089	<0.001	
	Unknown	-0.951	-0.013	0.577		-0.892	-0.012	0.576	
POLAR4 Quintile	2	-1.026	-0.043	0.205	0.002	-1.155	-0.049	0.127	0.003
	3	-0.674	-0.034	0.373		-0.679	-0.034	0.337	
	4	-0.612	-0.026	0.445		-0.585	-0.025	0.435	
	5	-0.165	-0.008	0.833		0.061	0.003	0.933	
Socio-Economic Class	2	-1.060	-0.043	0.127	0.014***	-0.872	-0.035	0.178	0.013***
	3	-0.889	-0.030	0.254		-1.149	-0.039	0.114	
	4	-0.141	-0.005	0.859		-0.178	-0.006	0.809	
	5	0.557	0.012	0.605		0.112	0.003	0.911	
	6	-1.883	-0.058	0.028		-2.059	-0.063	0.010	
	7	-0.273	-0.010	0.722		-0.479	-0.017	0.502	
	8	-4.218	-0.059	0.009		-2.934	-0.041	0.051	
	9	-1.083	-0.048	0.110		-1.102	-0.048	0.081	
	Attendance (%)	0.089	0.211	<0.001	0.037***	0.139	0.287	<0.001	0.098***
	Course Access	0.023	0.255	<0.001	0.054***	0.005	0.278	<0.001	0.080***
	Library	0.089	0.029	0.224	0.003	-0.017	-0.034	0.127	0.002*
	LCS	0.090	0.058	0.029	0.014***	0.027	0.085	<0.001	0.021***
	Assessment Access	0.081	0.057	0.028	0.013***	0.036	0.093	0.062	0.034***
	Course Identifier	-	-	-	0.119***	-	-	-	0.104***


(a) VLE logins
(b) VLE course accesses
(c) Attendance
(d) LCS views
(e) VLE assessment accesses
(f) Library
Mean/Week
Mean/Week (%)
Q1
Q2
Q3
Q4
Q5
Mark Quintile


Semester: 1.1
Semester: 1.2
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
Logins (#)
Course access (#)
Attendance (%)
LCS views (#)
Assessment acc. (#)
Library(#)
Week
Mark Quintile
Q1
Q2
Q3
Q4
Q5


Q5
Q4
Q3
Q2
Q1
Q5
Q4
Q3
Q2
Q1
Activity Quintile
Mark Quintile

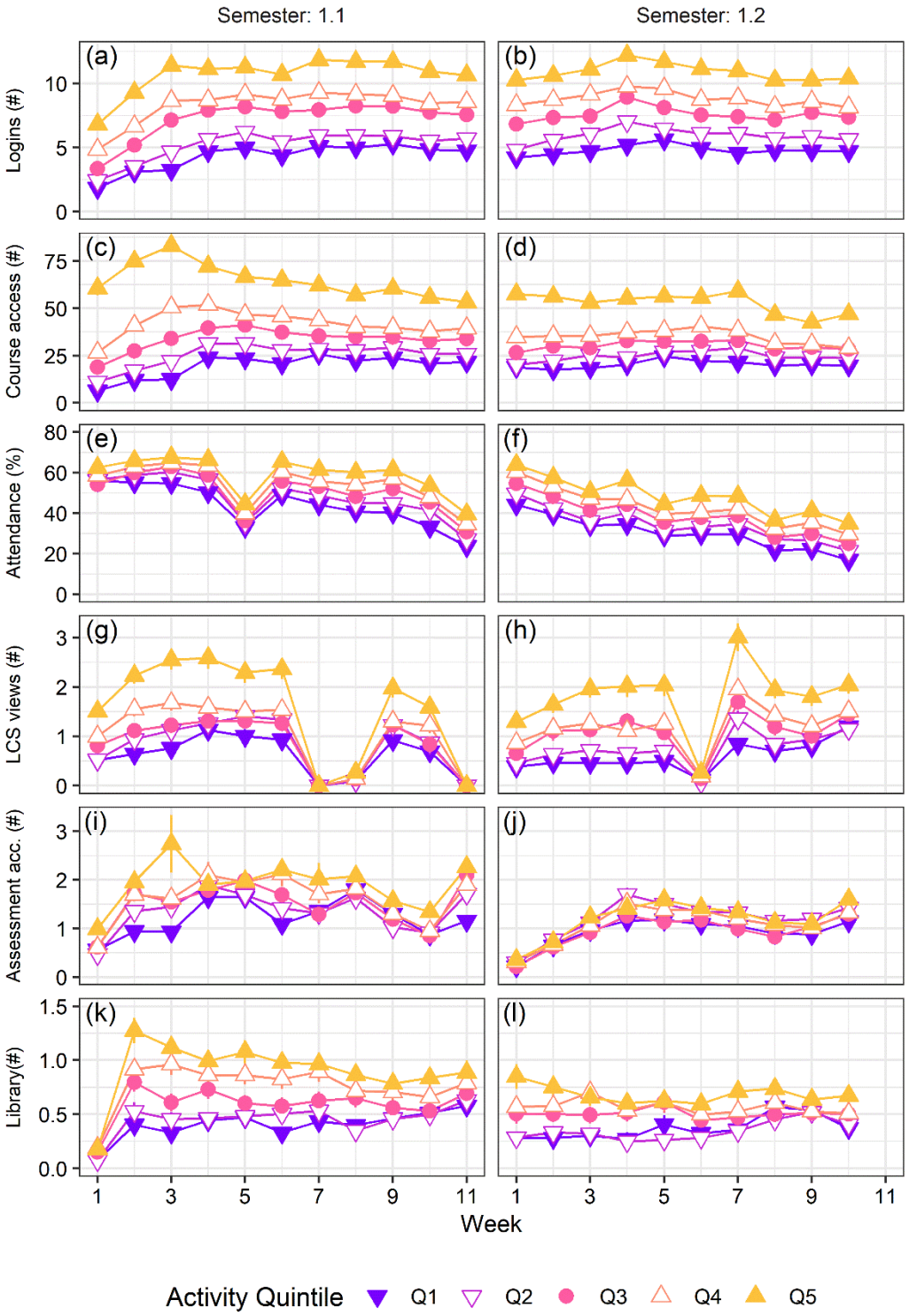

Semester: 1.1
Semester: 1.2
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
Logins (#)
Course access (#)
Attendance (%)
LCS views (#)
Assessment acc. (#)
Library(#)
Week
Activity Quintile
Q1
Q2
Q3
Q4
Q5

Figure 1: Violin plots showing the distribution of the mean weakly values for each data feed as a function of mark quintile. The mean value of each distribution is marked by the black diamond. The solid horizontal line indicates the median, and the dashed horizontal line below and above this are the 25^{th} and 75^{th} centiles.

Figure 2: Mean value of each data feed (rows) for each teaching week grouped by mark quintile (different colours/symbols) and semester (columns). Error bars indicate intra-student standard error of the mean. Dips in the metrics for attendance (semester 1 week 5) and LCS (Semester 1 weeks 7-8 & 11, and Semester 2 week 6) are caused by system problems.

Figure 3: Sankey diagram indicating the proportion of students from each activity quintile (left) that finish in each mark quintile (right).

Figure 4: Mean value of each data feed for each teaching week grouped by activity quintile (different colours/symbols) and semester (columns). Error bars indicate intra-student standard error of the mean. Dips in the metrics at for attendance (semester 1 week 5) and LCS (Semester 1 weeks 7-8 & 11, and Semester 2 week 6) are caused by system problems. As logins and course accesses are highly correlated the separation between these data feeds, particularly for weeks 1-3, is a direct result of the method used to compute activity quintile from the course access data.