Journal of Research on Educational Effectiveness

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uree20

Is Big Data Better? LMS Data and Prediction Accuracy in Postsecondary Education

Kelli A. Bird, Benjamin L. Castleman, Yifeng Song & Renzhe Yu

To cite this article: Kelli A. Bird, Benjamin L. Castleman, Yifeng Song & Renzhe Yu (15 Apr 2025): Is Big Data Better? LMS Data and Prediction Accuracy in Postsecondary Education, Journal of Research on Educational Effectiveness, DOI: 10.1080/19345747.2024.2308306

To link to this article: https://doi.org/10.1080/19345747.2024.2308306

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
https://doi.org/10.1080/19345747.2024.2308306

Routledge
Taylor & Francis Group

METHODOLOGICAL STUDIES

Is Big Data Better? LMS Data and Prediction Accuracy in Postsecondary Education

Kelli A. Bird[a], Benjamin L. Castleman[b], Yifeng Song[a], and Renzhe Yu[c,d]

[a]School of Education and Human Development, University of Virginia, Charlottesville, Virginia, USA; [b]Frank Batten School of Leadership and Public Policy, University of Virginia, Charlottesville, Virginia, USA; [c]Teachers College, Columbia University, New York, New York, USA; [d]Data Science Institute, Columbia University, New York, New York, USA

ABSTRACT

Colleges have increasingly turned to data-science applications to improve student outcomes. One prominent application is to predict students' risk of failing a course. In this article, we investigate whether incorporating data from learning management systems (LMS)—which captures detailed information on students' engagement in course activities—increases the accuracy of predicting student success beyond using just administrative data alone. We use data from the Virginia Community College System to build random forest models based on student type (new versus returning) and data source (administrative only, LMS only, or full data). We find that among returning college students, models that use administrative-only outperform models that use LMS-only. Combining the two types of data results in minimal increased accuracy. Among new students, LMS-only models outperform administrative-only models, and accuracy is significantly higher when both types of predictors are used. This pattern of results reflects the fact that community college administrative data contain little information about new students. Within the LMS data, we find that LMS data pertaining to students' engagement during the first part of the course has the most predictive value.

ARTICLE HISTORY
Received 26 August 2022
Revised 11 December 2023
Accepted 3 January 2024

KEYWORDS
Predictive analytics; college success; LMS data; big data; data science

Introduction

Colleges have increasingly turned to data-science applications and "big data" to better understand their students' needs, improve instructional delivery, and better target scarce resources (Fischer et al., 2020). These applications are both widespread and varied, ranging from adaptive learning algorithms that tailor instruction to students (e.g., Murphy et al., 2020); to natural language processing tools that automate writing guidance and assessment (e.g., Gayed et al., 2022; McNamara et al., 2013); to chatbots that respond to textual or verbal input and guide students through the college application process (Page & Gehlbach, 2017).

CONTACT Benjamin L. Castleman castleman@virginia.edu School of Education and Human Development, University of Virginia, Charlottesville, VA, USA.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/19345747.2024.2308306.

One of the most prominent applications of data science in higher education has been to predict students' risk of failing a course or dropping out. A third of all higher education institutions have invested in predictive analytics and collectively spend hundreds of millions of dollars to generate these predictions (Barshay & Aslanian, 2019). Most institutions use the "early alerts" generated by predictive models to initiate outreach from academic advisors or to encourage faculty to reach out to students in their classes who are struggling to succeed (Ekowo & Palmer, 2016; Klempin et al., 2018).

There is growing interest among higher education administrators and researchers in what combination of data sources can be leveraged to create meaningful predictors and, in turn, the most accurate predictions. The most common data source is institutions' administrative data, which include information on students' academic history (e.g., academic preparation, academic momentum, enrollment intensity) that education researchers have found to be strongly correlated with student success (Kuh et al., 2007; Pascarella & Terenzini, 1991, 2005; Tinto, 1994, 2012). More recently, the steady rise in digital learning systems (most prominently during the COVID-19 pandemic but also in the years preceding) has generated unprecedentedly rich data about students' moment-to-moment academic engagement. A prominent example is learning management software (LMS), which allows instructors to manage instructional content, organize learning activities, administer assessments, and monitor progress. Prior studies using various data-mining techniques demonstrate how fine-grained "behavioral traces" in LMS data can provide an understanding of students' learning processes and predict students' academic performance (e.g., Li et al., 2020; Lim, 2016; Park et al., 2018). However, the richness of the LMS behavioral trace data (referred to as "LMS data" in the remainder of this article) requires substantial analytic time and computing capacity. For instance, the raw LMS data we use in this study is roughly one to two terabytes for each term; working with data of this size requires significant storage space and processing power. In addition, the raw LMS data include records for each single action a student performs when interacting with the system; therefore, generating predictors that meaningfully describe students' experiences and actions often requires complex data transformations (Baker et al., 2020).

In this article, we systematically evaluate whether incorporating LMS data into course-performance prediction models substantially improves prediction accuracy beyond administrative data alone (henceforth "admin-only data"). Our primary goal is to inform the decisions of other researchers, policymakers, and administrators who are considering investing in the use of LMS data in predictive analytics. Our analysis builds on prior studies that have conducted exploratory analyses, at the level of a small number of courses, of the comparative predictive utility of LMS data to other data sources. For instance, in a study using data from 10 introductory STEM courses at a public research university, Yu et al. (2020) found that predictive models trained on small sets of predictors derived from admin-only data or LMS-only data both have reasonably strong accuracy, and that models trained on admin and LMS data together have the highest levels of accuracy. Aguiar et al. (2014) demonstrate that, among first-semester engineering students at Notre Dame, students' ePortfolio entries enhance predictive accuracy for whether students will persist into the next course in the engineering sequence. Crossley et al. (2016) use data from several hundred participants in a MOOC course to demonstrate that students' clicks in the MOOC interface and discussion content accurately predict whether students will complete the course.

While the predictive models investigated in these prior articles included several courses and hundreds to a couple thousand students, our article includes 2,646 courses across 23 institutions and 226,784 students across an entire state community college system, thus greatly increasing the generalizability of our results (though we discuss remaining considerations about generalizability in detail below). We also build on prior studies by conducting our analysis within the community college sector, which accounts for approximately 40% of all postsecondary enrollments, in which course failure and dropout rates are much higher, and where colleges have relatively limited information about their students before they begin their coursework. Insights on whether LMS data improves course performance predictions at the community college level could thus inform outreach and support efforts that have the potential to benefit a much larger and more at-risk population of students at broad-access institutions. Relative to prior articles, we also make the novel contribution of investigating whether the additional predictive value of LMS data varies for new versus returning college students. This distinction is important because, while colleges tend to have more information about returning students, new students are on average more likely to not succeed in their coursework, so colleges may have more interest in predicting success for new students.

We conduct our investigation using data from the Virginia Community College System (VCCS), which consists of 23 community colleges in the Commonwealth of Virginia. VCCS currently uses Canvas as their LMS. Across course modalities (i.e., in-person, online, or hybrid), instructors can use Canvas to organize and manage a variety of teaching and learning activities, such as submitting and grading assignments, sharing course materials, creating discussion forums, proctoring quizzes and exams, and, in the case of synchronous online courses, hosting virtual meetings.

Because the VCCS recently navigated to Canvas from a different LMS, we use data from after all colleges switched to the new system in Summer 2019 and extend the analytic sample through Spring 2021 (omitting Spring 2020, for reasons we detail below). We classify data we use to generate predictors into two broad categories: admin-only data and LMS-only data. Admin data includes measures such as student's cumulative GPA, prior credit accumulation, and current enrollment intensity, as well as course-level information like average historic grades and modality. LMS data include measures such as total time spent logged into the LMS and the number of on-time assignment submissions. There are some LMS predictors we include, such as the number of discussion forum posts, that are only applicable to the subset of course sections in which the instructor enabled that Canvas feature.[1] We use random forest prediction models to predict student performance, using a binary measure of success based on the student's final grade (A/B/C versus D/F/W). In order to assess the degree to which incorporating LMS data into course performance prediction models improves their accuracy, and how this differs based on student type, we build six separate

[1]As we describe below, in these cases we set the value of the LMS predictor to zero and include a separate missing-value indicator in the model.

models: (1) admin-only data, returning students; (2) LMS-only data, returning students; (3) both admin and LMS data ("full data"), returning students; (4) admin-only data, new students; (5) LMS-only data, new students; and (6) full data, new students. Finally, we explore the generalizability of our results to other contexts by showing how the predictive utility of the LMS data vary based on how Canvas is used by students and instructors.

Our article yields several primary conclusions and corresponding contributions to informing research and administrative practices. First, among returning VCCS students, the models trained on admin-only data substantially outperform models trained on LMS-only data and are reasonably accurate at predicting students' course performance: The admin-only model has a *C*-statistic (a general metric of prediction accuracy, which is also referred to as the AUC) of 0.855, while the LMS-only model has a *C*-statistic of 0.779.[2] Including both LMS and admin data results in only a slight, marginal improvement in prediction accuracy above the admin-only model (2% increase in *C*-statistic). This suggests that, for students with enrollment history in college, detailed measures of students' engagement derived from LMS data do not meaningfully improve our ability to predict their success in the course beyond the predictions we could generate just relying on measures of their prior academic performance. By contrast, among new VCCS students, the LMS-only model outperforms the admin-only model (*C*-statistics of 0.775 and 0.728, respectively), and combining the admin and LMS data results in a more significant increase in prediction accuracy (*C*-statistic of 0.825).[3]

Second, within the LMS data, we find that the predictors describing students' engagement during the first part of the target course have the most predictive value; predictors describing students' engagement in prior or concurrently taken courses are significantly less predictive of performance in the target course. This finding suggests that prediction-model developers could use a small fraction of the vast LMS data and achieve a similar level of accuracy. Third, the relative value of the LMS data in increasing prediction accuracy is highly variable across courses, with LMS data having the lowest value for predicting performance in math courses, in which students have the least amount of interaction with the LMS. Overall, our results show that the value of LMS data in predicting students' course performance is highest for new students and in courses where students and faculty engage more with the LMS.

[2]As we describe in the Methods section below, prediction accuracy is generally considered strong if the *C*-statistic is 0.8 or higher. The *C*-statistic is a metric ranging from zero to one but is not a percentage.

[3]We find a similar pattern when we define overall accuracy as the share of students for whom the model correctly predicted their outcome, which we present in Table 2. Assuming a population of 1,000 returning students, the admin-only model correctly predicts the outcome for 32 more students than the LMS-only model, and the full model correctly predicts the outcome for 15 more students than the admin-only model. Similarly, assuming a population of 1,000 new students, the LMS-only model correctly predicts the outcome for 49 more students than the admin-only model, and the full model correctly predicts the outcome for 34 more students than the LMS-only model.

Data

The data for this study come from two sources within the VCCS: (1) administrative records; and (2) behavioral trace data from Canvas LMS. The administrative data include detailed academic information from each term in which a student enrolls (beginning in Summer 2000), including their program of study, courses taken, grades earned, credits accumulated, and degrees or certificates awarded. Each student in the data has a unique anonymized identifier. Each instructor also has a unique anonymized identifier, which allows us to track instructors across courses and terms, beginning in 2008. We also observe whether instructors are full-time or adjunct. The LMS data come from Canvas Data, a Canvas service that provides institutions with optimized access to their data for reporting and queries. The raw data include detailed records of almost every single activity students perform in the system since Summer 2019. For example, when a student clicks into a specific page, the raw data will capture the time stamp of this visit, the content of the page, and the time when the page is available to students. We provide more detailed information about the raw data in Appendix A. The LMS data include the same anonymized identifier as the admin data, allowing us to join the two data sources.

Outcome

Our outcome of interest is a binary measure for successful course completion, and is equal to one if the student earned a grade of A, B, or C and equal to zero for grades of D, F, or W.[4] While a grade of D earns the student credit for the course and is considered a passing grade, within VCCS, students cannot satisfy some program requirements with a D, and other colleges and universities typically do not accept transfer credit for D grades.[5]

Sample

Our analytic sample includes students taking VCCS course sections that use Canvas from Summer 2019 through Spring 2021. Seventy-five percent of all VCCS course sections use Canvas, and our analytic sample consists of 81 percent of all VCCS

[4]We also built a multinomial random forest model with a categorical outcome of the six possible grades. We provide the confusion matrices for the models including the full set of predictors in Online Appendix Table A1. Comparing this output to Panel C of Table 2, we find that the multinomial random forest has nearly identical performance to the binary outcome specification. Because the multinomial random forest has slightly lower true negative rates, and is significantly more computationally intensive, we opted to use the binary outcome specification (instead of the multinomial) as our main model.

[5]Note that if we inverted the outcome such that we were predicting whether the student struggled in the course as observed by earning a D, F, or W, the *C*-statistics would be identical to what we report throughout the article, and the TPR for predicting failure would equal the TNR for predicting success (and vice-versa).

student × course section observations during this time frame.[6] We exclude Spring 2020 from the sample due to the extreme disruptions of COVID-19 on higher education, which included the VCCS shifting to an emergency grading policy that changed the standard grading scale such that the possible grades were P+, P-, Incomplete, or Withdraw. We further restrict the sample to focus on college-level coursework for regularly enrolled students. Specifically, we exclude observations corresponding to dual-enrollment (i.e., high school students taking college-level coursework). We also exclude all observations outside the traditional A–F grading scale. The vast majority of such observations correspond to developmental courses, which are graded as pass or fail.

As shown in Panel A of Table 1, the final sample includes 1,173,878 student × course section observations from Summer 2019 through Spring 2021. This translates to 226,784 unique students, 2,646 unique courses, and 63,994 unique course sections. We split the analytic sample into a training set and a validation set. The training set includes observations from the Summer 2019, Fall 2019, Summer 2020, and Fall 2020 terms; the validation set includes observations from Spring 2021. We use Spring 2021 as the validation sample with the intention of building a more generalizable model; specifically, having the observation window of the validation sample occur after the observation window of the training sample provides a more accurate estimation of model accuracy when applied to a practical setting (i.e., predicting current student success using a model trained on historical cohorts). We further split the analytic sample based on whether the student was enrolled at VCCS prior to the target term. As we detail below, if a student is in their first term and therefore has no prior academic history at VCCS, then we have far less information to include as predictors. Therefore, we build separate models for the observations in the analytic sample with no prior VCCS enrollment ("new-student" sample) versus observations in the analytic sample with at least one term of VCCS enrollment history ("returning-student" sample).

Panel B of Table 1 shows basic student characteristics for the full analytic sample and separately for the training and validation sets of the returning and new samples. Within the returning-student sample, students in the training and validation sets are similar on average. However, for the new-student sample, there are some differences in student-level characteristics. Compared to the training set, the validation set contains a significantly lower share of Hispanic students (13.1 percent versus 5.9 percent), a larger share of female students (55.2 percent and 58.4 percent) and significantly older students (22 versus 27 years old). These differences are likely due in large part to changes in the composition of the new-student population, with community colleges experiencing a 20.8 percent drop in new enrollments between Fall 2019 and Fall 2021 (National Student Clearinghouse, 2021). Panel C of Table 1 shows basic course characteristics across the relevant samples. Courses represented in the returning sample are more likely to be for 200-level, medical science, or applied technologies courses,

[6]Online Appendix Table A2 shows the summary statistics for the full VCCS population. Comparing this to Table 1, our analytic sample is quite similar to the full population

Table 1. Summary statistics of analytic sample.

	Full analytic sample	Returning-student sample		New-student sample	
		Training set	Validation set	Training set	Validation set
	(1)	(2)	(3)	(4)	(5)
Panel A: sample sizes					
Student × course × section observations	1,173,878	698,361	270,664	181,673	23,180
Unique students	226,784	164,245	87,022	63,603	8,196
Unique courses	2,646	2,246	1,989	1,399	966
Unique course sections	63,994	47,145	16,645	33,942	8,284
Panel B: student characteristics					
White	51.6%	52.3%	50.7%	48.9%	50.0%
Black	19.5%	19.0%	18.9%	19.1%	22.7%
Hispanic	13.0%	13.9%	14.4%	13.1%	5.9%
Asian	8.0%	7.9%	8.1%	8.6%	8.0%
Other	5.4%	5.2%	5.4%	6.1%	6.1%
Female	58.9%	59.9%	60.2%	55.2%	58.4%
Age	24.8	25.2	25.4	22	27
Cumulative GPA (at start of the target term)	2.91	2.91	2.88	N/A	
Credits accumulated prior to target term	32.6	32.6	32.5	N/A	
Panel C: course characteristics					
200-level	50.1%	48.5%	50.8%	39.3%	39.2%
Average course-level enrollment	153.9	156.4	147.7	257	276.2
Average section-level enrollment	18.3	18.6	17.7	20.6	20.4
Applied technologies	18.0%	16.7%	16.8%	14.2%	13.9%
Arts	9.7%	9.9%	10.1%	10.3%	7.8%
Business/finance	7.3%	7.4%	7.9%	8.6%	8.8%
Engineering	21.8%	21.0%	22.1%	20.4%	24.0%
Foreign languages	2.6%	2.8%	2.8%	3.6%	3.6%
Humanities	6.9%	7.2%	7.6%	9.3%	9.4%
Mathematics	1.0%	1.2%	1.1%	1.7%	2.1%
Medical sciences	19.9%	20.2%	17.5%	14.4%	10.8%
Natural sciences	3.3%	3.4%	3.6%	4.3%	5.4%
Social sciences	9.6%	10.2%	10.4%	13.3%	14.3%

Notes: student race and sex are averaged across unique students, while student age and prior academic history are averaged across unique student × term cells. Course characteristics are averaged at the course level (with the exception of section-level enrollment, which is averaged at the course × section level). The unit of observation in the prediction model is student × term × course × section. For both the 1st term and 2+ terms samples, the training set consists of data from the Summer 2019, Fall 2019, Summer 2020, and Fall 2020 terms; the validation set contains observations from the Spring 2021 term.

less likely to be social sciences or humanities, and have smaller enrollments; these differences reflect that returning students are further along in their academic careers.

Administrative Predictors

We construct 279 predictors from the admin data to characterize students' academic preparation, prior academic outcomes in community college classes, and enrollment timing and intensity choices. We describe these predictors here and motivate our predictor selection based on the longstanding literature devoted to understanding success factors of community college students. We also include a full list of our predictors in Online Appendix Table A3.

Our admin predictors can be divided into two broad categories: (1) non-course-specific academic records; (2) and course-specific academic records and characteristics. The non-course-specific category includes predictors which provide an overview of the students' academic progress thus far, such as cumulative GPA and total credits earned, all measured before the start of the target term (i.e., the term in which the student is taking the course for which we are predicting their performance). These predictors provide measures of students' overall academic preparation for the target term (Kuh et al., 2007). As Attewell et al. (2012) highlight, "academic momentum" is strongly related to student success, so we also include term-level GPA and credits attempted in the term prior to the target term. The non-course-specific category also includes a predictor equal to the share of a student's prior attempted credits that were developmental courses, which is another proxy for academic preparation (Boatman & Long, 2018). We also include information about students' previous enrollment patterns (e.g., whether the student has previously "stopped-out" of college) and their current enrollment intensity, both of which are linked to ongoing success (Crosta, 2014).

For the course-specific predictors, we include course-section characteristics, many of which are also related to student success according to prior literature. These include whether the course is taught online (Kofoed et al., Forthcoming), the enrollment count in the course section (Bedard & Kuhn, 2008; Cuseo, 2007), and whether the instructor is full-time versus adjunct (Bettinger & Long, 2010). Acknowledging that grading practices may differ substantially across subject areas—due potentially to differential grade inflation (Achen & Courant, 2009)—we also include the average grade in the target course in recent terms. We also include predictors related to the student's academic preparation for the specific target course, including the student's GPA in the target course's prerequisites and whether the student is retaking the target course.

We are limited in what financial aid and income information we can observe for students, so we are unable to include predictors in the model related to affordability or socioeconomic factors, though we acknowledge these are also important factors of student success (Bailey et al., 2004; Long, 2010). The one demographic factor we include as a predictor is student age, which is particularly important in the community college context where there are many non-traditionally aged students (CCRC, 2021). We are also limited in what we can observe for new students, as they have no prior academic records. For these new-student observations, we include 59 predictors related to general term-level information (e.g., the student's current enrollment intensity) and characteristics of the target course (e.g., average grade in the course in recent terms).

LMS Predictors

LMS data provide comprehensive and fine-grained information about how students engage with the system and can capture important constructs and processes in students' learning experience, such as cognitive strategies, affective states, and self-regulated learning (Fischer et al., 2020). Therefore, an abundance of learning-analytics research has created various behavioral predictors from such trace data to predict learning outcomes (Gardner & Brooks, 2018; Wang & Mousavi, 2022). Importantly, behavioral predictors can be highly contextualized: different systemic or instructional decisions define the scope of behavior that students can exhibit and the meaning of specific behavioral patterns (Gašević et al., 2016). Predictors that incorporate more contextual information (e.g., the downloading of a specific document referenced on a course's syllabus) tend to be more predictive of performance within a particular course but less universally available or meaningful (Arizmendi et al., 2022). In this study, we investigate the value of behavioral trace data in institution-level models, where using relatively generalizable predictors is preferable both from a prediction accuracy standpoint because of the large sample size it allows for, and from a feasibility standpoint due to the time required to generate highly contextual predictors for each course section. As such, we surveyed the literature that models student behavior in LMS in postsecondary contexts, and identified the most widely used behavioral measures that are relatively straight-forward to compute and that are predictive of academic performance: number of click actions, study sessions, and total time online (Cicchinelli et al., 2018; Conijn et al., 2017; Zacharis, 2015); average duration and irregularity of study sessions (Conijn et al., 2017); number of active days (McCuaig & Baldwin, 2012) and weeks (Choi et al., 2016); number of assignment submissions (Macfadyen & Dawson, 2010; Motz et al., 2019) and proportion of on-time submissions (Heo et al., 2019); number of original discussion posts and discussion replies (Sher et al., 2019); average length (Sher et al., 2019) and depth (Barbosa et al., 2020) of discussion posts. These cross-context measures are mostly derived from simple aggregation of the raw data (in contrast to complicated data-mining techniques) and capture engagement with either anything or the most popular instructional activities in the system.[7]

We compute the measures based only on students' early-course behavioral traces, because behavioral predictors are mostly meaningful to instructors if they can help predict course performance in an early stage to allow for targeted interventions. While we focus on performance prediction within a target course, we also take into account students' behavior in their concurrently and previously enrolled courses, which provides more comprehensive contextual information about their focal behavior, but is typically missing in learning analytics research. We compute the same behavioral measures from these additional courses and include them as predictors as well.

With these considerations, we construct a total of 50 predictors from the LMS data. We describe them below, and include a full list in Online Appendix Table A3. We also briefly describe the process of cleaning the raw data and generating these predictors in Appendix A to guide others attempting to do similar work.

- Early-term target course: measures of engagement in the target course during the first quarter of the course period (e.g., total number of click actions, total time

[7]While conceptually assignment grades would be good at predicting course performance, we did not include them due to critical concerns about data quality issues. Instructors use the "assignment" functionality in the LMS in various ways. While some use the function for what we may perceive as assignments, other instructors use them for class attendance, optional practice tests, or for other miscellaneous or idiosyncratic purposes and, in many of these cases, the grades are not consistently recorded. Therefore, using assignments grades for prediction would incur additional noise and biases.

spent online, percent of on-time assignment submissions), which we measure relative to the course section average[8];
- Early-term concurrent courses: the same early-term measures of engagement in all other courses taken in the same term as the target course, averaged across these courses;
- Prior early term: the same early-term measures of engagement in all courses taken in prior terms, averaged across these courses;
- Prior full term: the same measures of engagement metrics in all prior courses, computed across the entire term instead of the first quarter and averaged across these courses.

If a student is in their first term, we only include the 21 predictors measuring engagement in current courses. All of the LMS predictors are normalized using the z-score within each term × course × section cell; in other words, the LMS predictors measure a student's Canvas activity in a particular course, relative to other students in the same course section.[9] This standardization accounts for differences in engagement due to differences in the use of LMS across courses, instructors, and modalities.

Handling Missing Values

There is expected missingness in the data; for example, if a student has not taken prior courses in an academic cluster, then the average grade of prior courses in this cluster is missing. For another example, if an instructor does not enable the discussion feature of Canvas for their course section, then there is no information with which to compute the discussion-related predictors. All instances of missing predictors are due to these "not applicable" situations. We handle this missingness by setting missing values equal to zero and include indicators for whether a given predictor is missing.[10]

Methods

We use a random forest model to predict successful course completion. Random forest is a tree-based ensemble model commonly used in data-science research for predictive analytics. In other work where we investigate degree-completion prediction models using admin data from the VCCS (Bird et al., 2021), we find similar levels of accuracy for random forest and other commonly used models. For this article, we initially tested

[8]Within an academic term, different courses may vary in start date, end date, and length, so the measures are computed in relation to the specific period of each course. The first quarter of the course period is defined by dividing the total length of the course (in weeks) by four, rounded up to whole weeks. For example, for a course that lasted 10 weeks, early-term measures are computed based on the first 3 weeks of behavioral trace data.

[9]Z-score normalization is applied to the predictors within each term × course × section cell, such that normalized values of each predictor within the cell has mean 0 and variance 1.

[10]If we instead do not perform this imputation of missing values and allow the random forest model to differentiate between actual zeros and the missing values, our results are nearly identical (see Online Appendix Table A12).

other models to predict course success, and random forest slightly outperformed the others. We use 5-fold cross-validation to tune the random forest model (i.e., choosing the optimal number of decision trees, the maximum depth, and the number of random features to include at each node for splitting), which reduces the risk of model over-fitting (Breiman, 2002; Ghojogh & Crowley, 2019).[11] All evaluation metrics we report are from a hold-out validation sample.

Our primary objective is to compare the predictive accuracy of models using the admin versus LMS data. Therefore, we estimate models using (1) admin-only predictors; (2) LMS-only predictors; and (3) full set of predictors. For each of these three settings, we build separate models on the new-student and returning-student samples. To compare the accuracy of these six main models, we report the following evaluation metrics:

- *C*-statistic: a "goodness of fit" measure that is equal to the probability that a randomly selected positive observation (i.e., a student who passed a particular course) has a higher predicted score than a randomly selected negative observation. The *C*-statistic is also referred to as the AUC, which stands for area under the ROC curve. A *C*-statistic of 0.5 corresponds to a model being no better than choosing at random, while a *C*-statistic of 1 corresponds to a model perfectly predicted the outcome. A *C*-statistic of 0.8 or higher is considered strong performance, and a *C*-statistic of 0.9 or higher is considered outstanding (Hosmer et al., 2013). We provide standard errors for *C*-statistics following Hanley and McNeil (1982).[12]
- True positive rate (TPR): share of true positives that the model correctly predicts as succeeding (also called "recall").[13]
- True negative rate (TNR): share of true negatives that the model correctly predicts struggling (also called "specificity").

We also estimate course-specific models for five of the largest courses offered by VCCS: General Biology I (BIO101); College Composition I (ENG111); College Composition II (ENG112); Quantitative Reasoning (MTH154); and Pre-Calculus I (MTH161). We estimate admin-only, LMS-only, and full predictor models for each of these courses. With the exception of ENG111, the vast majority of observations (particularly in the validation sets) for these courses are from returning students; therefore, we combine the new- and returning-student samples for the course-specific models.[14]

[11]In Online Appendix Table A4, we display the optimized values of these parameters for the six main models. For other model parameters, we use the default values set by the Python scikit-learn library.

[12]This standard error measures the uncertainty in applying the fixed predictive model obtained from the current training set to a new validation set, rather than the uncertainty resulting from fitting the predictive model on a new training set.

[13]Another common evaluation metric is precision, which is the share of observations that the model predicts will succeed that actually succeed. The pattern of the precision values very closely follows our results for TPR; therefore, we display only TPR for concision.

[14]We include an indicator variable for whether each observation corresponds to a course taken during the first term. If a predictor (e.g., cumulative GPA prior to taking the course) is not available for the new-student observations, the value of that predictor is set to 0 for all new-student observations.

Because our outcome is binary, the immediate output of the model is a predicted score for each observation ranging from zero to one, with a value closer to one indicating a higher predicted probability of course "success" (earning an A, B, or C). Therefore, we set a threshold in predicted score to delineate observations into two categories: those predicted to successfully complete the course (i.e., those with a predicted score at or above the set threshold), and those predicted to not. We set the threshold equal to the course completion rate within the training sample used for each model (77.8% for the returning training sample, and 70.9% for the new-student training sample).[15]

Results

Comparing Prediction Accuracy Based on Types of Predictors

In Figure 1, we present several accuracy metrics for the prediction models using admin-only data, LMS-only data, or both admin and LMS data. Panel A presents *C*-statistics, while Panels B and C present true-positive and true-negative rates, respectively. Within each panel, we present accuracy metrics for the returning-student sample on the left and for the new-student sample on the right. As we show in Panel A, the prediction model trained on the returning-student sample with admin-only data achieves a high level of accuracy, with a *C*-statistic of 0.855. The model trained on LMS-only data and the returning-student sample has substantially lower accuracy, with a *C*-statistic of 0.779. Combining both admin and LMS data with the returning-student sample leads to modestly higher accuracy than we obtain with the admin-only data, with a *C*-statistic of 0.872. The standard errors of these *C*-statistics, shown in parentheses below the *C*-statistic values, are all less than 0.001, indicating that these are statistically distinct from one another.

Among the new-student sample, on the other hand, we find comparatively greater predictive value from the LMS-only data: whereas the model trained on admin-only data has a *C*-statistic of 0.728, the model trained on LMS-only data has a *C*-statistic of 0.775. Combining both admin and LMS data leads to a proportionally greater gain in accuracy (*C*-statistic of 0.825) than we observed in the returning-student sample. Each with standard errors of 0.003, these *C*-statistics are also statistically distinct from one another.

This pattern of relationships makes intuitive sense. First, the prediction models are more accurate for the returning-student sample than the new-student sample, which we would expect given that we have more data—and in particular more observed academic performance—on which to train the model.[16] Second, the comparative value of LMS measures of engagement is higher for new students, for whom baseline data on academic performance is much more limited.

[15]While *C*-statistics are independent of the threshold, TPR and TNR can be highly sensitive to the threshold chosen. Other common methods used to set the threshold, such as maximizing the F1 score, can result in significant differences in thresholds set from model to model. Our approach allows for better comparison of TPR and TNR across models.

[16]It is possible that the lower accuracy is due (at least in part) to the smaller number of observations for new students. However, as we show in Table 1, there are still over 200,000 observations in the new-student sample, so we think it is unlikely that the smaller sample size is driving the lower accuracy of the new-student models.

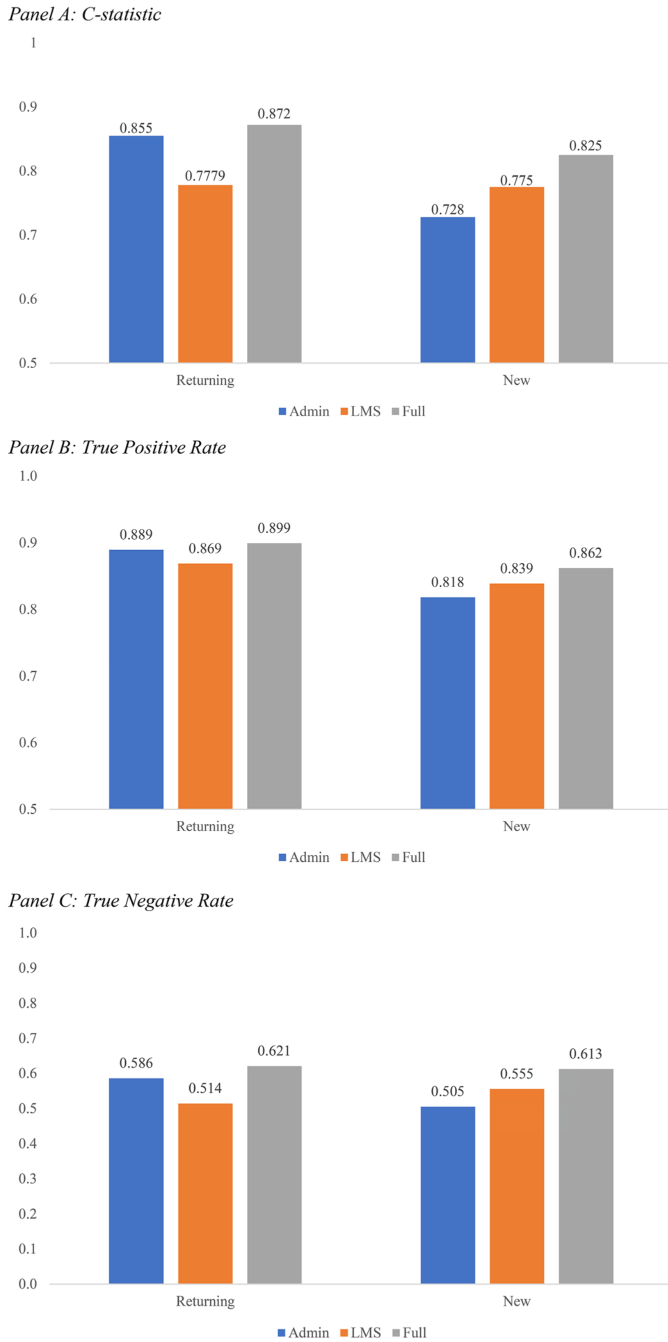

Figure 1. Prediction model accuracy by category of predictors and sample of students.
Notes: each bar corresponds to a separate random forest prediction model using the set of predictors indicated by the color of the bar, and observations from the sample of students based on academic history indicated by the x-axis label.

This same basic pattern of relationships holds when we consider TPR or TNR as our accuracy metrics instead of *C*-statistics (Panels B and C of Figure 1). All models achieve high TPRs, with the highest TPR among the returning-student sample and with the model trained on both admin and LMS data. Specifically, 89.9% of students who actually earn an A, B, or C are predicted by the model to complete the course. The TNRs are significantly lower than the other model evaluation metrics we consider (ranging from 50.5–62.1%) but follow the same pattern across the six model variations. The relatively low TNR is an expected result when a large majority of observations achieve the outcome, as is the case with our models (Spelmen & Porkodi, 2018).[17]

Table 2 displays confusion matrices for the six models, which provide a more fine-grained comparison of the models' predictions with students' actual outcomes. These matrices show the share of observations for each combination of predicted outcome (A/B/C versus D/F/W) and actual grade received (A, B, C, D, F, or W). Intuitively, we find that the model is more accurate for the highest and lowest grades compared to middle grades. For example, for the returning-student model with full predictors, 95.5% of students who actually received an A were Pred(ABC) compared to 74.4% of students who received a C. Most starkly, only 37% of students who received a D were Pred(DFW), compared to 62% of students who received an F and 77.9% of students who received a W.[18] We also report the overall accuracy at the bottom of each matrix, which is equal to the percent of observations with accurate predictions: Pred(ABC) with actual grades of A, B, or C-plus Pred(DFW) with actual grades of D, F, or W. These overall accuracy rates follow a very similar pattern to the other evaluation metrics shown in Figure 1.

We acknowledge that colleges are most interested in using course performance-prediction models to identify students at risk of failing, and that these relatively low TNR values may raise concerns about the overall practical value of these models.[19] Generally, though, our models' accuracy is similar to (or better) than other those found in other articles predicting course accuracy that incorporate LMS data. In Yu et al. (2020), the TPRs of their models using institutional or click-data range from 0.687 to 0.750, and the TNRs range from 0.515 to 0.649. In Aguiar et al. (2014), the *C*-statistic for the model with all-academic data is 0.654, and the *C*-statistic for their best performing model that includes top-academic + engagement data is 0.929. In Crossley et al. (2016), their overall accuracy rate is 76.1 percent. In assessing prediction accuracy, it is also important to consider the counterfactual of how scarce student supports would be

[17]In instances where the failure rate is higher, we would expect to see larger TNR values. we observe this when we build similar models using admin data to predict degree completion (Bird et al., 2021).

[18]These grade-specific accuracy rates are quite similar to the alternative multinomial random forest model. Our main model (which uses the binary outcome ABC versus DFW) has slightly lower accuracy rates for students with grades A, B, and C, but slightly higher accuracy rates for students with grades D, F, and W. See Online Appendix Table A1.

[19]We attempted to improve the models' true negative rates by upweighting observations of students who did not achieve the outcome. Specifically, we multiplied each observation by a factor that is equal to the inverse of the frequency of its outcome. For instance, if 70% of observations in the training sample whose outcome is success and 30%of observations whose outcome is failure, then each success observation will be multiplied by 1/0.7 = 1.43, and each failure observation will be multiplied by 1/0.3 = 3.33. Online Appendix Table A5 shows the results of these upweighted models. We find very similar levels of prediction accuracy in the upweighted model compared with our main version (shown in Figure 1); the TNRs of the upweighted model are all slightly lower than the main version.

Table 2. Confusion matrices.

Panel A: admin predictors

Returning-student sample				New-student sample			
Actual grade	Pred (ABC)	Pred (DFW)	Total	Actual grade	Pred (ABC)	Pred (DFW)	Total
A	40.1%	2.4%	42.4%	A	37.4%	6.0%	43.3%
B	20.2%	3.1%	23.3%	B	13.8%	4.1%	17.9%
C	8.9%	3.1%	12.1%	C	6.9%	2.8%	9.7%
D	2.9%	1.7%	4.6%	D	3.0%	1.7%	4.7%
F	4.7%	6.2%	10.8%	F	9.7%	6.3%	16.1%
W	1.6%	5.2%	6.8%	W	1.7%	6.6%	8.3%
Total	78.4%	21.6%	100.0%	Total	72.4%	27.6%	100.0%
% Observations with accurate prediction = 82.2%				% Observations with accurate prediction = 72.7%			

Panel B: LMS predictors

Returning-student sample				New-student sample			
Actual Grade	Pred(ABC)	Pred(DFW)	Total	Actual Grade	Pred(ABC)	Pred(DFW)	Total
A	38.9%	3.5%	42.4%	A	38.3%	5.1%	43.3%
B	19.6%	3.7%	23.3%	B	14.2%	3.7%	17.9%
C	9.0%	3.0%	12.1%	C	6.9%	2.7%	9.7%
D	3.0%	1.6%	4.6%	D	3.0%	1.7%	4.7%
F	4.8%	6.0%	10.8%	F	6.8%	9.3%	16.1%
W	2.9%	3.9%	6.8%	W	3.2%	5.1%	8.3%
Total	78.4%	21.6%	100.0%	Total	72.4%	27.6%	100.0%
% Observations with accurate prediction = 79.0%				% Observations with accurate prediction = 75.6%			

Panel C: full predictors

Returning-student sample				New-student sample			
Actual grade	Pred (ABC)	Pred (DFW)	Total	Actual grade	Pred (ABC)	Pred (DFW)	Total
A	40.5%	1.9%	42.4%	A	39.7%	3.6%	43.3%
B	20.4%	2.9%	23.3%	B	14.6%	3.3%	17.9%
C	9.0%	3.0%	12.1%	C	6.8%	2.8%	9.7%
D	2.8%	1.7%	4.6%	D	2.9%	1.8%	4.7%
F	4.1%	6.7%	10.8%	F	6.6%	9.5%	16.1%
W	1.5%	5.3%	6.8%	W	1.8%	6.5%	8.3%
Total	78.4%	21.6%	100.0%	Total	72.4%	27.6%	100.0%
% Observations with accurate prediction = 83.7%				% Observations with accurate prediction = 79.0%			

Notes: each of the six groupings shows the confusion matrix for the prediction model that includes the set of predictors indicated by the column heading (admin, LMS, full), and the sample of observations based on timing (returning, new). Within a confusion matrix, each cell contains the share of observations in the validation sample who received a grade as indicated by the row labels, and was predicted to receive a grade as indicated by the column labels. Note that the column "Total" contains the sum of observations within each row, while the row "Total" contains the sum of observations within each column.

allocated in the absence of the prediction model. As we show in Online Appendix Table A6, the TNRs of our full models are 23 –37% higher than if we used cumulative GPA or total Canvas clicks alone to identify at-risk students, and two to three times greater than if we used random guessing.[20] While the models clearly provide a meaningful improvement above these basic targeting strategies, it is ultimately up to the

[20]Specifically, the TNRs in columns (2) and (3) are based on predicting that the students with values of cumulative GPA or total clicks below the Xth percentile would earn a D/F/W, where X is equal to the mean D/F/W rate in the training samples. These mean D/F/W rates are also equal to the TNR values in column (4).

discretion of administrators whether the amount of model error is acceptable within their context.[21]

Next, we explore whether certain categories within the admin and LMS data are more predictive of course performance. If institutional researchers face time or computational constraints in constructing predictors, or if they wish to generate predictions prior to or at the beginning of a term, we believe it would be informative to understand the relative predictive value of these various categories. We divide the admin predictors into two separate categories: non-course-specific records and course-specific records. We divide the LMS data into four categories: early-term target course, early-term concurrent, prior early term, and prior full term. Each of these categories are described in the Data section above. Some categories contain more-complexly specified predictors than others, and some categories are only available after the target course is under way (i.e., the early-term categories). We display results in Table 3, where each row corresponds to a separate model trained just on the categories of predictors indicated. Panel A presents the *C*-statistics for the sample of students with prior VCCS experience while Panel B presents the *C*-statistics for the sample of students in their first term. The first two rows in Panel A repeat what we have shown earlier: we obtain the highest accuracy level from the model that leverages all admin and LMS predictors, but the model trained on admin-only predictors achieves nearly the same level of accuracy. Within this total set of admin predictors, a subset of 41 predictors that measure a combination of overall (i.e., not course-specific) academic performance and students' age achieves similar accuracy (*C*-statistic = 0.841). By comparison, a model trained on 238 course-specific predictors has notably lower accuracy (*C*-statistic = 0.778). The remaining rows in Panel A present *C*-statistics for models trained on different combinations of LMS predictors. The subset of LMS predictors that measure students' engagement in the target course contribute substantially more to prediction accuracy than LMS measures of students' engagement in prior or concurrent courses. Among the new-student sample (Panel B), we observe a generally similar pattern of results, though, as we show earlier, overall prediction accuracy among this sample of students is lower.[22]

[21]For instance, a college may be willing to accept lower levels of model accuracy if they are using the predictions to target a low-cost messaging campaign to at-risk students for which there is little perceived downside if well-performing students also receive the messages. However, if a college instead wants to target resource-intensive supports to at-risk students and they are tightly constrained in the number of students they can provide these supports to, then the college would likely want to invest substantial time in improving their prediction accuracy.

[22]We complement this analysis by also calculating feature importance (FI) scores for the models with full predictors; we report the top 30 predictors of the returning-student and new-student models in Online Appendix Table A7. The FI scores are based on the mean decrease in impurity and provide a metric of each predictor's contribution to the overall model's accuracy (Breiman 2002). Consistent with what we show in Table 3, most of the highest feature-importance predictors for the returning-student model capture some aspect of students' prior credit accumulation and GPA, while most of the highest feature-importance predictors for the new-student model are LMS measures from the target course. Five of the top 10 predictors in terms of feature importance are common between the two samples of students: the number of total credits attempted in the target term, the two measures of historic performance in the course, and the two LMS measures of student engagement.

Table 3. *C*-statistics of models using different predictor subcategory combinations.

Panel A: model with returning-student observations

Predictor categories	# predictors	*C*-statistic	Std Err
All	329	0.872	(0.0007)
All admin	279	0.855	(0.0007)
Non–course specific	41	0.843	(0.0008)
Course-specific	238	0.778	(0.0009)
All LMS	50	0.778	(0.0009)
Early-term target course + early-term concurrent	21	0.751	(0.0010)
Early-term target course	12	0.733	(0.0010)
Early-term concurrent	9	0.604	(0.0012)
Prior early-term + prior full-term	29	0.713	(0.0011)
Prior early-term	13	0.665	(0.0012)
Prior full-term	16	0.709	(0.0011)

Panel B: model with new-student observations

Predictor categories	# predictors	*C*-statistic	Std Err
All	80	0.825	(0.0027)
All admin	59	0.728	(0.0034)
Non–course specific	34	0.602	(0.0039)
Course-specific	25	0.664	(0.0037)
All LMS (early-term target course + concurrent)	21	0.775	(0.0031)
Early-term target course	12	0.754	(0.0032)
Early-term concurrent	9	0.595	(0.0040)

Notes: each row corresponds to a separate random forest prediction model using the set of predictors indicated in the first column. All prior LMS predictors and course-subject-specific predictors are not available for new-student observations; some course-specific and non–course specific academic records are unavailable for new-student observations.

Generalizability

Given our primary goal of informing research and practice and the value-addition of LMS data in predictive modeling, it is important to consider how we expect our results to generalize in different settings. There are several important contextual considerations: (1) our setting is a community college system, which is open-access and enrolls a diverse student body; (2) our data spans a relatively short time window, which includes COVID; and (3) VCCS instructors have a great amount of flexibility in how they set up their Canvas pages, and these course design decisions are likely related to the type of content covered in the course. In general, we would expect greater additional predictive value of the LMS predictors within other contexts where the LMS is used more comprehensively or consistently. For example, if a college required that all instructors maintain accurate gradebooks on Canvas (which is not the case for VCCS, at least during our sample window), then the predictive value of the LMS data could be substantially higher. Conversely, if a college has more comprehensive information about students before they matriculate (e.g., high school transcripts and entrance exam scores, which are often required by selective four-year institutions), then it's likely that the LMS data would add less predictive value beyond the administrative data for first-term students.

While it is not possible to provide definitive answers to how our results will generalize in other contexts, we can provide some additional details to consider. First, we provide summary statistics for all early-term LMS predictors for the target course in Table 4. Column (1) shows the mean, median, and standard deviation of the 12 predictors for the full analytic sample. If the values of these LMS predictors are substantially different in another context, then our main conclusions may not hold in that context. In columns (2) through (6) of Table 4, we further show how these predictor values vary across the time line of our analytic sample. Not surprisingly, the mean and median values are typically higher for the COVID-impacted terms of Summer 2020, Fall 2020, and Spring 2021, when most instruction was still occurring online. Summer 2020 and Fall 2020 make up roughly half of the training sets, and Spring 2021 constitutes the full validation sets. Therefore, if a college has since departed from its COVID-influenced online instruction practices, then again our main conclusions may not hold in that context.[23]

To provide some concrete examples of how context can influence our results, we next explore how the relative value of the LMS data varies across VCCS courses. As other researchers have noted (e.g., Baker et al., 2020), the value of LMS predictors is driven in some part due to course-specific context. English instructors may structure their courses on the LMS significantly differently than math instructors. These differences mean that some LMS predictors may be more or less valuable; for instance, a low value for frequency of discussion-forum posts could indicate either a student is either unengaged, or alternatively that discussion forums are not an important part of the course design. We explore the question of how results differ across courses in two ways. First, we compute a separate *C*-statistic for each of the top 50 courses by applying our models trained on the full sample to course-specific validation sets. We present these results in Figure 2, where we observe substantially more variation in the accuracy of the LMS-only models compared to the admin-only or full predictor models. Within the LMS-only models, math and science courses consistently have the lowest *C*-statistic—all courses with *C*-statistics below 0.75 are either math, chemistry, biology, or IT. This is not the case for the admin-only models, with several chemistry and biology courses having *C*-statistics above 0.85. Second, we compare the accuracy of models trained on course-specific samples (e.g., all students who enroll in English 111, the College Composition course offered across the VCCS). We focus this analysis on five large-enrollment courses in core subjects that typically function as “gateways” for students to take higher-level courses and fulfill degree requirements across most VCCS programs of study. Specifically, we build course-specific performance prediction models for the two-course sequence of College Composition (ENG111 and ENG112); General Biology (BIO101); and two introductory, college-level math courses, Quantitative Reasoning (MTH154) and Pre-Calculus I (MTH161).[24]

[23]Because all course sections, regardless of modality, may use the LMS for a variety of course aspects, we include observations from online, in-person, and hybrid course sections in our analytic sample. We explore differences in prediction accuracy for modality-specific models in Appendix B and intuitively find that the accuracy of the LMS-only models are substantially higher for the online sample compared to the in-person sample, and that the predictive value-add of the LMS data is higher for the online sample.

[24]In Online Appendix Table A8, we present summary statistics for these courses. Each course is offered in hundreds of sections each term across the 23 VCCS colleges, and each enrolls thousands or even tens of thousands of students per term. Student performance across these courses tends to be relatively low, with mean GPAs ranging from 2.22 in MTH161 to 2.73 in ENG112. All five courses have a high rate of students earning a D, F, or W, which range from 26.7 percent in BIO101 to 41.5 percent in MTH161. A sizable share of enrollments in four of the five courses (all except ENG112) are students in their first term at VCCS. For instance, 25% of students in MTH154 and 52.1% of students in ENG111 are in their first term.

Table 4. Summary statistics for early-term LMS predictors for the target course.

		Full Sample	Summer 2019	Fall 2019	Summer 2020	Fall 2020	Spring 2021
		(1)	(2)	(3)	(4)	(5)	(6)
Panel A: term-specific samples							
Total time, in minutes	(Mean)	518	399	357	563	704	519
	(Median)	335	255	209	396	497	344
	(Std dev)	599	471	455	578	720	583
Total click count	(Mean)	875	746	619	933	1115	920
	(Median)	641	535	421	721	862	687
	(Std dev)	918	808	687	902	1075	922
Average session length, in minutes	(Mean)	12.86	15.57	9.29	17.13	14.73	12.15
	(Median)	8.35	10.95	5.77	11.47	9.99	8.16
	(Std dev)	15.01	16.94	11.74	18.73	16.04	13.67
Standard deviation of session length	(Mean)	26.30	28.20	20.03	31.18	30.80	25.68
	(Median)	20.53	22.34	14.91	24.72	25.16	20.23
	(Std dev)	22.14	23.73	19.24	24.54	23.09	20.63
Assignment Submission count	(Mean)	8.11	8.37	7.94	7.83	8.08	8.39
	(Median)	5.00	5.00	5.00	4.00	5.00	4.00
	(Std dev)	10.68	11.11	10.20	10.38	10.55	11.40
Assignment submission available	(Mean)	50.7%	45.7%	53.7%	45.2%	54.5%	47.3%
Share assignment submissions on-time	(Mean)	0.66	0.64	0.58	0.72	0.67	0.68
	(Median)	0.71	0.67	0.57	0.83	0.75	0.75
	(Std dev)	0.32	0.33	0.32	0.31	0.32	0.32
On-time assignment submission available	(Mean)	33.0%	25.0%	27.4%	32.6%	40.1%	34.6%
Discussion post count	(Mean)	1.19	1.19	0.87	1.47	1.33	1.27
	(Median)	0.00	0.00	0.00	1.00	1.00	0.00
	(Std dev)	1.81	1.73	1.61	1.83	1.92	1.88
Discussion reply count	(Mean)	1.47	1.56	1.06	1.93	1.63	1.54
	(Median)	0.00	0.00	0.00	0.00	0.00	0.00
	(Std dev)	2.95	2.92	2.53	3.20	3.15	3.03
Average discussion post depth	(Mean)	1.47	1.49	1.46	1.51	1.47	1.46
	(Median)	1.50	1.57	1.50	1.57	1.50	1.50
	(Std dev)	0.36	0.35	0.36	0.37	0.36	0.35
Average discussion post length, in words	(Mean)	661	705	663	679	638	663
	(Median)	548	597	548	567	525	550
	(Std dev)	508	508	503	510	506	511
N		1,173,878	102,744	332,187	120,040	325,063	293,844
Panel B: course-specific samples							
		ENG 111	ENG 112	BIO 101	MTH 154	MTH 161	
		(7)	(8)	(9)	(10)	(11)	
Total time, in minutes	(Mean)	655	604	533	532	501	
	(Median)	477	444	330	321	254	
	(Std dev)	653	587	628	628	671	
Total click count	(Mean)	1013	948	996	680	665	
	(Median)	800	765	698	478	432	
	(Std dev)	920	860	1029	705	771	
Average session length, in minutes	(Mean)	13.30	13.47	11.02	12.02	11.69	
	(Median)	9.47	9.62	6.96	7.41	6.33	
	(Std dev)	13.80	13.87	12.94	14.27	15.51	
Standard deviation of session length	(Mean)	27.26	27.58	24.05	25.11	24.91	
	(Median)	22.46	22.26	18.98	18.89	17.47	
	(Std dev)	20.73	20.78	20.52	21.89	23.91	
Assignment Submission count	(Mean)	7.55	7.44	9.59	9.54	6.82	
	(Median)	5.00	5.00	5.00	4.00	3.00	
	(Std dev)	7.71	7.12	12.68	13.29	9.51	
Assignment submission available	(Mean)	66.0%	60.3%	66.5%	51.4%	53.1%	
Share assignment submissions on time	(Mean)	0.61	0.65	0.64	0.56	0.65	
	(Median)	0.64	0.67	0.67	0.50	0.67	
	(Std dev)	0.30	0.30	0.33	0.36	0.33	

(*Continued*)

Table 4. Continued.

		Full Sample	Summer 2019	Fall 2019	Summer 2020	Fall 2020	Spring 2021
		(1)	(2)	(3)	(4)	(5)	(6)
On-time assignment submission available	(Mean)	48.0%	48.7%	43.8%	25.2%	24.4%	
Discussion post count	(Mean)	1.83	2.10	0.53	0.50	0.36	
	(Median)	1.00	2.00	0.00	0.00	0.00	
	(Std dev)	2.30	2.11	1.32	1.00	0.90	
Discussion reply count	(Mean)	2.01	2.73	0.81	0.61	0.45	
	(Median)	0.00	1.00	0.00	0.00	0.00	
	(Std dev)	3.22	3.62	2.59	1.66	1.48	
Average discussion post depth	(Mean)	1.47	1.50	1.48	1.46	1.45	
	(Median)	1.50	1.59	1.60	1.50	1.50	
	(Std dev)	0.35	0.33	0.37	0.40	0.38	
Average discussion post length, in words	(Mean)	751	826	488	459	388	
	(Median)	648	740	406	375	317	
	(Std dev)	531	514	354	345	307	
N		54,232	31,457	38,806	25,175	19,981	

Notes: columns 1 and 7–11 contain student × course-section observations from both the training and validation sets. If an observation has no online sessions, then their values of "average session length, in minutes" and "standard deviation of session length" is set to missing. If a student has no discussion posts, then their "average discussion post depth" and "average discussion post length, in words" is set to missing. Discussion post depth is defined such that a value of 1 corresponds to the original post, a value of 2 corresponds to the first reply, and so forth.

In Figure 3 we present *C*-statistics and their standard errors for course-performance models trained separately on the sample of students enrolled in each of the five courses. Across all five courses, models that combine admin and LMS data achieve the highest levels of accuracy, and accuracy levels are generally high for the course-specific models.[25] Across four of the five courses (all except ENG111), we find that models trained on admin-only measures meaningfully outperform models trained on LMS-only data. In the case of ENG111, the model trained on LMS-only data does outperform the model trained on admin-only data (*C*-statistic of 0.81 compared to 0.78); this makes intuitive sense as a sizeable share of ENG111 students (56.2 percent of the training sample) are in their first term at VCCS.[26]

The differences in LMS-only models across courses in Figure 3 reflect the pattern from Figure 2, with *C*-statistics ranging from 0.81 for ENG111 to 0.70 for MTH161 for models using all LMS predictors. This finding is directly related to the values of the LMS predictors, which we provide for each of the five courses in columns (7) through (11) of Table 4. Overall, we see that the courses for which LMS predictors add the greatest value are those with the highest averages of LMS predictors.

[25]In Online Appendix Table A9, we show which groups of predictors contribute most to overall prediction accuracy within the course-specific performance prediction models. We again observe a very similar pattern to what we found with the prediction model trained on all courses.

[26]Online Appendix Figure A1 shows very similar patterns for TPR and TNR across the 15 course-specific models represented in Figure 3. Online Appendix Figure A2 shows the course-specific performance of the models trained on the full training sample for the five courses (i.e., the same metric described in Figure 2). These *C*-statistics are slightly higher (one percent or less) than the *C*-statistics from the course-specific models shown in Figure 3. In other words, the models trained on the full training set have very similar levels of accuracy to course-specific data, regardless of which type(s) of predictors are included.

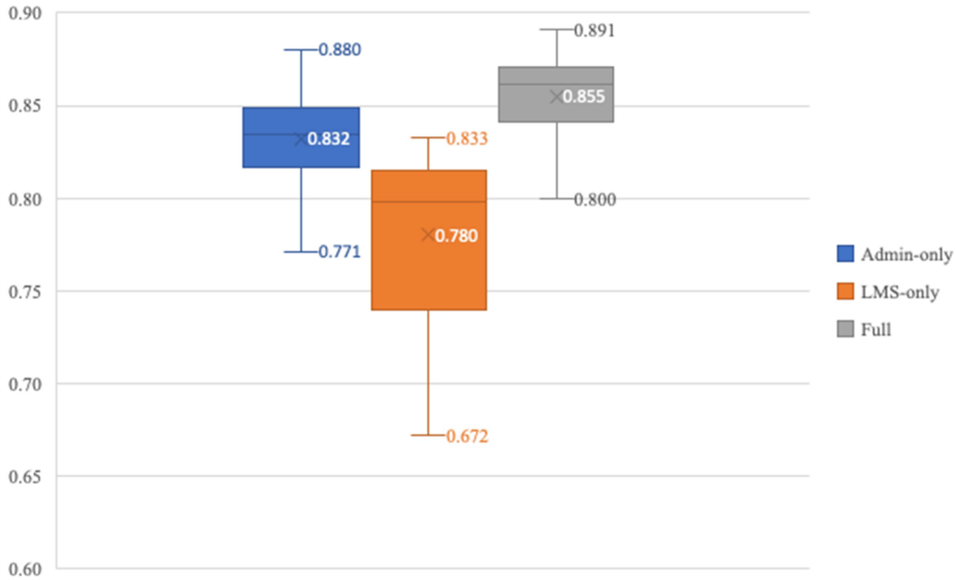

Figure 2. Distribution of *C*-statistics for course-specific validation samples, top 50 courses.
Notes: Top 50 courses determined by number of observations in the validation sample. For each course, we compute a separate *C*-statistic by applying the models trained on the full sample to only observations in the validation set corresponding to that course. The value labels indicate the minimum, mean, and maximum values of the distribution of c-statistics.

We see that while total time online is more similar across courses (ranging from 507 minutes for MTH161 to 641 minutes for ENG111), the two math courses have approximately one-third fewer click actions than the English and Biology courses. The starkest difference is average word count in the predictors describing discussion posts: the average number of discussion posts is roughly four times higher for the English versus math courses, and, within discussion posts submitted, those for the English courses are roughly twice as long as for the math courses. These results further support the intuitive hypothesis that the value-add of the LMS data is greatest when students and instructors engage more with the LMS data through the coursework.

It is worth noting that our analyses used a limited set of LMS predictors based on existing learning-analytics research. Because there are an infinite number of predictors one could construct from the LMS data, any prediction model will necessarily rely on a subset of these possible predictors; the choice of which predictors to generate and include could impact the results. However, we believe that our results reasonably reflect the overall predictive utility of LMS data. This is because both existing research we draw on and our large sample cover a broad range of instructional and institutional contexts and represent the majority of use cases of Canvas LMS. In addition, our selected predictors cover student behavior around different functionalities of the LMS. While there can be variants or more complicated forms of predictors, they might either be highly correlated with what we included or only available for a small number of courses or students, and therefore might not provide substantial marginal predictive value when adopted at the institutional level.

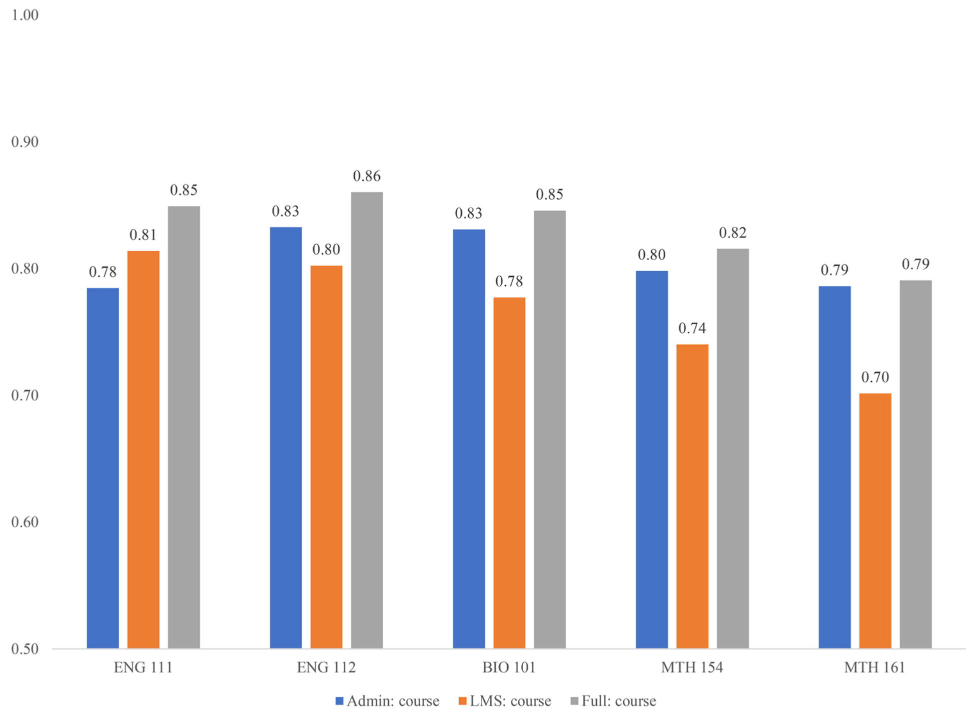

Figure 3. *C*-statistics for course-specific models, by predictor category.
Notes: each bar corresponds to a separate random forest prediction model using the set of predictors indicated by the color of the bar, and observations from the course indicated by the x-axis label. Standard errors of the *C*-statistics are in parentheses.

Discussion

As LMS software is becoming increasingly more prevalent in higher education—particularly in a post-COVID era characterized by flexibility of instruction modality—researchers and higher education institutions are increasingly interested in harnessing the LMS-generated data for various instructional and analytic purposes. However, making use of LMS data can be very costly in terms of personnel time, data storage, and computing power. For example, the VCCS LMS data for a single term is roughly one to two terabytes. Converting the raw data (which includes a row for each navigation or "click" a student makes within the LMS) into usable predictors requires expertise and a significant time investment. Particularly given limited resources at institutions like community colleges, it is important to understand the potential value of LMS data in predictive analytics.

In this article, we show how including LMS data improves the accuracy of models predicting course performance, relative to models using only admin data. We find that the accuracy gain from LMS data varies significantly across contexts, even within a community college system that uses the same LMS software across all courses and institutions. Specifically, LMS data add little value in predicting course performance for returning students. Including LMS predictors to the admin-only returning-student model increases the share of students with accurate predictions by 1.8%. However, in the case of new students, LMS-only data outperform admin-only data, and the combination of LMS and admin data has significantly higher accuracy compared with using only one data source.

Again, adding LMS predictors to the admin-only new-student model increases the share of students with accurate predictions by 8.6%. We also find significant variation across courses in the accuracy from and value-addition of LMS data. Intuitively, the prediction accuracy is highest for courses where students are more engaged within the LMS system, as observed by time spent logged in and contributing to discussion boards. Overall, our results suggest that colleges should be mindful of the types of courses for which they plan to use predictive analytics when deciding whether to invest in LMS data. Specifically, we suggest that LMS data add substantial predictive value and may be worth the investment for courses that (1) enroll many new students; (2) actively use LMS for instructional design; and (3) a significant share of students do not succeed in the course. The relatively poor accuracy of the admin-only data for new students that we find (*C*-statistic of 0.728, which we would classify as insufficient accuracy for implementation) suggests that if LMS data are not available for new students, then other data-collection efforts (e.g., incorporating high school transcripts) could substantially benefit predictive analytics in that setting. More broadly, our results demonstrate that researchers and educators should continue to critically investigate whether making use of these data results in meaningfully better models or accuracy than can be achieved with more traditional data sources and methods. Still, it is important to note that in this article we are solely focused on the predictive value-addition of LMS data in terms of increasing overall prediction accuracy. There are other potential benefits to incorporating the LMS data into prediction models. Specifically, incorporating LMS data into existing predictive models could decrease algorithmic bias (Yu et al., 2020). In a recent exploration of algorithmic bias in admin-only models, our results suggest that administrative predictors are less useful at predicting Black student outcomes compared with White students, which suggests that including additional data sources has the potential to mitigate bias (Bird et al., 2024). In future work, we will test this point explicitly within the VCCS context. For researchers or administrators interested in learning more specifically about how we work with the LMS and admin data to construct predictors, and how we build the predictive models described in this article, we have made our codebase public at https://github.com/nudge4/admin_vs_lms_data_public.

Open Research Statements

Study and Analysis Plan Registration

There is no study and analysis plan registration associated with this manuscript.

Data, Code, and Materials Transparency

The code used to generate the results reported in this manuscript are available on GitHub: https://github.com/nudge4/admin_vs_lms_data_public. The data and materials underlying the results reported in this manuscript are not openly available.

Design and Analysis Reporting Guidelines

Not applicable.

Transparency Declaration

The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Replication Statement

This manuscript reports an original study.

Acknowledgements

We are very grateful to Dr. Cat Finnegan and the team at the Virginia Community College System for their collaboration on this project. We are grateful to our colleagues at VCCS and University of California–Irvine for their assistance obtaining the LMS data. We appreciate thoughtful feedback from the journal editor and two anonymous referees.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

References

Achen, A. C., & Courant, P. N. (2009). What are grades made of? *The Journal of Economic Perspectives*, *23*(3), 77–92. https://doi.org/10.1257/jep.23.3.77

Aguiar, E., N. V., Chawla, Jay Brockman, G., Alex Ambrose, & V., Goodrich. (2014). *Engagement vs performance: Using electronic portfolios to predict first semester engineering student retention* [Paper presentation]. LAK '14: Proceedings of the Fourth International Conference on Learning Analytics and Knowledge, March 2014 (pp. 103–112). https://doi.org/10.1145/2567574.2567583

Arizmendi, C. J., Bernacki, M. L., Raković, M., Plumley, R. D., Urban, C. J., Panter, A. T., Greene, J. A., & Gates, K. M. (2022). Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work. *Behavior Research Methods*, *55*(6), 3026–3054. https://doi.org/10.3758/s13428-022-01939-9

Attewell, P., Heil, S., & Reisel, L. (2012). What is academic momentum? And does it matter? *Educational Evaluation and Policy Analysis*, *34*(1), 27–44. https://doi.org/10.3102/0162373711421958

Bailey, T., Alfonso, M., Calcagno, J. C., Jenkins, D., Kienzl, G., & Leinbach, T. (2004). *Improving student attainment in community colleges: Institutional characteristics and policies*. Community College Research Center. http://files.eric.ed.gov/fulltext/ED484346.pdf.

Baker, R., Xu, D., Park, J., Yu, R., Li, Q., Cung, B., Fischer, C., Rodriguez, F., Warschauer, M., & Smyth, P. (2020). The benefits and caveats of using clickstream data to understand student self-regulatory behaviors: Opening the black box of learning processes. *International Journal of Educational Technology in Higher Education*, *17*(13), 1–24. https://doi.org/10.1186/s41239-020-00187-1

Barbosa, G., Camelo, R., Cavalcanti, A. P., Miranda, P., Mello, R. F., Kovanović, V., & Gašević, D. (2020). *Towards automatic cross-language classification of cognitive presence in online discussions* [Paper presentation]. Proceedings of the 10th International Conference on Learning Analytics & Knowledge (LAK '20) (pp. 605–614). https://doi.org/10.1145/3375462.3375496

Barshay, J., & Aslanian, S. (2019). *Under a watchful eye: Colleges are using big data to track students in an effort to boost graduation rates, but it comes at a cost* (APM Reports). https://www.apmreports.org/story/2019/08/06/college-data-tracking-students-graduation

Bedard, K., & Kuhn, P. (2008). Where class size really matters: Class size and student ratings of instructor effectiveness. *Economics of Education Review*, *27*(3), 253–265. https://doi.org/10.1016/j.econedurev.2006.08.007

Bettinger, E. P., & Long, B. T. (2010). Does cheaper mean better? The impact of using adjunct instructors on student outcomes. *Review of Economics and Statistics*, *92*(3), 598–613. https://doi.org/10.1162/REST_a_00014

Bird, K. A., Castleman, B. L., Mabel, Z., & Song, Y. (2021). Bringing transparency to predictive analytics: A systematic comparison of predictive modeling methods in higher education. *AERA Open*, *7*, 233285842110376. https://doi.org/10.1177/23328584211037630

Bird, K. A., Castleman, B. L., & Song, Y. (2024). Are algorithms biased in education? Exploring racial bias in predicting community college student success. *Journal of Policy Analysis and Management*, 1–24. https://doi.org/10.1002/pam.22569

Boatman, A., & Long, B. T. (2018). Does remediation work for all students? How the effects of postsecondary remedial and developmental courses vary by level of academic preparation. *Educational Evaluation and Policy Analysis*, *40*(1), 29–58. https://doi.org/10.3102/0162373717715708

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3.1. Statistics Department University of California Berkeley, CA, USA 1 58.

CCRC. (2021). *An introduction to community colleges and their students.* Policy Fact Sheet, July 2021. https://ccrc.tc.columbia.edu/media/k2/attachments/introduction-community-colleges-students.pdf.

Choi, H., Lee, J. E., Hong, W. J., Lee, K., Recker, M., & Walker, A. (2016). *Exploring learning management system interaction data: Combining data-driven and theory-driven approaches.* International Educational Data Mining Society.

Cicchinelli, A., Veas, E., Pardo, A., Pammer-Schindler, V., Fessl, A., Barreiros, C., & Lindstädt, S. (2018). *Finding traces of self-regulated learning in activity streams* [Paper presentation]. In Proceedings of the 8th international conference on learning analytics and knowledge (pp. 191–200). https://doi.org/10.1145/3170358.3170381

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies*, *10*(1), 17–29. https://doi.org/10.1109/TLT.2016.2616312

Crossley, S., Paquette, L., Dascalu, M., McNamara, D., & Baker, R. (2016). *Combining click-stream data with NLP tools to better understand MOOC completion* [Paper presentation]. LAK '16: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, April 2016 (pp. 6–14). https://doi.org/10.1145/2883851.2883931

Crosta, P. M. (2014). Intensity and attachment: How the chaotic enrollment patterns of community college students relate to educational outcomes. *Community College Review*, *42*(2), 118–142. https://doi.org/10.1177/0091552113518233

Cuseo, J. (2007). The empirical case against large class size: Adverse effects on the teaching, learning, and retention of first-year students. *Journal of Faculty Development*, *21*(1), 5–21.

Ekowo, M., & Palmer, I. (2016, October). T*he promise and peril of predictive analytics in higher education: A landscape analysis* (New America Policy Paper). https://www.newamerica.org/education-policy/policy-papers/promise-and-peril-predictive-analytics-higher-education/

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education*, *44*(1), 130–160. https://doi.org/10.3102/0091732X20903304

Gardner, J., & Brooks, C. (2018). Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction*, *28*(2), 127–203. https://doi.org/10.1007/s11257-018-9203-z

Gašević, 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. https://doi.org/10.1016/j.iheduc.2015.10.002

Gayed, J. M., Carlon, M. K. J., Oriola, A. M., & Cross, J. S. (2022). Exploring an AI-based writing Assistant's impact on English language learners. *Computers and Education: Artificial Intelligence*, *3*, 100055. https://doi.org/10.1016/j.caeai.2022.100055

Ghojogh, B., & Crowley, M. (2019). The theory behind overfitting, cross validation, regularization, bagging, and boosting: Tutorial. https://arxiv.org/abs/1905.12787:arXiv

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Heo, J., Lim, H., Yun, S. B., Ju, S., Park, S., & Lee, R. (2019). *Descriptive and predictive modeling of student achievement, satisfaction, and mental health for data-driven smart connected campus life service* [Paper presentation]. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge (pp. 531–538). https://doi.org/10.1145/3303772.3303792

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons, Inc.

Klempin, S. C., Grant, M., & Ramos, M. (2018). *Practitioner perspectives on the use of predictive analytics in targeted advising for college students* (CCRC Working Paper No. 103). https://ccrc.tc.columbia.edu/publications/practitioner-perspectives-predictive-analytics-targeted-advising.html

Kofoed, M. S., Gebhart, L., Gilmore, D., & Moschitto, R. (Forthcoming). Zooming to class?: Experimental evidence on college students' online learning during COVID-19. *American Economic Review: Insights*.

Kuh, G. D., Kinzie, J., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2007). Piecing together the student success puzzle: Research, propositions, and recommendations. *ASHE Higher Education Report*, *32*(5), 1–182. https://doi.org/10.1002/aehe.3205

Li, Q., Baker, R., & Warschauer, M. (2020). Using clickstream data to measure, understand, and support self-regulated learning in online courses. *The Internet and Higher Education*, *45*, 100727. https://doi.org/10.1016/j.iheduc.2020.100727

Lim, J. M. (2016). Predicting successful completion using student delay indicators in undergraduate self-paced online courses. *Distance Education*, *37*(3), 317–332. https://doi.org/10.1080/01587919.2016.1233050

Long, B. T. (2010). Higher education finance and accountability. In Kevin Carey and Mark Schneider (Eds.). *Accountability in American higher education*. Palgrave Macmillan and American Enterprise Institute.

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

McCuaig, J., & Baldwin, J. (2012). *Identifying successful learners from interaction behaviour*. International Educational Data Mining Society.

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*(2), 499–515. https://doi.org/10.3758/s13428-012-0258-1

Motz, B., Quick, J., Schroeder, N., Zook, J., & Gunkel, M. (2019). *The validity and utility of activity logs as a measure of student engagement* [Paper presentation]. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge (pp. 300–309). https://doi.org/10.1145/3303772.3303789

Murphy, R., Roschelle, J., Feng, M., & Mason, C. A. (2020). Investigating efficacy, moderators and mediators for an online mathematics homework intervention. *Journal of Research on Educational Effectiveness*, *13*(2), 235–270. https://doi.org/10.1080/19345747.2019.1710885

National Student Clearinghouse. (2021). COVID-19: Stay informed with the latest enrollment information, November 18, 2021. Retrieved June 13, 2022, from https://nscresearchcenter.org/stay-informed/

Page, L. C., & Gehlbach, H. (2017). How an artificially intelligent virtual assistant helps students navigate the road to college. *AERA Open*, *3*(4), 233285841774922. https://doi.org/10.1177/2332858417749220

Park, J., Yu, R., Rodriguez, F., Baker, R., Smyth, P., & Warschauer, M. (2018). Understanding student procrastination via mixture models. *Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018)* (pp. 187–197).

Pascarella, E. T., & Terenzini, P. T. (1991). *How college affects students: Findings and insights from twenty years of research*. Jossey-Bass.

Pascarella, E., & Terenzini, P. (2005). *How college affects students: A third decade of research*. Jossey-Bass.

Sher, V., Hatala, M., & Gašević, D. (2019). On multi-device use: Using technological modality profiles to explain differences in students' learning. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK '19)* (pp. 1–10).

Spelmen, V. S., & Porkodi, R. (2018). *A review on handling imbalanced data* [Paper presentation]. 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, India (pp. 1–11). https://doi.org/10.1109/ICCTCT.2018.8551020

Tinto, V. (1994). Discovering the sources of student success. In D. Floyd (Ed.), *From vision to reality: Student affairs agenda in the '90s*. American College Testing Program.

Tinto, V. (2012). Enhancing student success: Taking the classroom success seriously. *The International Journal of the First Year in Higher Education*, *3*(1), 1–8. https://doi.org/10.5204/intjfyhe.v2i1.119

Wang, Q., & Mousavi, A. (2022). Which log variables significantly predict academic achievement? A systematic review and meta-analysis. *British Journal of Educational Technology*, *54*(1), 142–191. https://doi.org/10.1111/bjet.13282

Yu, R., Li, Q., Fischer, C., Doroudi, S., & Xu, D. (2020). Towards accurate and fair prediction of college success: Evaluating different sources of student data. In: Proceedings of the 13th International Conference on Educational Data Mining (A. N. Rafferty, J. Whitehill, V. Cavalli-Sforza, & C. Romero (Eds.), pp. 292–301).

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. https://doi.org/10.1016/j.iheduc.2015.05.002

Appendix A: Canvas Data and Generation of LMS predictors

The Canvas Data service organizes information across 95 tables at the time of this study, following typical data-warehouse conventions.[27] Each table tracks a specific aspect of user activity, and usually a few relevant tables need to be joined together to capture the full picture of even one type of activity (e.g., assignment submissions need to be joined with the assignment information). The most granular information about student activities comes from one table (requests) which includes click action records (a.k.a., clickstream data) (see Online Appendix Figure A3 for a snapshot and Baker et al. (2020) for a comprehensive introduction). While different courses may be organized and designed differently, the data they generate all fit into the same schema of 95 tables.

We purchased a read-only Amazon Redshift data-warehouse instance where the 95 tables are hosted and ran a series of SQL queries to aggregate raw tables by area of activity (e.g., click actions, discussion posts, assignment submissions). After transferring the aggregated tables to a high-performance computing (HPC) environment where the administrative data were hosted, we identified the full course period and the span of "early term" for each course from the administrative data. Finally, we ran a series of Python codes to compute the 50 measures (predictors) according to the measure definitions and specified time spans.

The organization of 95 tables in Canvas Data is identical across institutions that adopt the system, so the processes described above are largely applicable in different institutional contexts. However, there are some potential obstacles that institutions may face when replicating this work. For example, the cost of technical infrastructure to store and process the gigantic raw data can be unaffordable, especially for low-resourced institutions. Also, the mechanism of connecting LMS and administrative data varies across institutions and can be complicated. Specifically, the organization of courses on the administrative side and the LMS side might be misaligned. To accurately figure out this connection usually requires different administrative offices, such as registrar and IT, to coordinate, which adds to the logistic costs.

[27]The documentation of these tables can be found at https://portal.inshosteddata.com/docs.

Appendix B: comparison of predication accuracy for online versus in-person observations

While all VCCS courses can use Canvas's LMS features, online courses typically require more LMS interaction with the student.[28] We show the number of online versus in-person observations in Panel A of Online Appendix Table A9. The majority (73.8%) of the student-by-course section observations in our analytic sample are online, which is driven in some part by the inclusion of Fall 2020 and Spring 2021, during which most coursework was still online due to the COVID pandemic, in our analytic sample. Indeed, 94.4 %of observations in the validation set, which consists entirely of Spring 2021 observations, are online. Online enrollment in the validation set is over 99% for ENG 111, ENG 112, and BIO 101.

Panel B of Online Appendix Table A6 shows that for most (but not all) of the early target term LMS predictors, the online observations have considerably higher mean values. For example, the average total minutes spent logged in was 655 minutes for online observations and 279 for in-person observations. However, assignment submission data is available for more in-person observations (57.3 percent) compared to online (49.2 percent).

Given these differences, we explore whether the added value of LMS predictors differs for online versus in-person observations. To do so, we calculate separate *C*-statistics online versus in-person subsets of the validation sample. We present these results in Online Appendix Table A10. The *C*-statistics for the online observations closely mirrors the results in Figure 1. However, we observe a significant drop in the *C*-statistic for the LMS-only model for the in-person observations, equal to 0.647 for the new-student sample and 0.708 for the returning-student sample. Interestingly, the in-person *C*-statistic is higher for admin-only models and is only slightly lower for the full predictor models (compared to Figure 1). These results suggest that LMS-only models are of significantly less value for in-person observations; however, given that the validation sample from Spring 2021 contains only 5.6% in-person observations, we caution against drawing strong conclusions from this particular comparison.

Because the training set contains a significantly larger share of in-person observations (31.5% for returning-student sample and 37.2% for new-student sample), and because the computation of feature importance scores are not reliant on the validation sample, we build modality-specific models with the full set of predictors and compare the feature importance scores in Online Appendix Table A12. We find that the LMS predictors have higher feature importance for the online observations compared with the in-person observations. Comparing Panels A and B, which show the top 30 predictors for the modality-specific models using the returning-student sample, respectively, we see that there are four LMS predictors in the top 10 predictors for online observations, but only two LMS predictors in the top 10 for in-person observations. Similarly, the top-rated LMS predictor has a ranking of two (i.e., second most important feature) for online observations, but a ranking of seven for in-person observations. We find similar patterns when comparing Panels C and D, which show the same set of results using the new-student sample.

[28]We classify all hybrid courses, which VCCS defines as having 50–99 % of course instruction occurring online, as online courses.