


A learning analytics case study: On class sizes in undergraduate writing courses

Richard A. Levine^{1,2}  | Patricia E. Rivera¹ | Lingjun He^{2,3} | Juanjuan Fan^{1,2} | Marilee J. Bresciani Ludvick^{2,4}

¹Department of Mathematics and Statistics, San Diego State University, San Diego, California, 92182, USA

²Analytic Studies and Institutional Research, San Diego State University, San Diego, California, 92182, USA

³Machine Learning Staff, Affirm, Inc., San Diego, California, 92182, USA

⁴Institutional Effectiveness, Loyola University Chicago, Chicago, Illinois, 60660, USA

Correspondence

Richard A. Levine, Department of Mathematics and Statistics, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA.

Email: rlevine@sdsu.edu

Funding information

NSF, Grant/Award Number: 1633130

With the collection and availability of data on student academic performance and academic background, higher education institutions have recently stepped up initiatives in and infrastructure for learning analytics, leveraging this deluge of data to inform student success. With definitions of student success varying from analyses of what predicts levels of specific career readiness competencies to degree completion, the environment is a fertile ground for statistical practice and collaboration among a statistically savvy yet diverse clientele of instructors, programme advisors and administrators. In this paper, we discuss our experiences to this end through a consulting project evaluating the impact of writing course class size on students achieving a graduation writing requirement. In detailing the workflow for and challenges in this project, we share aspects of statistical communication and reporting, applications of innovative statistical methodology developed by our research group for handling confounding factors and correlated inputs and training through an interdisciplinary applied institutional research professional development programme. This paper illustrates how instilling an appreciation for statistical inference through each of these components is invaluable for capturing institutional buy-in for data-informed decision-making in general statistical practice.

KEYWORDS

data-informed decision-making, educational data mining, observational study, random forest, student success, writing placement assessment

1 | INTRODUCTION

The terms “learning analytics” and “educational data mining” have entered the mainstream of studies designed to accurately predict student success, often defined by persistence, degree completion, and cumulative grade point average (GPA). A number of universities have been positioning themselves to take advantage of the ever-increasing collection of student academic and demographic data for instructors, programme directors, advisors and administrators to retain students, facilitate attainment of course and programme learning outcomes and help students succeed in programmes of study towards a successful graduation outcome (Romero & Ventura, 2020). Crick Deakin et al. (2017) and Macfadyen et al. (2014) take a complex systems perspective, describing how to incorporate learning analytics into existing processes and arguing that more institutions are engaging in this kind of “systems” improvement. That said, Gasevic et al. (2019) states that “the vast majority of institutions are yet to exploit the full use of learning and organizational data to address institutional and educational challenges.”

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Stat* published by John Wiley & Sons Ltd.

Tsai et al. (2018) argues that a collaborative team of “data experts ... and users” are needed “in order that learning analytics may have positive impact on informing decisions and changing behavior. This has been identified as a common gap between needs and solutions in institutional analytics capacity.” Along these lines, Piety et al. (2014) proposes a field of “Educational Data Sciences” ... “will benefit both those producing and consuming information from these practices as well as those developing education programs aimed at building the human capital necessary to work with educational data.” Both articles emphasize that such collaborative efforts are not a new endeavour, but engaging campus stakeholders is extremely challenging.

In recent years, San Diego State University (SDSU) has established a collaborative team of student success leaders to work with and learn from data scientists and statisticians to improve data-informed decision-making on campus. This infrastructure includes establishing a data warehouse and data governance plan, student success dashboards, the roles of key academic and student affairs units and administrative leadership and cross-college collaborations relative to student success. A focal point of these efforts is a data science team in the institutional research office called the Statistical Modeling Group, collaborating on campus needs in statistical training, statistical communication and reporting, data curation and analytics with an eye on actionable outcomes.

In this paper, we illustrate this collaborative effort in a higher education setting through a specific student success project. The SDSU Academic Senate convened and charged a Class Size Task Force with studying the impact of class size on success in first- and second-year composition courses that satisfy general education (GE) composition requirements. A previous task force on the subject made a recommendation in 2014 to “reduce the size of classes fulfilling the Composition and Intermediate Composition and Critical Thinking General Education Foundations requirements from 30 to 18.” However, this recommendation was largely ignored as being too costly and critiqued as anecdotal since the definition of student success remained elusive.

Regardless of how student success is defined and measured in writing courses, concern around class size in writing courses is growing. For instance, the 2020 Association of Departments of English through the Modern Language Association states Associate of Departments of English (2020) “... No more than 20 students should be permitted in any writing course. ... No writing instructor should teach more than sixty students in one term.” This recommendation continued the Statement of Principles for Postsecondary Teaching by the 2015 National Council of Teachers of English. A similar statement by UT Austin in 2020 (UT Austin Undergraduate Studies, 2020) states that “... departments staff Writing Flag classes at no more than a 25:1 student-instructor ratio.” The SDSU Academic Senate recognized a need to design a study to analyse these issues given seemingly conflicting opinions of key stakeholders. The university administration expressed concerns about the rather large resources required to implement even a cap of 25 students on writing courses and allocated larger classrooms. Rhetoric & Writing Studies (RWS) Department faculty expressed concerns that optimizing student success in such ad hoc class size selection is hard. While divergent view points existed, published data on the class size in writing, as it relates to student success, was nascent.

From a statistical practice viewpoint, we will leverage this surprisingly rich problem to exposit the cross-campus, interdisciplinary collaborations in this project, introduce extensions and application of statistical learning innovations our research team previously developed and briefly delve into a Data Champions programme we spearheaded to resolve similarly confusing discussion surrounding data-informed student success requests on campus. Our methods, in particular, must control for confounding and multicollinearity present in such educational data sets. The data used for the statistical analysis presented herein are derived from a broader Data Champions programme data set. Data collection for this specific project was not onerous, entailing a subsetting of the Data Champions programme data. However, confirming the demand for human resources in learning analytics projects discussed in Tsai et al. (2018), curation of and documentation for the Data Champions data set was a monumental task led by a full-time data engineer in our Institutional Research office in collaboration with the campus information technology team, subject-specific learning experts and programme data science and institutional research experts (including all the authors on this paper). While a significant upfront cost in data set curation, the benefit is gained by its continued use for campus student success studies.

In Sections 2 and 3, we will focus on the RWS problem at hand laying out the collaborative process and challenges therein. In Section 4, we will detail recommendations the Task Force made to design a follow-up prospective study of class size in writing courses. We will also come back to the broader role of data science training and practice, developing a data-informed decision-making infrastructure for scientific problem solving not just limited to higher education institutions.

2 | DATA AND METHODS

This project came at a fortuitous time in the evolution of a data-informed decision-making culture on our campus. In a joint effort among the offices of institutional research, of institutional effectiveness, of instructional technology services (ITS) and of student affairs research and assessment, in collaboration with Statistics faculty and graduate students, the university offered an interdisciplinary applied institutional research professional development programme called Data Champions, started in 2017. This year-long programme trains faculty, staff and administrators on data resources; various student development measures known to correlate with and predict student success, such as sense of belonging, academic self-efficacy and well-being; study design; statistical analysis and effective statistical communication for student success studies. Particular to this project, a shared data set was curated for Data Champions' projects, and campus student success leaders gained appreciation for challenges

in observational study design and analysis inherent in student success decision-making. At the same time, the campus was vetting EAB software, used to gather academic advising and engagement data for predictive analytics use. The company EAB (name derived from the original moniker the Education Advisory Board) was open about the machine learning engine underlying its data processing, analyses and reporting. Our research group also had published a number of student success studies leaning heavily on propensity score methods and random forest machinery to evaluate a variety of pedagogical reforms and a Supplemental Instruction near-peer tutoring service, for example, Guarcello et al. (2017) and Pelaez et al. (2019). Our ITS unit in particular made a strong push for growing the learning analytics infrastructure on campus by popularizing these papers, the Data Champions programme training and educational data mining research in the community. Consequently, we had a relatively clean data set from which to work and campus buy-in for the analytics approach, using machine learning tools for quasi-experimental study data, we proposed.

Of particular importance for the study herein, the campus devoted resources for a full-time data engineer and released the time of information technology, institutional research and data science experts on campus to curate a common data set for the Data Champions programme. This challenging and time-consuming task entailed selecting features desired, merging queries on the student information database with data silos in student success centres and individual departments around campus, organizing these diverse data sets into identical temporal structure (by week, by semester or by year as needed), identifying data anomalies and inconsistencies particularly for supposedly identical elements among the data sets, automating semesterly updates of the data set and composing a detailed data documentation. As mentioned in the introduction, though a high initial expense in data curation, the result is a reliable data set for use by campus student success leaders. In the case of this class size study, we have a comparatively more reliable and complete data set than previous studies on the topic.

The official client for this project was the Academic Senate. The Senate convened a Class Size Task Force with the last author as Chair, the first author as Statistics expert, a RWS faculty member and an Associate Dean from the College of Arts & Letters (CAL) in which the RWS Department is housed. The Task Force was tasked with examining the impact of class size on student success as defined by writing comprehension/writing skill ability. The Task Force met on average every other week and the Chair regularly reported progress back to the Senate executive committee and the Provost. During this period, the Task Force, through the RWS faculty representative, queried the RWS Department Chair and, when needed, faculty. The Task Force also consulted the previous task force, chaired by a CAL Women's Studies faculty member and consisting of, to our recollection, at least 10 faculty members representing a diverse set of departments from every college on campus. While the level of statistical training varied among these groups, all individuals had training in statistical design and analysis at the level of regression inferences and all had performed statistical studies for research publications in their respective fields.

These collaborations provided critical insight on writing course offerings and pedagogy, a writing placement assessment (WPA) exam offering and scoring and desired learning outcomes for writing success the RWS faculty derived from the literature and the higher education writing community. As our analysis matured and findings came into focus, the Class Size Task Force provided executive summaries (no more than two pages) to the RWS Department, Senate executive committee and the Provost for feedback and any further direction. These executive summaries typically included a statement of problem, a brief data description, a very brief (typically two sentences) layman's summary of statistical methods, a single graphic describing inferences drawn and a paragraph summarizing key takeaways. Section 3 is really a compendium of these executive summaries and thus the reader can get a taste for the analyses performed and summaries presented there. We include a representative example of a progress report as supplementary information.

We emphasize that the clients consisted of an administrative stakeholder concerned with allocating precious campus resources in very short supply, the faculty Senate with the bird's-eye view of student success in degree programmes broadly and the RWS Department of subject-specific experts passionate about student training in writing and discourse. The administration thus moved for larger writing class sizes and the RWS Department presented literature and arguments for small writing class sizes. Though a diverse clientele, all parties appreciated and respected the challenges of optimizing student success at a large, diverse, public institution. As an observational study with changing student population, course offerings and degree programme goals, there were two primary statistical challenges: study design and statistical communication.

In collaboration with RWS Department expertise, the first statistical challenge entailed narrowing in on a study frame, namely, time period, student cohorts and writing courses and a writing faculty accepted measure of student success for analysis. For example, while defining writing competency with a writing course grade initially seemed logical and easy, the team soon discovered understandable faculty resistance of doing so. As class sizes increased, writing assignments and feedback on writing had to be reduced. Such significant changes to course syllabi in the form of specific assignments and grading schemes are not captured in typical learning analytics and would have been lost on data scientists had a collaborative team not been engaged in this process. We also found that there was limited data the writing faculty believed could speak to writing competency. After extensive dialogue, an upper level writing (UW) placement test score was chosen. We shall detail this issue more when discussing future directions later in the paper.

The second statistical challenge was reporting of findings to this really three-headed client. On the one hand, the clients had limited time, nor could we expect them to study detailed statistical visualizations, analyses and inferences. On the other hand, as statistically savvy clients within its numbers displaying contrary views on appropriate writing class sizes, we needed to successfully communicate rather subtle inferences, in terms of assumptions and methodology, in visualizations and brief reports that were quickly understood.

In the remainder of this Section 2 and the next Section 3, we detail the data, methods and results of this project. We will highlight areas where statistical challenges arose in the collaborative process. We emphasize that the analytics pipeline came about from extended, but excellent

conversations and feedback from the clients, helping to shape the work flow and ultimately the statistical reporting. In Section 4, we will present a higher level view of this collaborative process for learning analytics training and projects more broadly in higher education and beyond.

2.1 | Outcome: Graduation writing assessment requirement and the WPA

Students entering the university must demonstrate writing competency by achieving a specified threshold on a standardized exam (SAT, ACT, English Placement Test [EPT], English Language Arts California Standards Test, AP Language and Composition or Literature and Composition tests, College Board English Composition test and International Baccalaureate English exam) or earning a C or better in an approved remedial writing course. As part of the GE communication and critical thinking component for graduation, students must complete one first-year course in composition and one second-year course in intermediate composition and critical thinking. There is a list of courses across campus, including but not limited to courses in the RWS Department, which satisfy these two requirements. In the remainder, we will call the first of this sequence of two courses as lower level writing (LW) and the second as UW.

Students must take the LW and UW courses as part of their first 60 units at the university. The semester of or the immediate semester after completing 60 units, students must take the WPA offered by the RWS Department. The WPA asks students to read an argument and then write an essay that responds to questions related to the argument. The WPA is a two hour exam.

The WPA is scored out of 10. The graduation writing assessment requirement (GWAR) entails three levels.

- Score of 10: GWAR is satisfied, no additional writing courses required unless mandated by a student's major.
- Score of 8 or 9: student must take and earn a C or better in one of a list of approved upper division writing courses.
- Score of 7 or below: students must take and earn a C or better in RWS 280: Academic Reading and Writing and in one of a list of approved upper division writing courses.

We note that RWS 280 and the upper division writing courses used to satisfy the GWAR for a score of 9 and below are tailored specifically to students earning these scores and are not among the LW nor UW courses taken prior to sitting for the WPA. The WPA is scored by two trained readers blinded to the student and each other's score. When the two readers have a score disagreement, a third independent, trained reader reviews and assigns a score to that essay.

From 2007, the university instituted registration blocks to ensure the WPA is taken after 60 units, strictly enforcing that requirement and removing all but a single repeat. During this period, 81% of students scored an 8 or above on the WPA (standard deviation over years of 9 percentage points) and 19% of students scored a 10 on the WPA (standard deviation over years of 3 percentage points). As the RWS faculty shared with the Task Force, "the percentage of juniors writing at a level considered appropriate for graduation has had little fluctuation during this period. This limited fluctuation is not surprising given that juniors taking the exam have completed only two years of their coursework and need additional writing experience to be able to write at a graduation-appropriate level." The RWS Department has inferred that the WPA is serving the role in which it was designed to achieve and considers an 8 or above as identifying a student writing at a level appropriate to their class standing.

We will thus consider the period since 2007, a data set of 22,162 students. In Section 4, we will delve further into alternatives to and implications of using the WPA as an indicator of student success. Given that the WPA is the measure chosen by the university to determine whether the GWAR is met, the goal of the analysis herein is to ascertain the influence of class size on student success as defined by the WPA. Interestingly, the WPA serves as both a placement exam (intended to place students into a second sequence writing course) and the evaluation method intended to satisfy GWAR. We will consider two outcomes: binary WPA pass/fail (scoring an 8 or above; 7 or below) and trinary WPA high/medium/low (scoring 10; 8 or 9; 7 or below). Figure 1 presents a pie chart of the latter categorization. A summary of the WPA scores is included in Table 1.

2.2 | Inputs: Class size and covariates

Figure 2 presents histograms of the class sizes over the years of study. Placing into the context of the thresholds presented in Section 1, for LW courses, 36% of the classes are 30 or more students, 17% of the classes are below 25 and 3% of the classes are below 20. For UW courses, 38% of the classes are 30 or more students, 14% of the classes are below 25 and 2% of the classes are below 20. Though the university administration desired larger writing class sizes, for the most part they limited those classes to at most 35 students. That said, on the extreme end, there are LW and UW classes with over 40 and even over 50 students.

As an initial exploration on the relationship between WPA performance and class size, Figure 1 presents the distribution of high, medium and low WPA scores by class size groupings. There is not a strong apparent relationship, though the large class sizes show a higher proportion of students passing the WPA (high and medium scores), the small class sizes show a higher proportion of students failing the WPA (low score) and the

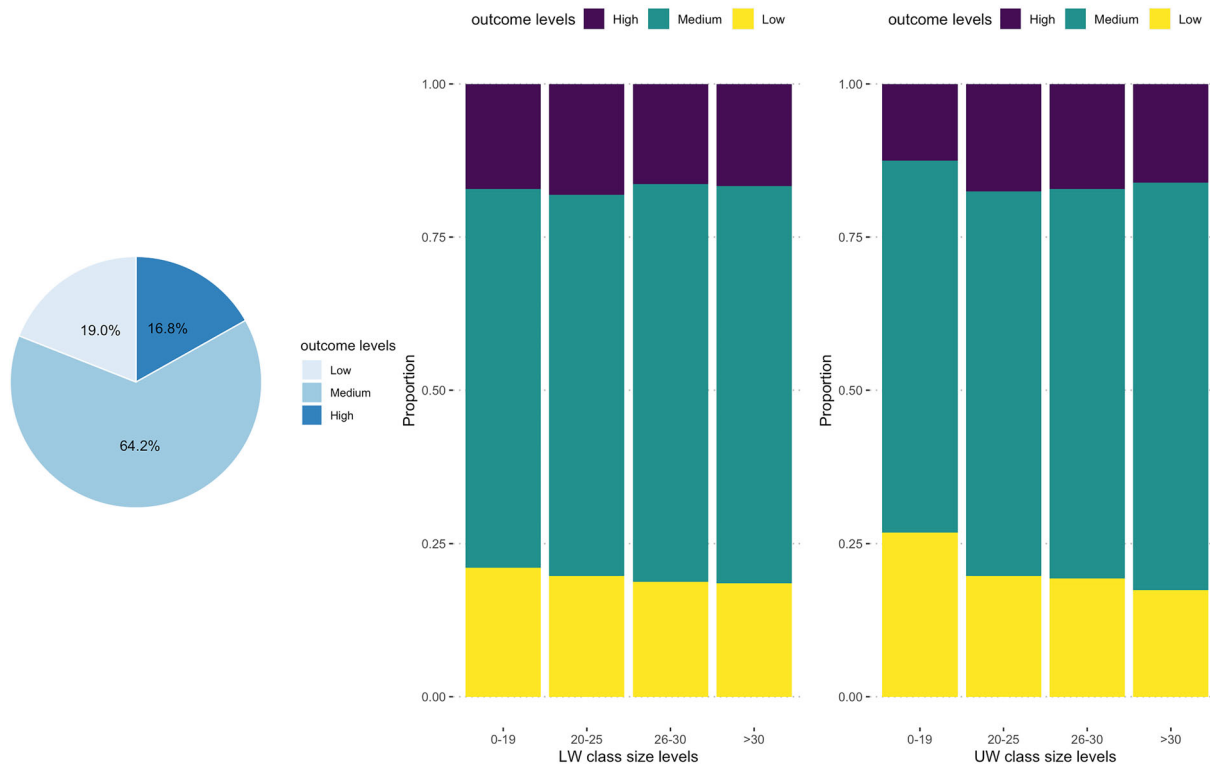


FIGURE 1 Exploratory data analysis on the response variable writing performance assessment (WPA) performance. The left graphic presents the proportion of students scoring a 10 (high), 8 or 9 (medium) and 7 or below (low) on the WPA. The middle graphic presents the proportion of students scoring a 10 (high), 8 or 9 (medium) and 7 or below (low) on the WPA by lower level writing (LW) course class size groupings (# of students in the writing course). The right graphic presents the proportion of students scoring a 10 (high), 8 or 9 (medium) and 7 or below (low) on the WPA by upper level writing (UW) course class size groupings (# of students in the writing course)

TABLE 1 Summary statistics for the continuous variables in the data set

Variable code	Min	Max	Mean	Median	SD	Description
WPA score	2	10	7.9	8	1.3	Score obtained on the writing placement assessment
stu_term_units	1	24	13.9	15	2.4	Total units enrolled in semester
age_entry	16	28	17.9	18	0.4	Age in years when student started at SDSU
satcomp	450	1520	1050	1050	138	SAT composite score
hsgpa	0	4.50	3.54	3.56	0.32	High school GPA
trangpa	0.23	4.00	3.29	3.35	0.70	GPA from transfer units
elmscore	0	80	21.8	0	25.1	Entry-level Mathematics test score
tran_units	4	63	48	48	12	Total transfer units accepted
LW.gradepoint	0	4	3.11	3	0.65	Gradepoint student received in lower level writing class
LW.course.size	4	69	27	28	3	Lower level writing class size
UW.gradepoint	0	4	3.10	3	0.67	Gradepoint student received in upper level writing class
UW.course.size	5	68	28	28	3	Upper level writing class size

Note: The output is the WPA score. The other variables are inputs.

moderate size classes perhaps show an increasing trend in poorer performance as the number of students increases. A mixed message bears out in univariate analyses. For example, logistic and multinomial regressions indicate that for each additional student in a LW or UW course, students are 1.02 times more likely to pass the WPA, students are twice (LW) and 1.5 times (UW) as likely to score at the medium level than the low level; but for a reduction of one student in the class, students are 1.43 times (LW) and twice as (UW) likely to score at the high level than the low level (all *p* values less than 0.01). These seemingly counterintuitive inferences emphasize the importance of controlling for confounding factors in

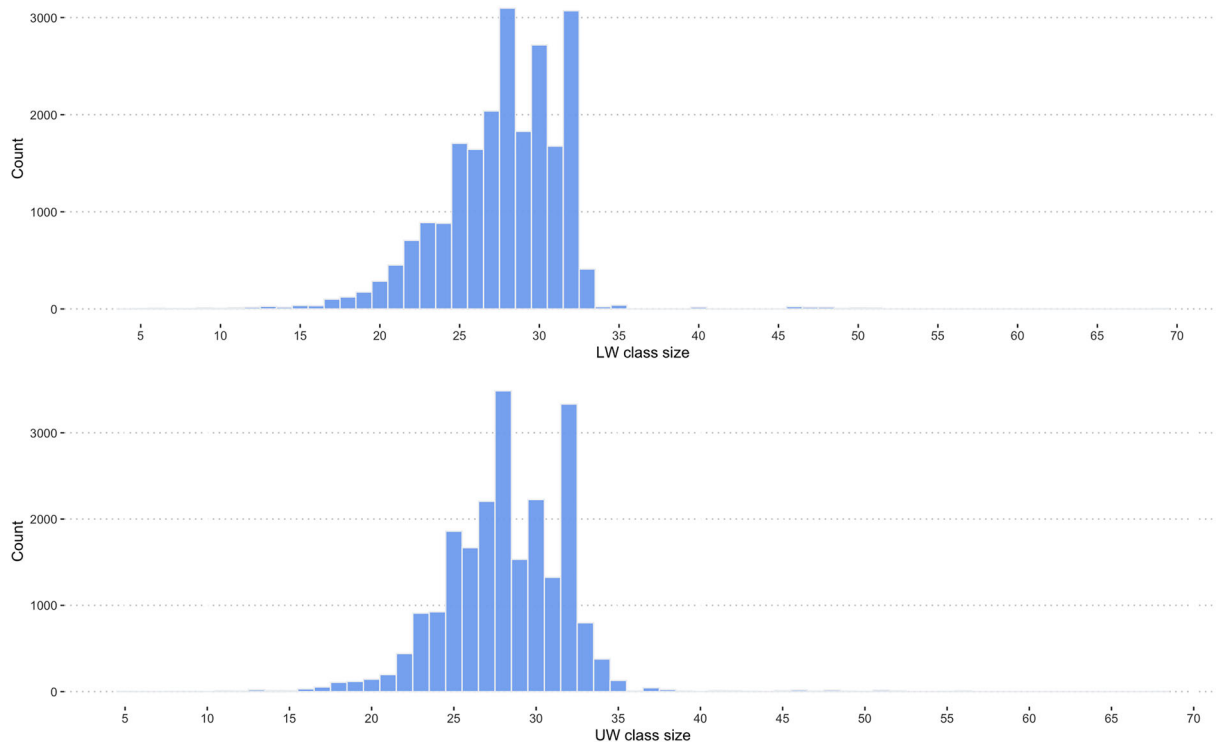


FIGURE 2 Distribution of class sizes for LW (above) and for UW (below) courses. The bar heights (y axis) are the number of writing courses at a given class size (x axis)

analyses of observational data. This is a common finding in student success research (Alyahyan & Dustegor, 2020; Shahiri et al., 2015). Once we control for covariates in subsequent analyses, relationships between class sizes and WPA performance conform to expectations.

Tables 1 and 2 present summaries of covariates available to us for modelling. The tables categorize the inputs in continuous variables, exposing the minimum, maximum, mean, median and standard deviation, and categorical variables, exposing group percentages. The data were obtained from a query of the student information database and curated for the Data Champions programme. The only processing for this project was creation of the outcome levels and class size categories. The only missing data occurred in SAT scores. Some students reported an ACT score, we used tables from the College Board to convert these to SAT scores. After that, 0.063% of the SAT scores remained missing for which we imputed using the `mice` package in R (van Buuren & Groothuis-Oudshoorn, 2011).

In the modelling, we first include these inputs individually to study important predictors. But given focus is on class size, all other inputs serve as covariates in the modelling process. We thus perform a data engineering step to identify features combining these inputs for the sake of dimension reduction in subsequent regression modelling.

2.3 | Analytics workflow

Recall that our penultimate goal is to study the relationship between WPA performance and writing course (LW and UW) class sizes, controlling for student characteristics. Our research questions are as follows:

- What predicts WPA performances? More specifically, does class size predict WPA performance upon controlling for academic performance and demographic variables?
- If class size does predict WPA performance, to what extent does it do so?

Our RWS Department and SDSU administrative clients wished to estimate an optimal class size for writing courses. Given the retrospective nature of the study and lack of data on instructors, curriculum/curricular changes and student writing analyses other than the WPA, we believe that this goal is a bit of a pipe dream; we will discuss these issues more in Section 4. Nonetheless, our analytics workflow progressed, in collaboration with our subject and administrative partners, with an eye on the impact of larger class sizes on student success. We note that the analytics

TABLE 2 Summary statistics for the categorical inputs in the data set

Variable code	Label	Proportion	Description
sex	Female	60%	Assigned gender at birth
	Male	40%	
URM	0 is no	59%	Underrepresented minority student
	1 is yes	35%	
	2 is other	6%	
stem	0 is not STEM major	78%	Science, Technology, Engineering, & Mathematics major
	1 is STEM major	22%	
stu_college	Arts & Letters	8%	College of major at SDSU
	Business	17%	
	Education	1%	
	Engineering	9%	
	Health & Human Services	11%	
	Professional Studies & Fine Arts	19%	
	Sciences	18%	
	Undergraduate Studies	17%	
preMajor_stat_des	Premajor	73%	Major status
	Major	27%	
honors	0 is no	98%	Enrolment in the honours programme
	1 is yes	2%	
disability_des	0 is no	99%	Registration with disability services
	1 is yes	1%	
eop	0 is no	80%	Participation in the educational Outreach programme
	1 is yes	20%	
dorm	0 is no	64%	On-campus dorm housing
	1 is yes	36%	
timestat_des	Full-time	85%	Attendance status at SDSU
	Part-time	15%	
low_income_EFC	0 is no	63%	Low income student
	1 is yes	37%	
pell_indicator	0 is no	80%	Recipient of Pell Grant
	1 is yes	20%	
admbas_des	FTF from CA	94%	First time freshmen admissions
	FTF not from CA	6%	
	FTF foreign apps	0.1%	
first_gen_nces	0 is no	55%	First-generation student, parents had no college
	1 is yes	45%	
first_gen_some_college	0 is no	80%	First-generation student, parents had some college
	1 is yes	20%	
learning_community	0 is no	88%	Education programme in dorm
	1 is yes	12%	
compact	0 is no	95%	COMPACT scholar programme
	1 is yes	5%	
math_prof_des	SAT/ACT	54%	Math proficiency for admission
	ELM	17%	
	Early start programme	3%	
	Remediation	26%	

(Continues)

TABLE 2 (Continued)

Variable code	Label	Proportion	Description
engl_prof_des	SAT/ACT	33%	English proficiency for admission
	EAP/EPT test	26%	
	Early start programme	2%	
	Remediation	39%	
vet_stat_des	Active duty	0.01%	Veteran status description
	Not military	98.5%	
	Military dependent	1.5%	
study_abroad	0 is no	94%	Study abroad
	1 is yes	6%	
local	0 is no	60%	From SDSU service region
	1 is yes	39%	
efc_category	0	28%	Expected family contrib, financial aid; for confidentiality, categories collapsed
	not 0	72%	
LW.course	RWS 101	11%	Lower level writing course name
	RWS 100	84%	
	CCS 111B	2%	
	LING 100	0.4%	
	AFRAS 120	2%	
	GEN S250A	0.04%	
	AMIND 120	0.6%	
UW.course	RWS 200	90%	Upper level writing course name
	AFRAS 200	1%	
	PHIL 110	6%	
	GEN S250D	0.1%	
	CCS 200	2%	
	LING 200	0.5%	
	GEN S260D	0.07%	
AP_indicator	0 is no	99.8%	AP credit for writing requirement
	1 is yes	0.2%	

workflow presented here grew organically from regular reports to and feedback from our clients. We will first outline our statistical analysis and then briefly detail the methods and innovations therein.

1. Exploratory data analysis (EDA), briefly summarized in Sections 2.1 and 2.2: response; class size
2. First pass on identifying a class size threshold: random forest importance ranking
3. First pass on inference:
 - factor analysis of mixed data (FAMD) to reduce the dimension of the feature space
 - regression of WPA performance on LW and UW course threshold variables and FAMD components
4. Quantify a treatment effect:
 - partial dependence plot from a random forest fit of WPA performance on LW and UW course sizes and FAMD components
 - covariate balancing generalized propensity scores (CBGPSs) for studying the impact of LW and UW course sizes on WPA performance

We performed a detailed EDA studying the relationship between WPA performance and class size with respect to the covariates collected; we do not have the space to fully exposit that work. As suggested in the EDA of Section 2.2, class size may impact WPA performance, but the message is not clear given the strong predictive power of academic performance measures and student academic characteristics. We expanded a random forest approach we developed previously (He, Levine, Bohonak, et al. 2018), whereby we create dummy variables for the class size thresholds, for example, $c_{25} = I(\text{class size} \leq 25)$. The class size categories we use are derived from recommendations by the RWS faculty and the

previous Senate task force on the subject. As random forest can handle correlated inputs as part of predictive modelling (He, Levine, Fan, et al. 2018), we can take advantage of the variable importance ranking functionality to identify the class size threshold that is most important in predicting WPA performance after controlling for all covariates in the data set.

Upon identifying this important class size threshold, we wish to infer the strength of the relationship with WPA performance. Given the size of the covariate feature space in this study, regression modelling is problematic due to effect modification and multicollinearity, to say the least. We are considering the academic performance inputs and the student academic characteristics as covariates in this study. In particular, drawing inferences on the relationship between WPA performance and these variables could have potential ramifications of enrolling students in courses with specific enrolment numbers based on identities and academic preparedness. This was not a direction the clients wanted to entertain. We thus performed a variation of principal components regression (James et al., 2021). We have mixed continuous and categorical data and thus chose to apply FAMD to identify a small set of linear combinations of the covariates that explained the most variance in the covariates (Lê et al., 2008). In brief, FAMD uses principal component analysis for quantitative variables and multiple correspondence analysis for qualitative variables, standardizing/scaling each variable to place their influence on the analysis on an equal footing. In our case, FAMD was able to reduce our covariates down to a set of three factors. Of course we cannot interpret these factors, but again, our goal was merely to control for the covariates not interpret specific relationships between individual covariate inputs and WPA performance. With this dimension reduction, regression models of WPA performance on LW course size, UW course size and the FAMD factors (so five inputs in all) were easily fit, including regression assumption diagnostics and regression inferences.

The analysis to this point identifies an important course size threshold and infers its impact on predicting WPA performance. But the course size is a discrete integer variable, and we wish to quantify how WPA performance may change over the course size. We consider two modelling approaches for this problem. We first study partial dependence plots (Greenwell, 2017) off a random forest fit, with which we estimate a marginal effect for the writing course sizes after controlling for all other covariates. We will have to fit a random forest now on the actual course sizes, rather than the class size thresholds, as we want to estimate a treatment effect, namely, the impact of LW course size and of UW course size on the probability of passing the WPA.

As a second approach to estimating a treatment effect, note that we do not have a randomized controlled study where students are randomly placed into writing courses of specific sizes. We may approach such an analysis through the lens of an observational study. As such, we are concerned about selection bias, namely, students characterized by specific academic performance and/or academic demographics may be more likely to choose a large or a small writing class. Propensity score methods have proved valuable in the personalized medicine literature for balancing covariates between treatment groups, for example see Austin and Stuart (2015). In our setting, the propensity score is the probability a student will select a course of a given class size given their covariate make-up. We have a discrete treatment and class size and must resort to so-called generalized propensity scores. We have extended the CBGPS approach (Fong et al., 2018) to an educational data mining setting (Shao et al., 2022) and thus used this method here to quantify a treatment effect. We note that our current area of research is extending CBGPS to multinomial outcomes and multiple treatments. For the analyses on the class size project collaboration presented herein, we thus present results from separate models on LW class size and UW class size and draw inferences for the binary WPA outcome (pass/fail). Ultimately, the treatment effect is estimated from a generalized additive model (James et al., 2021) of the WPA outcome on the treatment (so a single predictor, either LW class size or UW class size), with weights based on the estimated generalized propensity scores.

3 | RESULTS

Figure 3 presents variable importance plots from random forest fits of the binary WPA outcome (pass/fail) and the trinary WPA outcome (high/medium/low), respectively, on all inputs. In both cases, academic performance measures of high school GPA and SAT scores and student academic characteristics of college major and financial aid category are the strongest predictors. We note too that the course size thresholds do not rank among the top 20 inputs in the random forest models. This is a common finding when test results are used as definitions of student success (Bresciani Ludvik, 2018). Academic preparation and performance prior to taking the WPA are potential confounding factors, in our case perhaps overwhelming the impact of class size on the WPA outcome. Figure 3 thus provides further evidence on the need to control for these factors in subsequent analyses. The most important class size cut point is a class size of 27, for both courses in the writing sequence (LW and UW).

Tables 3 and 4 present regression inferences of the binary WPA outcome (pass/fail) and the trinary WPA outcome (high/medium/low), respectively, on the three FAMD factors and the LW and UW class size threshold of 27. Interactions among FAMD factors and class size thresholds were not predictive of WPA performance. Recall that the FAMD factors summarize the covariates, inputs other than class size, into three scores for this regression analysis. A relationship between WPA performance and class size exists, even after controlling for these academic performance measures and student academic characteristics. As predicted by the EDA and modelling to this point, the impact of class size on WPA performance is not strong. However, students in LW and UW classes of smaller size are more likely to attain a score of 8 or above on the WPA. Students in a LW and a UW class below 27 are 1.22 times more likely (95% CI of 1.12 to 1.32) and 1.12 times more likely (95% CI of 1.03 to 1.21), respectively, to pass the WPA; odds ratios computed by exponentiating the slope estimates on Table 3. Students in a LW and a UW class

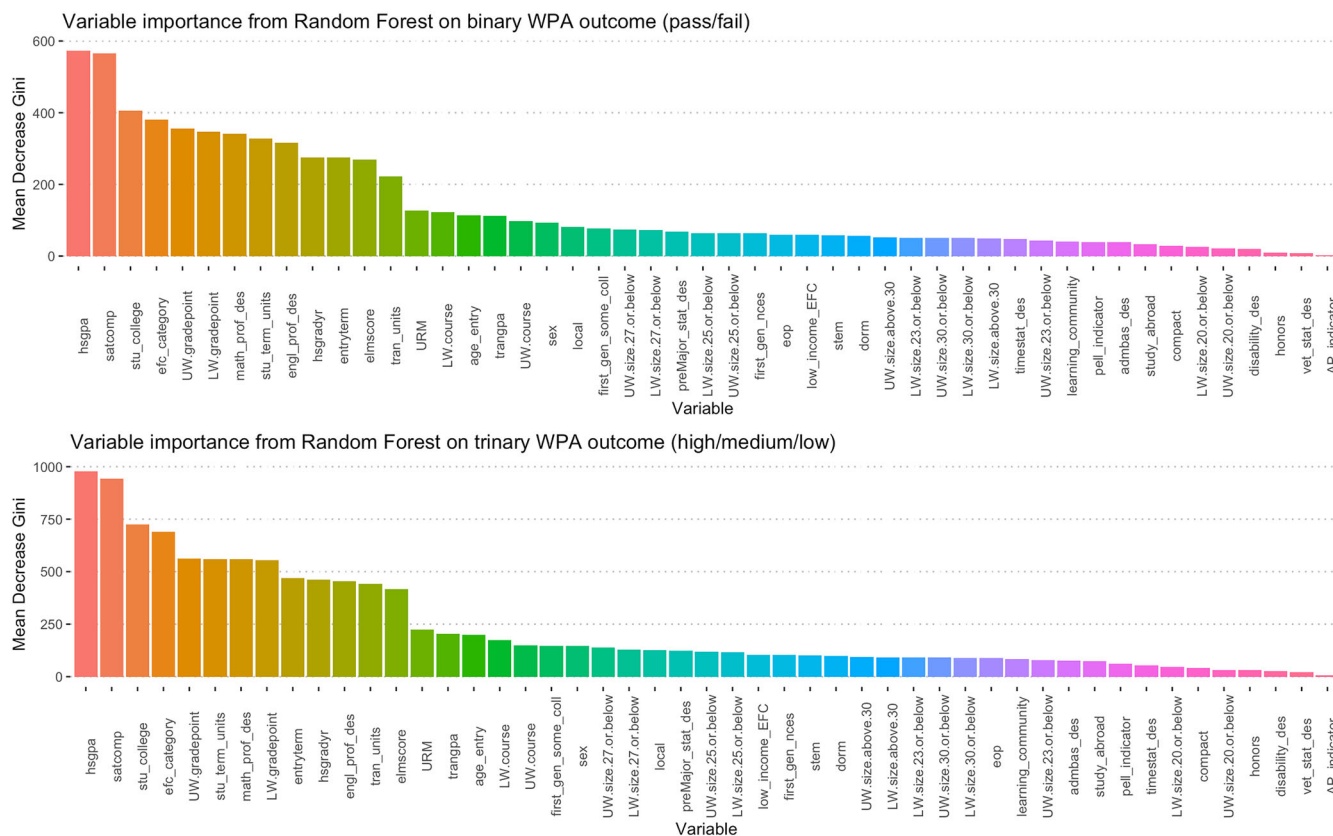


FIGURE 3 The left graphic presents random forest variable importance plot of binary WPA performance outcome (pass/fail) on all inputs. The right graphic presents random forest variable importance plot of the three-category WPA performance outcome (high/medium/low) on all inputs. See Tables 1 and 2 for variable definitions

TABLE 3 Inferences from regression of *binary WPA performance (pass/fail)* on the three FAMD scores, LW course size and UW course size

Input	Estimate	SE	p value	95% CI
Intercept	1.46	0.028	<0.0001	(1.41,1.52)
FAMD 1	0.27	0.008	<0.0001	(0.26,0.29)
FAMD 2	0.12	0.011	<0.0001	(0.10,0.14)
FAMD 3	0.15	0.011	<0.0001	(0.13,0.17)
LW course size <27	0.20	0.042	<0.0001	(0.11,0.28)
UW course size <27	0.11	0.041	0.010	(0.03,0.19)

below 27 are 1.20 times more likely (95% CI of 1.09 to 1.30) and 1.11 times more likely (95% CI of 1.02 to 1.20), respectively, to score a medium mark compared to a low mark on the WPA. Students in a LW and a UW class below 27 are 1.32 times more likely (95% CI of 1.19 to 1.48) and 1.16 times more likely (95% CI of 1.04 to 1.28), respectively, to score a high mark compared to a low mark on the WPA. These latter odds ratios are computed by exponentiating the slope estimates on Table 4.

Figure 4 presents partial dependence plots for LW and for UW class sizes from a random forest fit on the binary WPA outcome (pass/fail), controlling for all other covariates. Large class sizes negatively impact the pass rate on the WPA. Students in small class sizes, less than 20, have a lower pass rate on the WPA. Class sizes between 20 and 30 have a similar and desirably high pass rate on the WPA.

We also estimate a generalized propensity score for class size via two CBGPS fits of WPA outcome (pass/fail) on LW class size and on UW class size, each model controlling for all other covariates. Figure 5 presents the treatment effect estimated from generalized additive models of the WPA binary outcome on LW class size and on UW class size, with weights based on the generalized propensity scores estimated from CBGPS. Again, students in LW class sizes between 20 and 30 are predicted to pass at a rate above 80%, and the pass rate declines for larger (above 30) and smaller (below 20) LW class sizes. This CBGPS-based treatment effect plot estimates an even sharper decline for large LW class sizes than

TABLE 4 Inferences from regression of *trinary WPA performance (high/medium/low)* on the three FAMD scores, LW course size and UW course size

Input	Estimate	SE	p value	95% CI
Medium vs. low				
Intercept	1.25	0.028	<0.0001	(1.20,1.31)
FAMD 1	0.24	0.008	<0.0001	(0.22,0.26)
FAMD 2	0.13	0.011	<0.0001	(0.11,0.15)
FAMD 3	0.14	0.015	<0.0001	(0.11,0.16)
LW course size <27	0.18	0.043	<0.0001	(0.09,0.26)
UW course size <27	0.10	0.042	0.017	(0.02,0.18)
High vs. low				
Intercept	-0.26	0.038	<0.0001	(-0.34,-0.19)
FAMD 1	0.42	0.011	<0.0001	(0.40,0.44)
FAMD 2	0.05	0.015	0.001	(0.02,0.08)
FAMD 3	0.22	0.015	<0.0001	(0.19,0.25)
LW course size <27	0.28	0.056	<0.0001	(0.17,0.39)
UW course size <27	0.15	0.054	0.005	(0.04,0.25)

Note: The low WPA scoring group is the reference level for inference.

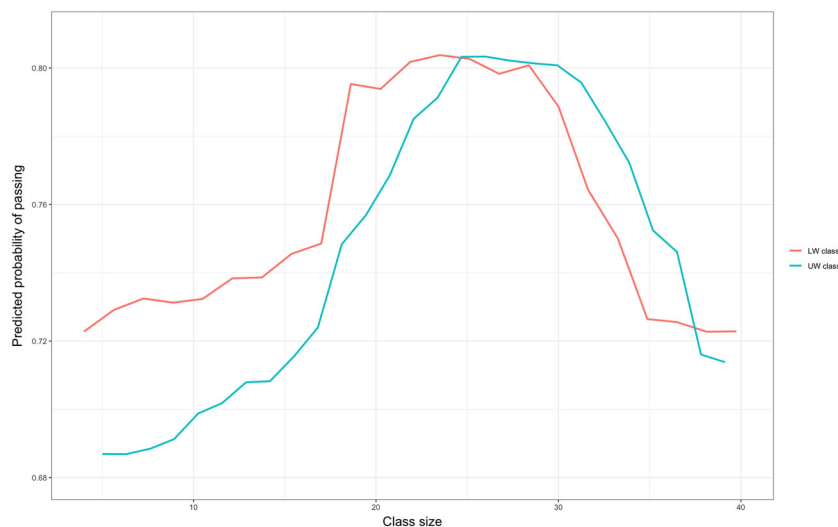


FIGURE 4 Partial dependence plots from random forest fits predicting WPA pass rate over LW (red) and over UW (turquoise) class sizes

the corresponding partial dependence plot in Figure 4. Figure 5 also shows that students in UW class sizes above 20 pass the WPA at a rate above 80%, the treatment effect estimate flat even out to larger class sizes above 30. This relationship is not surprising as students taking the UW course have passed the LW course and are more mature (older, longer tenure at the university) than the LW students. As such, we may expect the impact of class size to be less for the UW course.

We present these treatment effect visualizations, Figures 4 and 5, over a class size range of 10 to 40 since the sample size is too small to produce reliable treatment effect estimates for small and large class sizes outside this range. (See Figure 2 for the class size distributions.) As an overall summary, students seem to perform better on the WPA when the LW class size is below 27. The impact of UW class size on WPA performance is not as clear cut: there is some indication that UW class size below 27 is preferable, but the effect size is small. Interestingly, smaller LW and UW class sizes, below 20, are negatively related to WPA performance. We shall provide more explanation for this finding in Section 4 next.

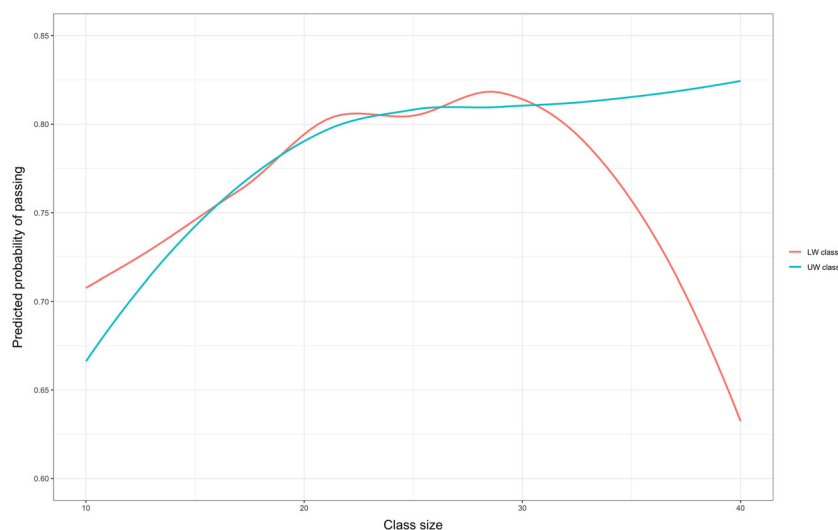


FIGURE 5 Predicted WPA pass rate over LW (red) and UW (turquoise) class sizes by balancing covariates using CBGPS

4 | DISCUSSION

4.1 | On statistical practice

In this paper, we present an example of a learning analytics project representative of the cross-disciplinary collaborations we have found becoming increasingly common on our campus, and we suspect higher education institutions broadly. The specific problem was posed by the University Senate: how does class size impact student success in required GE writing courses. This project was broached due to the realization that a quantitative assessment of writing course class size at a large, hispanic-serving, public institution and urban campus was needed. Of note, on the one hand, the administration was unwilling to devote significant resources on the foundation of decades old recommendations from the writing community, as cited in Section 1 earlier, nor on what they perceived as qualitative assessments scored using faculty agreed upon rubrics and anecdotal evidence provided by a previous task force and the writing community. On the other hand, the RWS Department and writing instructors across campus documented hampered student success and attainment of learning outcomes in larger enrolment writing courses.

Our key finding was that writing classes need to enroll between 20 and 27 students. In such class sizes, students are 1.11 to 1.32 times more likely to succeed on an UW test. This inference came out of a collaborative effort between analysts, education experts and administrative stakeholders. We perceive a number of lessons learned that may be absorbed by readers in statistical practice more broadly. These are already centrepieces of successful statistical consulting relationships and statistical collaboration projects in the field, and we think it is worth expounding within the context of our project here in the remainder of this subsection.

There has been a growing data-informed decision-making culture on our campus, with an appreciation of statistical visualizations and statistical learning machinery for understanding data nuances/complexities, quantifying uncertainty and drawing inferences. This path was paved by a Learning Analytics Task Force constructed in 2013 by forward-thinking leadership in the offices of instructional technology and institutional effectiveness. The group consisted of faculty with Statistics expertise and key administrative stakeholders in institutional research, instructional technology, student affairs and the student information system with interest and expertise in the nascent discipline of learning analytics. One significant product of the Learning Analytics Task Force was funding and development of a Data Champions programme starting in 2017. The Data Champions programme embodies active problem solving and collaborative discussions in the classroom. Each academic year, the programme enrolls campus student success leaders to propose and complete a student success project. These so-called Data Champions are trained on data confidentiality and privacy, survey design and survey instruments and statistical design, visualizations, analysis and communication. Each Data Champion is paired with a Data Coach, a position held by most authors on this paper. The Data Champion and Data Coach met regularly throughout the year as research questions were derived, data needs were established, study designs were formulated and executed, data were curated and analysed and findings were reported. Monthly Data Champions programme meetings included the 11 Data Champions, student success teams from their respective units/colleges and the Data Coaches. These interactions spurred cross-disciplinary discussions on the student success projects pursued, challenges therein and related problem solving. A primary goal of the programme was to equip campus student success leaders with the tools to run their own student success studies. Additionally, the programme provided a collaborative infrastructure with dedicated resources and a group of people with a common interest in student success.

The Data Champions programme provided a shared data set from which to draw the features and outputs for the writing course class size analysis herein. Furthermore, the Data Champions are now a large cohort of campus student success leaders with an appreciation of the subtleties of institutional data, the inherent biases in observational studies, the value of EDA and the importance and potential drawbacks of machine learning machinery; the programme mantra is that “data informs decisions, it should not drive it!” As such, a collaborative effort for the writing course class size project was an expectation, not a proposal requiring justification. There was no pushback in the proposal for using random forest and propensity score-based innovative educational data mining methods our research team developed. To the contrary, Data Science students were explicitly invited to derive methodology angles for thesis research and gain invaluable applied statistics training by participating as Data Coaches on student success projects. On the flip side, there was regular and transparent communication among the analysts and clients, from which challenges in the data collection were resolved. For example, the Class Size Task Force gained an understanding of changes in the WPA administration and grading, changes in the sequences and student learning outcomes in writing courses and penultimately zeroed in on a time period where the class offerings and WPA were stable. Furthermore, the Task Force emphasized the desire for punchlines, messaging and visualizations to maximize absorption by the clients. The graphics and inference tables in Section 3 illustrate the statistical reporting that derived from these collaborations. And our current work entails developing interactive dashboards for users to parse results by student demographics, when it is beneficial for such dialogue to occur.

This project was a by-product of the culture that grew out of the Data Champions programme. An institution organized in a way to cultivate these kinds of conversations and invest in professional development to this end will lend to a sustainable data-informed decision-making environment.

4.2 | Future directions

A multi-institutional study of good practice analysis in how institutional leaders use data to inform decision-making revealed the importance of building interdisciplinary teams to discuss with clients the problems that need to be resolved and the methodological approaches that will best inform decisions to improve those problems (Bresciani Ludvik, 2006). Senge (2006) wrote about how such use of data would illustrate how the institution itself is learning and improving. However, with the advancement of learning analytics to access increasingly sophisticated data mining approaches to refine that decision-making (Deakin Crick, 2017), we believe it is good statistical practice to discuss, with clients, statistical issues that may be addressed in future data collection and analyses. While not a novel discovery, Knight et al. (2014) posited that one way through this is in dialogue with the analysts and the decision-makers.

While most organizations begin the dialogue once the data scientists have mined and analysed the data, this team worked closely with the clients throughout the process. What began as a question of how to better utilize institutional resources resulted in a greater understanding of the stress caused to writing faculty. Writing instructors had to—often quickly—redesign courses and grading systems. Furthermore, a placement exam that was out of their locus of control, and remained somewhat static, continued to be used to measure high stakes students' success and potentially inform organizational resource allocation.

The analytics team applied novel random forest and propensity score methods to reveal that a class size for student success, as measured by a writing placement exam, was between 20 and 27 students. While this finding was of interest to the client, of greater interest was our ability to help the clients identify potential confounding from impacts of curriculum, instructors and campus student success programmes. The RWS faculty shared that increases in class size caused significant pedagogical adaptations that deviated from the best practices recommended by the writing education literature and professional organizations (on, e.g., individual meetings with students, number of assignments, magnitude of assignments—small stakes writing vs. longer essays, feedback detail and frequency, peer evaluations and amount of revision). Also, smaller classes were often taught in computer labs capped by the number of computers and less conducive to some forms of group work, collaboration, discussion and peer review work. The most experienced and strongest lecturers were often assigned the large classes whereas graduate teaching assistants with less experience and training were assigned the smaller classes. Finally, during the period of study, the campus introduced a writing centre and funded a number of tutoring programmes for underserved students. The relationship between WPA performance and writing course class size may thus be confounded with mode of instruction, instructor and supplemental instruction.

The team was able to make recommendations to proactively collect data on instructors, student writing preparedness characteristics and writing centre/tutoring centre attendance. The Class Size Task Force also recommended standardizing common learning outcomes, assessments and grading rubric across writing courses, collecting evidence of writing proficiency from graded student papers (Bresciani Ludvik, 2018). With these data in hand, a more detailed longitudinal study may be designed to infer optimal class sizes (treatments) in student subgroups and the impact of instructor workload on student learning outcomes in writing classes. A resource analysis was also recommended as it is not clear if the cost in offering large writing classes in tandem with supplemental instruction programmes offsets the additional cost in offering smaller writing classes, where perhaps less extensive supplemental instruction support is needed. Beyond cost, measures of student success to evaluate include attaining specific course learning outcomes, achieving a desired level of writing and communication skill, the attainment of and interaction with career readiness skills and other skills known to correlate with and predict student success such as sense of belonging and academic self-efficacy,

as well as progress through a degree programme (e.g., probation, retention, course repeats, performance in major, graduation and impact on internship/job prospects) (Bresciani Ludvik, 2018). Of note, a smaller sized writing course may not have a large impact on WPA scores but may greatly enhance students' sense of belonging at school and in their peer groups. The goal is to balance economic stability of an institution while attaining student learning goals and maintaining student growth.

These recommendations—those found most useful by the clients—would not have been possible without the innovative statistical learning tools developed for quantifying treatment effects and the collaborative process used to design and pursue the analysis.

ACKNOWLEDGEMENT

This work is funded in part by NSF Grant 1633130.

DATA AVAILABILITY STATEMENT

Data not available due to FERPA, student confidentiality.

ORCID

Richard A. Levine  <https://orcid.org/0000-0002-7553-4264>

REFERENCES

- Alyhayan, E., & Dustegor, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17, 3. <https://doi.org/10.1186/s41239-020-0177-7>
- Associate of Departments of English. (2020). ADE guidelines for class size and workload for college and university instructors of english: A statement policy. <https://www.maps.mla.org/Resources/Policy-Statements/ADE-Guidelines-for-Class-Size-and-Workload-for-College-and-University-Instructors-of-English-A-Statement-of-Policy>. Accessed: 2022-05-01.
- Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34(28), 3661–3679.
- Bresciani Ludvik, M. J. (2006). *Outcomes-based academic and co-curricular program review: A compilation of institutional good practices*: Stylus.
- Bresciani Ludvik, M. J. (2018). *Outcomes-based program review: Closing achievement gaps in and outside the classroom with alignment to predictive analytics and performance metrics* (2nd ed.). Stylus.
- Crick Deakin, R., Knight, S., & Barr, S. (2017). Towards analytics for wholistic school improvement: Hierarchical process modeling and evidence visualization. *Journal of Learning Analytics*, 4, 160–188.
- Deakin Crick, R. (2017). Learning analytics: Layers, loops, and processes in a virtual learning infrastructure, *Handbook on learning analytics* (pp. 291–308). SOLAR.
- Fong, C., Hazlett, C., & Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12, 156–177.
- Gasevic, D., Tsai, Y.-S., Dawson, S., & Pardo, A. (2019). How do we start? An approach to learning analytics adoption in higher education. *The International Journal of Information and Learning Technology*, 36, 342–353.
- Greenwell, B. M. (2017). pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal*, 9(1), 421–436. <https://doi.org/10.32614/RJ-2017-016>
- Guarcello, M. A., Levine, R. A., Beemer, J., Frazee, J. P., Laumakis, M. A., & Schellenberg, S. A. (2017). Balancing student success: Assessing supplemental instruction through coarsened exact matching. *Technology, Knowledge and Learning*, 22(3), 335–352.
- He, L., Levine, R. A., Bohonak, A. J., Fan, J., & Stronach, J. (2018). Predictive analytics machinery for stem student success studies. *Applied Artificial Intelligence*, 32(4), 361–387.
- He, L., Levine, R. A., Fan, J., Beemer, J., & Stronach, J. (2018). Random forest as a predictive analytics alternative to regression in institutional research. *Practical Assessment, Research & Evaluation*, 23(1), 1. Available online: <http://pareonline.net/getvn.asp?v=23&n=1>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R* (2nd ed.): Springer.
- Knight, S., Buckingham Shum, S., & Littleton, K. (2014). Epistemology, assessment, pedagogy: Where learning meets analytics in the middle space. *Journal of Learning Analytics*, 1(2), 23–47.
- Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1), 1–18.
- Macfadyen, L. P., Dawson, S., Pardo, A., & Gasevic, D. (2014). Embracing big data in complex educational systems: The learning analytics imperative and the policy challenge. *Research & Practice in Assessment*, 9, 17–28.
- Pelaez, K., Levine, R. A., Guarcello, M. A., & Fan, J. (2019). Using a latent class forest ensemble to identify at-risk students in higher education. *Journal of Educational Data Mining*, 11, 18–46.
- Piety, P. J., Hickey, D. T., & Bishop, M. J. (2014). Educational data sciences—Framing emergent practices for analytics of learning, organizations, and systems. In *LAK '14: Proceedings of the Fourth International Conference on Learning Analytics And Knowledge* (Vol. 72, pp. 193–202). Association for Computing Machinery.
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10, e1355.
- Senge, P. M. (2006). *The fifth discipline: The art and practice of the learning organization*. Doubleday.
- Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 414–422.
- Shao, L., Levine, R. A., Guarcello, M. A., Wilke, M. C., Stronach, J., Frazee, J. P., & Fan, J. (2022). Estimating a dose-response relationship in quasi-experimental student success studies. *International Journal of Artificial Intelligence in Education*.

- Tsai, Y.-S., Moreno-Marcos, P. M., Jivet, I., Scheffel, M., Tammets, K., Kollom, K., & Gasevic, D. (2018). The SHEILA framework: Informing institutional strategies and policy processes of learning analytics. *Journal of Learning Analytics*, 5, 5–20.
- UT Austin Undergraduate Studies (2020). Recommendations on class size of writing flagged courses. <https://ugs.utexas.edu/flags/class-size-statement>. Accessed: 2022-05-01.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Levine, R. A., Rivera, P. E., He, L., Fan, J., & Bresciani Ludvick, M. J. (2023). A learning analytics case study: On class sizes in undergraduate writing courses. *Stat*, 12(1), e527. <https://doi.org/10.1002/sta4.527>