UNIVERSITY OF CALIFORNIA,
IRVINE

Equity-Oriented Educational Data Science

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Education

by

Renzhe Yu

Dissertation Committee:
Professor Mark Warschauer, Chair
Associate Professor Di Xu
Assistant Professor Rene Kizilcec
Chancellor's Professor Padhraic Smyth
Associate Professor Sameer Singh

2022

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

ACKNOWLEDGMENTS

The completion of this dissertation and my doctoral journey is not possible without many people who have been supportive in various ways. My deepest gratitude goes to Professor Mark Warschauer, who has provided every possible form of professional and life support to help me become an independent researcher. I would also like to thank my other dissertation committee members, especially Professor Di Xu and Professor Rene Kizilcec who have been my closest research collaborators and sources of inspiration. Moreover, I am extremely grateful to Professor Richard Arum for creating an open space to scale up my research in real-world contexts. Lastly, I want to recognize the intellectual and personal support from my peers and collaborators across the world, especially those from the Digital Learning Lab at UC Irvine and the Future of Learning Lab at Cornell.

My research accomplishment is also credited to two wonderful fellowship experiences at Alan Turing Institute and IBM Research. I must also acknowledge the organizational and financial support from UC Irvine School of Education, National Science Foundation, and Andrew W. Mellon Foundation. Without such support I cannot even build a research career.

Finally, I would like to thank my family for their unwavering support and company through my doctoral journey. I love you all.

The text of Chapter 4 in this dissertation is a reprint of the material as it appears in *Towards Accurate and Fair Prediction of College Success: Evaluating Different Sources of Student Data*, (2020), Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020), 292-301, used with permission from the International Educational Data Mining Society (IEDMS). The co-authors listed in this publication are Qiujie Li, Christian Fischer, Shayan Doroudi, and Di Xu. The text of Chapter 5 in this dissertation is a reprint of the material as it appears in *Should College Dropout Prediction Models Include Protected Attributes?* (2021), Proceedings of the 8th ACM Conference on Learning @ Scale (L@S '21), 91-100, https://doi.org/10.1145/3430895.3460139. The co-authors listed in this publication are Hansol Lee and René F. Kizilcec.

VITA

Renzhe Yu

EDUCATION

Doctor of Philosophy in Education	**2022**
University of California, Irvine	*Irvine, CA, USA*
Master of Arts in Economics of Education	**2017**
Peking University	*Beijing, China*
Bachelor of Science in Artificial Intelligence	**2014**
Bachelor of Arts in Economics	**2014**
Peking University	*Beijing, China*

RESEARCH EXPERIENCE

Graduate Student Researcher	**2018–2022**
University of California, Irvine	*Irvine, California*
Research Intern	**2020**
IBM Research	*Remote*
Summer Fellow	**2019**
The Alan Turing Institute	*London, England, UK*
Visiting Researcher	**2018**
University of California, Berkeley	*Berkeley, CA, USA*
Graduate Student Researcher	**2014–2017**
Peking University	*Beijing, China*

TEACHING EXPERIENCE

Guest Lecturer	**2019**
University of California, Irvine	*Irvine, CA, USA*
Teaching Assistant	**2019**
University of California, Irvine	*Irvine, CA, USA*
Teaching Assistant	**2016**
Peking University	*Beijing, China*

SELECTED HONORS & FELLOWSHIPS

Dissertation Fellowship (Declined) **2022**
National Academy of Education / Spencer Foundation

David P. Gardner Fellow **2021–2022**
University of California, Berkeley

Public Impact Fellow **2021**
University of California, Irvine

Best Paper Nomination **2021**
ACM Conference on Learning at Scale (L@S)

Best Paper Honorable Mention **2020**
AERA Educational Data Science Conference

Science for Social Good Fellow **2020**
IBM Research

Data Science for Social Good Fellow **2019**
The Alan Turing Institute

Best Paper Award **2018**
International Conference on Educational Data Mining (EDM)

REFEREED JOURNAL PUBLICATIONS

Moeller, J., von Keyserlingk, L., Spengler, M., Gaspard, H., Lee, H., Yamaguchi-Pedroza, K., **Yu, R.**, Fischer, C., & Arum, R. (2022). Risk and protective factors of college students' psychological well-being during the COVID-19 pandemic: Emotional stability, mental health, and household resources. *AERA Open*.

Umarji, O., Day, S., Xu, Y., Zargar, E., **Yu, R.**, & Connor, C. (2021). Opening the black box: User-log analyses of children's e-Book reading and associations with word knowledge. *Reading and Writing*, 34(3): 627–657.

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.

Fischer, C., Pardos, Z., 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.

Ha, W., & **Yu, R.** (2019). Quasi-experimental evidence of a school equalization reform on housing prices in Beijing. *Chinese Education & Society*, 52(3-4): 162–185.

Ha, W., & **Yu, R.** (2017). How much is an improved school worth? Evidence from the comprehensive reform in compulsory education in Beijing. *Peking University Education Review*, 15(03): 137–153.

Ha, W., Wu, H., & **Yu, R.** (2015). A new research on the capitalization of school quality in housing prices: An empirical study based on repeated cross-sectional data in Beijing. *Education & Economy*, 05: 3–10.

REFEREED CONFERENCE PUBLICATIONS

Chopra, H., Lin, Y., Samadi, M., Cavazos, J., **Yu, R.**, Jaquay, S., & Nixon, N. (2022). Modeling student discourse in online discussion forums using semantic similarity based topic chains. In *Proceedings of the 23rd International Conference on Artificial Intelligence in Education (AIED '22), Part II* (pp. 453–457).

Sabnis, S., **Yu, R.**, & Kizilcec, R. (2022). Large-scale student data reveal sociodemographic gaps in procrastination behavior. In *Proceedings of the 9th ACM Conference on Learning at Scale (L@S '22)* (pp. 133–141).

Yu, R., Das, S., Gurajada, S., Varshney, K., Raghavan, H., & Lastra-Anadon, C. (2021). A research framework for understanding education-occupation alignment with NLP techniques. In *Proceedings of the 1st ACL Workshop on NLP for Positive Impact* (pp. 100–106).

Yu, R., Lee, H., & Kizilcec, R. (2021). Should college dropout prediction models include protected attributes? In *Proceedings of the 8th ACM Conference on Learning at Scale (L@S '21)* (pp. 91–100).

Li, X., & **Yu, R.** (2021). Construction of weighted course co-enrollment network. In *Companion Proceedings of the 11th International Conference on Learning Analytics & Knowledge (LAK '21)* (pp. 464–469).

Kung, C., & **Yu, R.** (2020). Interpretable models do not compromise accuracy or fairness in predicting college success. In *Proceedings of the 7th ACM Conference on Learning at Scale (L@S '20)* (pp. 413–416).

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 (EDM '20)* (pp. 292–301).

Lin, Y., **Yu, R.**, & Dowell, N. (2020). LIWCs the same, not the same: Gendered linguistic signals of performance and experience in online STEM courses. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education (AIED '20)* (pp. 333–345).

Yu, R., Pardos, Z., & Scott, J. (2019). Student behavioral embeddings and their relationship to outcomes in a collaborative online course. In *Joint Proceedings of the Workshops of the 12th International Conference on Educational Data Mining (EDM '19)* (pp. 23–29).

Rodriguez, F., **Yu, R.**, Park, J., Rivas, M., Warschauer, M., & Sato, B. (2019). Utilizing learning analytics to map students' self-reported study strategies to click behaviors in STEM courses. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK '19)* (pp. 456–460).

Yu, R. (2019). Deconstructing the evolution of collaborative learning networks. In *Companion Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK '19)* (pp. 741–745).

Park, J., **Yu, R.**, Rodriguez, F., Baker, R., Smyth, P., & Warschauer, M. (2018). Understanding student procrastination via mixture models. In *Proceedings of the 11th International Conference on Educational Data Mining (EDM '18)* (pp. 187–197).

Yu, R., Jiang, D., & Warschauer, M. (2018). Representing and predicting student navigational pathways in online college courses. In *Proceedings of the 5th ACM Conference on Learning at Scale (L@S '18).*

PREPRINTS & REPORTS

Lastra-Anadon, C., Das, S., Varshney, K., Raghavan, H., & **Yu, R.** (2021). *How universities can mind the skills gap. Higher education and the future of work.* Center for the Governance of Change, IE University.

Yu, R., Scott, J., & Pardos, Z. (2021). *Unsupervised representations predict popularity of peer-shared artifacts in online learning environment.* ArXiv.

SOFTWARE

Yu, R., & Kung, C. (2020). *College success prediction and fairness evaluation* [Source code]. `https://github.com/renzheyu/fair-college-success-prediction`.

Calikus, E., Trimarco, J., Tseng, T., **Yu, R.**, de Unanue, A., & Sipka, A. (2019). *Understanding and reducing inequities in transportation in the West Midlands* [Source code]. `https://github.com/alan-turing-institute/DSSG19-WMCA-PUBLIC`.

ABSTRACT OF THE DISSERTATION

Equity-Oriented Educational Data Science

By

Renzhe Yu

Doctor of Philosophy in Education

University of California, Irvine, 2022

Professor Mark Warschauer, Chair

As educational institutions increasingly adopt digital tools for daily operations, unprecedented amounts of data are generated at different levels of the education system. The granularity of these big data makes it possible to understand and support educational processes in a data-informed, easy-to-scale manner, and educational data science (EDS) has emerged as a nascent field to realize this potential. This dissertation specifically focuses on the promise of EDS to address issues related to educational equity, a central theme of education research. To begin with, a two-dimensional taxonomy is presented to characterize equity-oriented EDS research – whether the work is focused on explanation or prediction, and whether the problem of interest takes a micro- or macro-level perspective of education research. The interaction of these two dimensions partitions the research space into four quadrants, and one empirical study in higher education contexts is presented to illustrate each quadrant. The first two explanatory studies leverage novel data sources (i.e., digital behavioral traces) to understand systematic sociodemographic gaps in 1) peer interaction experience in virtual learning environments (micro-level), and 2) academic engagement over time at the institutional level (macro-level). The latter two prediction-oriented studies investigate algorithmic fairness from the perspectives of 1) choice of data sources in online learning performance prediction (micro-level), and 2) use of sensitive attributes in early warning systems (macro-level). These studies highlight how EDS research can advance the understanding of existing educational inequalities and guide preventive action against potential inequities. Finally, recommendations for future research on equity-oriented EDS are discussed.

Chapter 1

Introduction

In the era of “datafication”, educational institutions are increasingly dependent on digital tools to organize and manage teaching and learning, student affairs, finance, and other aspects their daily routines (Selwyn and Gašević, 2020; Jarke and Breiter, 2019). These tools generate unprecedented amounts of data, often in real time, which opens up new possibilities of understanding what happens every day at different levels of the education system. For example, students’ learning behavior which was mostly captured through labor-intensive classroom observations in the past is now substantiated by their clicks, content submissions and other activities in digital learning platforms. Each of these activities is logged by the system and can easily amount to thousands of data points for just one student (Fischer et al., 2020). Besides their volume, granularity, and low costs of collection, most “big data” in education are generated in a minimally intrusive manner and therefore less prone to some of the common biases in traditional data sources, such as reflection bias and social desirability bias (Miller, 2012; Choi and Pak, 2005).

To translate “big data” into fine-grained knowledge about educational processes, new generations of computational resources and methodologies are needed. Since the late 2000s, learning analytics and educational data mining researchers have made significant contributions to the computational understanding of human learning (Romero and Ventura, 2020). Beyond the science of learning, less effort has been made to answer questions related to other aspects of education such as organizational and policy contexts, although novel data sources have been increasingly available with the potential to uncover more nuanced dynamics of educational practices. In this context, the inaugural conference on Educational Data Science (EDS)[1] was jointly hosted by Stanford University, American Educational Research Association (AERA) and ETS, and formally established the nomenclature for this nascent crossroads of education and data science. In parallel, one of the flagship AERA journals also published the first special issue on EDS and the editorial described EDS as "an umbrella for a range of new and often nontraditional quantitative methods (such as machine learning, network analysis, and natural language processing) applied to educational problems often using novel data" (McFarland et al., 2021). In addition to being interdisciplinary, EDS research tends to produce more actionable insights than previous education research because the scale and granularity of the analyses can directly augment the capacity of limited human resources in the education system. For instance, predictive algorithms can identify the most at-risk students just when they are struggling based on their behavioral traces in digital learning systems, which is almost impossible for a teacher or academic counselor who faces dozens to hundreds of students at the same time (Ekowo and Palmer, 2016).

The interplay between education and social inequalities places the pursuit of equity at the core of education research (Blanden, 2020). Because EDS research is a new addition, it remains an open question how researchers should approach educational equity with novel data and computational methods. A recent survey of the past decade of educational data mining research finds that only fewer than 20% of empirical studies incorporated students' sociodemographic information in their analyses (Paquette et al., 2020), which is a prerequisite for investigating equity issues. As EDS continues to become integrated into different strands

[1] https://iriss.stanford.edu/css/conferences/conference-educational-data-science

Table 1.1: A taxonomy of equity-oriented EDS research and mapping to chapters in this dissertation

	Explanatory	Predictive
Micro-level	Chapter 2	Chapter 4
Macro-level	Chapter 3	Chapter 5

of education research, it is crucial to establish the paradigms and guidelines for embedding equity-oriented themes.

This dissertation aims to illustrate *equity-oriented EDS* research through empirical studies. Given the breath of what can be seen as EDS research, Table 1.1 presents a heuristic two-dimensional taxonomy to help navigate the landscape. The first dimension (in rows) concerns the educational perspective of the focal problem. This can be 1) micro-level perspective that focuses on individual differences in learning and development, and the immediate environments, such as instruction and family, that shape these differences, or 2) macro-level perspective that highlights organizational, social, and policy contexts that shape systematic educational inequalities. The second dimension (in columns) includes two common paradigms of applied data science research: explanatory and predictive (Hofman et al., 2021). In the context of education and equity, explanatory research focuses on computational understanding of the sources and mechanisms of existing inequalities, whereas predictive research develops algorithms that guide preventive action against inequities in a reliable and ethical manner. The interaction of these two dimensions partitions the research landscape into four quadrants. Each of the following four chapters will exemplify a quadrant via an empirical study in higher education contexts:

- Chapter 2 presents a micro-level explanatory study that leverages students' postings in course discussion forum to understand individual differences in peer interaction experience in virtual learning environments.
- Chapter 3 is a macro-level explanatory study that highlights the use of large-scale

digital behavioral trace data to depict systematic inequalities in academic engagement patterns over time.

- Chapter 4 is micro-level, predictive in nature and examines algorithmic fairness in academic performance prediction as a function of the choice of data sources in online learning contexts (Yu et al., 2020).

- Chapter 5 presents a macro-level predictive study that scrutinizes the equity consequences of using sensitive attributes in early warning algorithms at the institutional level (Yu et al., 2021).

Finally, Chapter 6 presents high-level reflections on equity-oriented EDS research and highlights a few important directions for the future.

Chapter 2

Quantity versus Quality: How Do Peer Interactions in Online Discussion Forums Contribute to Academic Performance?

2.1 Introduction

Distance learning through fully online coursework is becoming a normal part of college students' learning experiences. As of Fall 2016, almost one third of college students in the U.S. took at least one fully online course (Seaman et al., 2018), and the more recent COVID-19 pandemic further moved the entire college experience to the online space. Yet, online learning is associated with unique challenges, one of which being the lack of interpersonal interactions. The physical distance between individuals and the nature of asynchronous communications that dominate the majority of online courses often lead to reduced psychological connections between students and the learning community, which may demotivate learners from optimal engagement and further result in lower academic performance or early course withdrawal (Bettinger et al., 2017; Xu and Jaggars, 2013).

In light of the importance of interpersonal connections in engaging learners, researchers and practitioners have proposed an array of instructional practices conductive to strengthening interpersonal interactions and fostering social presence – the degree to which a person is perceived as a “real person” in mediated communication – more visibly and intentionally in an online setting (Tu and McIsaac, 2002; Pacansky-Brock et al., 2020; Ragusa and Crampton, 2018). Among these efforts, online discussion forum is one of the most widely adopted tools. For students, communication via asynchronous posting does not involve a steep learning curve and makes it convenient to archive, retrieve, and reflect on ideas at any time (Balaji and Chakrabarti, 2010). For instructors, assigning and organizing online discussions is handy especially with the advent of modern learning management systems (LMS). As a result, discussion forums continue to serve as the central medium in online courses to achieve interpersonal, especially student-student interactions. Across different institutional settings, subject matters and course design contexts, however, findings about the relationship between forum-based peer interaction and learning outcomes have been mixed (Picciano, 2002; Kent et al., 2016). Along this line of research, the current study presents an in-depth examination of how exposure to peer responses in discussion forums may support learning at the individual student level.

We contribute to the literature on peer interaction and online learning in two respects. First, going beyond correlational analysis, we employ a quasi-experimental instrumental variable approach that leverages a randomized grouping instructional design across multiple course offerings and the temporal nature of posting records, in order to estimate the causal effects of peer influence on learning gains. Second, we tease out the quantity and quality of peer interaction, which respectively map to two intertwined theoretical benefits, enhancing social presence and facilitating knowledge construction, and therefore may contribute deeper understanding of the mechanism of peer interaction.

2.2 Related Work

Learning theories that emphasize the integral role of social interaction in the learning process are often traced back to Vygotsky (1978), who famously situated learning in the interaction between people mediated by tools and signs, which in turn shapes how learning and world views become internalized. With the later application of computers and internet to education in the following decades, some research on computer-supported learning started to decipher how social interactions among online learners contribute to learning and what instructional strategies can mediate this process (Stahl et al., 2014). More recently, the concept of connectivism was coined as a new learning theory in the digital era (Siemens, 2005). The underlying assumption is that knowledge lies within the connections between nodes of information. Learning therefore occurs through social interactions within a knowledge community. While these theories were developed in different contexts, they provided a common ground from which the widespread effort of promoting peer interaction in online learning environments builds.

Learning scientists and educational psychologists have explicated multiple channels through which peer interaction can benefit learning. Specific to discussion scenarios, effective peer interaction has been regarded as one approach to boosting social presence, i.e. learners' ability to project themselves socially into a community of "real people" (Garrison and Arbaugh, 2007). An adequate level of social presence then secures higher levels of motivation and engagement (Richardson et al., 2017). Likewise, asynchronous online discussions enable students to share own ideas and read and reflect on each other's thoughts. Through this process, they learn from peers and build knowledge collectively (Stahl et al., 2014).

Research on distance and online learning has empirically examined the role of peer interaction in a variety of learning contexts. Overall, there is some consistent evidence that effective interaction between students is associated with objective and subjective measures of learning outcomes including course grades (Kent et al., 2016), knowledge construction (Wang and Noe, 2010) and student satisfaction (Ho and Swan, 2007), but most of the studies measure peer interaction as individual participation in interactive activities (e.g. discussions) instead of actual influence from peers. For exampleWise and Cui (2018) find that making forum contribution is associated with passing the course in the context of MOOCs. In addition, most existing studies performs correlational analysis, including the relative few of them that examine peer influence (Kent et al., 2016). One methodological limitation of this analytical paradigm comes from the reflexive nature of peer interaction. While a student's learning is affected by the classmates she interacts with, she simultaneously exerts her influence on their performance. In this case, the observed amount of peer influence is a sum of the actual influence from her peers and the contribution of her own qualities by way of her direct influence on those peers. The recent availability of granular traces of online learning behavior can inform partial solutions to this issue. In the most closely relevant studyBettinger et al. (2017) disentangle each student's inherent tendency to interact (prior to exposure to reflexive influence) from discussion logs, and use this estimated exogenous quality to instrument peer interaction and estimate causal effects.

This study builds upon the theoretical affordances of peer interaction and the empirical methods that aim at causal estimates of peer effects. From the theoretical perspective, we focus on two aspects of peer responses to forum posts: quantity and quality. Receiving more responses from peers is likely to strengthen a focal student's psychological connection to the course via increased social presence, and the quality of these responses might facilitate reflection on her previous ideas and foster higher-order knowledge construction. Before causally estimating peer effects, we also intend to gain a descriptive understanding of to what extent discussion behavior differs across student subpopulations, because this would inform us of the student-level characteristics to examine at the time of peer effect estimation. Therefore, the following research questions are proposed:

1. Do students' individual characteristics relate to their posting behavior and their chances of attracting peer responses in college online courses?
2. To what extent does receiving more peer responses affect course performance?
3. To what extent does the quality of peer responses affect course performance?

2.3 Data and Research Context

2.3.1 Course Context

This study is focused on repetitive iterations of two fully online courses offered to residential college students at a four-year public institution in the United States. Both courses were gateway courses for lower-division students majoring in public health. Their topics were complementary in that Course 1 introduced basic concepts and principles in the field while Course 2 presented case studies in practice to demonstrate how the principles are applied. Both courses lasted for ten weeks (an academic quarter) and were taught repetitively each quarter by the same instructor. The course design (introduced below) was virtually identical across these two courses and across different quarters. We initially examined both courses in three academic terms in 2017 (Winter, Spring and Fall), summing to a total of six separate classes[1]. This multi-class sample would help to improve the generalizability of our findings.

The majority of course activities were organized in the Canvas LMS. Within each class, there were weekly requirements of watching lecture videos, finishing quizzes, authoring posts in

[1] In the remainder of this paper, we continue use the hierarchy of "course" and "class" to clarify our data structure where necessary (two courses with three classes each).

the discussion forum and synthesizing course material. In addition, there were: an individual presentation with peer reviews, a research paper, a midterm exam, and a final exam. For forum discussion assignments, students were given a specific question related to the course content each week. They were required to post their opinions over the question in 150-200 words by Wednesday and make a comment on a peer post with the same length requirement by Sunday. Students could engage in further discussions based on these posts and comments, but this was not required. Within the first two weeks where students were still actively adding or dropping courses, each student was open to all their classmates' posts. Starting from Week 3, the instructor randomly assigned all the enrolled students into groups of around 10 (in four classes) or around 20 (in two classes) and each student could only see and respond to the posts from their group members. Students received an overall discussion score (out of 10) each week for the original post and the reply they authored. This score was based on whether the student made the posts as required and the quality of her posts (e.g., exhibiting sufficient reasoning and new ideas around the topic). Detailed rubrics (Table 2.1) of this discussion assignment was included in the syllabus and all these forum postings made up 9% of the final course grade. The random group assignment was implemented using a built-in function of Canvas. In Course 2 of the Fall quarter, however, the function failed to work for a few weeks due to technical glitches, and students were exposed to all of their classmates' posts like in the first two weeks. This glitch undermined the random group assignment, the basis of our analysis, so we dropped this class from our final dataset, leaving only five classes.

2.3.2 Data and Key Metrics

We acquired the dataset of 21,410 raw discussion posts from the five classes. Each entry contained the message content (including title), posting time, author ID, course ID, parent post ID (if any), among others. This allowed us to identify the time-stamped response structure among students, i.e. who responded to whom at what time. Additionally, we got student-

Table 2.1: Instructor's rubric for discussion posts (including original posts and replies)

Points	Category	Explanation
5	Thought provoking or challenging new idea informed by reading or lesson	This rating is given to posts that, in addition to responding to the prompt, present a new idea based on information from a scholarly source. The new idea must be clearly marked and include an in-text citation for the scholarly source (i.e., peer reviewed journal article) and the reference at the end of the post. The new idea must expand on the information in the cited source in some way rather than repeat the information in the source.
3	Opinion based on information from reading or lesson	This rating is given when a person writes a fact-based forum post. The facts could come from a lesson or a chapter from the textbook, or another scholarly external source.
2	Answered as required, but nothing more	This rating is given when a post answers all parts of the question but does nothing more. May show an absence of depth or thought.
0	Inappropriate or insufficient postings	This rating is given to posts that do not meet my grading requirements. Used for: agreement without new substance, general humor, posts that do not fit into the current discussion.

level information including group ID, detailed course grades, demographic information and academic history. We matched these datasets and only kept the students who finished the course with a valid grade as well as the posts among them. This process left us with 1,091 students and 20,996 posts. Note that the 1,091 students across five classes included only 989 unique student IDs because some students enrolled in more than two classes (e.g., Course 1 and 2, or failing one course and retaking it). Because in this paper we use within-class variation to estimate peer effects and there is no duplicate student within each class, the cross-class sample correlation due to multiple enrollment would not induce bias to our estimates. As such, we treat observations of the same student in multiple classes as separate students in the remainder of this paper, unless otherwise clarified.

Table 2.2 reports the summary statistics of student-level variables across the five classes. The class size (number of students) ranged from 183 to 275. In general, student profiles were highly diverse within each class and this diversity was similar between classes. More than half of the students came from various disadvantaged backgrounds, including underrepresented ethnic groups, low-income family with no previous college attendees and/or non-English-speaking environment at home. This pattern is aligned with the general student demographics of this university.

The second and the third parts under Panel A of Table 2.2 include selected measures of students' behavior and outcomes in the class. The total number of posts authored counts anything a student posted in the discussion forum, including both initial posts and replies to others' posts, while the total number of replies received only counts replies authored by other students (i.e., excluding self-replies). Based on the course design, students were required to make 20 posts (10 initial ones and 10 replies) throughout the ten-week term, which was slightly greater than what students actually finished on average, with small standard deviations (around 3). On the other hand, the average volume of peer responses received per student (around 9.5) was also aligned with the course requirement, but its dispersion was as large as half of its mean value. These figures combined suggest that while most students might not have engaged in deeper interactions than what was required, they might have read others' initial posts and choose whom to respond to. This variation allows us to examine the effects of the quantity of peer interactions. We also report metrics of students' posting behavior during the first two weeks of each term, which we deem as reflecting their prior study habits. The average post length was mostly around 180 words except for one class and this was largely shaped by the course requirement (150-200 words). Students in two classes were generally more active, authoring their posts more than one day ahead of the deadline, while in the most procrastinating class, students waited on average until four and a half hours before deadline. In terms of outcomes, we use the total course score instead of GPA because it can capture more nuances of students' performance and can be easily converted to the 0-4 scale. Lastly, the discussion points reflect the quality of peer interactions.

2.4 Modeling Strategy

We use different empirical strategies to answer each of the three research questions. Below we briefly explain the empirical model for each of our analyses.

To understand systematic differences, if any, in discussion patterns across subgroups of students (RQ 1), we examine the relationship between student-level characteristics and students' volumes of posts and replies in the classes. We run the following linear regressions model:

Table 2.2: Summary statistics of student characteristics in each class

Class	Winter 1	Winter 2	Spring 1	Spring 2	Fall 1
(A) Continuous and binary variables (mean/SD)					
Personal background					
Age	20.6	20.4	20.6	20.1	20.1
	(1.7)	(1.9)	(1.9)	(1.6)	(2)
Female	0.735	0.756	0.711	0.781	0.71
	(0.44)	(0.43)	(0.45)	(0.41)	(0.45)
Low-income family	0.413	0.444	0.437	0.445	0.388
	(0.49)	(0.5)	(0.5)	(0.5)	(0.49)
First generation college student	0.577	0.574	0.597	0.582	0.525
	(0.5)	(0.5)	(0.49)	(0.49)	(0.5)
Transfer student	0.163	0.324	0.141	0.131	0.128
	(0.37)	(0.47)	(0.35)	(0.34)	(0.34)
SAT total score[2]	1715	1708	1686	1685	1728
	(216)	(219)	(205)	(214)	(206)
College GPA cumulative[3]	2.93	2.94	2.93	2.91	2.88

[2] SAT scores are not recorded for many transfer students. Overall missing rate: 14%.

[3] Students in their first term of school do not have cumulative GPAs. Overall missing: 5.7%.

	(0.58)	(0.58)	(0.54)	(0.57)	(0.52)
Current units attempted	14.6	14.7	14.3	15	14.8
	(2.8)	(2.6)	(3.2)	(2.6)	(2.7)
Course behavior					
Total # of posts authored	19.4	19.2	19.2	18.6	19
	(2.4)	(2.9)	(3.1)	(3.5)	(3)
Total # of replies received	9.52	9.51	9.45	9.08	9.03
	(4.9)	(5.3)	(6.6)	(5.2)	(5.9)
Total # of posts authored (weeks 1-2)	3.85	3.91	3.92	3.84	3.93
	(0.89)	(0.62)	(0.66)	(0.56)	(0.77)
Avg # of words of posts authored (weeks 1-2)	182	183	182	206	181
	(11)	(14)	(13)	(25)	(14)
Avg hours before deadline of posts authored (weeks 1-2)	12.5	30.3	4.31	29.5	12.1
	(56)	(42)	(82)	(40)	(107)
Course performance					
Avg. discussion points (out of 10) (weeks 3-10)	9.55	8.58	8.56	8.28	8.3
	(1.2)	(1.3)	(1.8)	(1.8)	(1.7)
Total score (out of 100)	88.4	95.4	81.6	91.3	81.6
	(9.6)	(8)	(16)	(10)	(15)

(B) Categorical variables (%)[4]					
Personal background					
Ethnic group					
American Indian / Alaskan Native	0.00	0.00	0.47	0.00	0.88
Asian / Pacific Islander	55.26	53.45	52.11	51.91	44.69
Black, non-Hispanic	1.58	6.91	2.82	4.92	5.31
Hispanic	25.26	24.00	20.66	30.05	27.43
International student	5.79	2.55	7.51	2.19	8.85
White, non-Hispanic	9.47	12.00	9.86	8.74	10.18
N	191	275	215	183	227

[4] Categories equivalent to missing values (e.g., declined to state) are removed from the table.

$$y_{ic} = \beta_0 + \beta_1 \ X_{ic} + \theta_c + \varepsilon_{ic} \tag{2.1}$$

where y_{ic} is the outcome of interest for student i in class c, including the number of posts authored and the number of replies received. We look at these measures for both the entire term and weeks 3-10 where enrollment has largely been finalized. X_{ic} is the same set of student-level covariates in Table 2.3. Because randomized grouping occurred within classes, we add θ_c, the class-level fixed effect. Finally, the model also includes the random error term ε_{ic}.

As mentioned above, self-selection into peer contexts is a common challenge to causal estimates of peer effects (Bettinger et al., 2016). In our analysis, the randomized group assignment should eliminate this issue. To test this assumption, we performed the following randomization checks. Within each class, we fitted ANOVA models (for continuous variables) or chi-squared tests (for categorical variables) on an array of student-level background characteristics[5] against the random groups. These tests checked whether the covariates have identical means or distributions across groups. Table 2.3 reports their results, including relevant statistics and their statistical significance. Most of these tests fail to reject the null hypothesis, suggesting successful randomization in general. For Course 1 offered in Winter, however, the groups were not balanced on three attributes: college GPA cumulative, current units attempted and average post length in the first two weeks.

For any individual student, how much does she benefit from receiving peer responses in terms of learning gains? This is the motivation of RQ 2 and RQ 3. The general goal of this inquiry

[5] Here we recode ethnicity into a binary variable "underrepresented minority" which includes American Indian / Alaskan Native, Black and Hispanic.

Table 2.3: Balance checks for randomized grouping in each class

Class	Winter 1	Winter 2	Spring 1	Spring 2	Fall 1
AVOVA (F-statistic reported)					
Age	0.835	0.962	1.032	0.967	1.137
SAT total score	1.196	0.974	1.016	0.845	1.180
College GPA cumulative	1.921**	1.044	0.879	1.022	0.427
Current units attempted	1.598*	1.230	1.551	0.968	1.575
Total # of posts authored (weeks 1-2)	1.117	1.152	0.606	0.997	1.321
Avg # of words of posts (wks 1-2)	1.844**	1.296	1.549	1.459	1.155
Avg hrs before deadline of posts (wks 1-2)	0.801	1.164	0.620	1.043	0.847
χ^2 *test*					
Female	12.84	27.61	5.763	16.92	8.162
Underrepresented minority	19.13	25.52	13.15	19.82	8.351
Low-income family	20.64	22.19	7.647	23.82	6.665
First generation student	17.01	29.66	8.139	10.06	7.624
Transfer student	17.55	31.20	2.791	15.33	6.915
$N_{students}$	191	275	215	183	227
N_{groups}	18	28	10	20	11

Each ANOVA/χ^2 test is performed on one variable across all discussion groups in each class.
* 0.1, ** 0.05, *** 0.01

is to estimate the following equation:

$$y_{ic} = \alpha + \beta P_{ic} + \gamma Q_{ic} + \delta X_{ic} + \theta_c + \varepsilon_{ic} \tag{2.2}$$

where y_{ic} is the measure of student i's academic performance in class c, in our case the final course grade (out of 100). P_{ic} is some measure of peer interaction that targets student i in class c. We restrict any specific P_{ic} to weeks 3-10 for each class because, as mentioned earlier, this period is when changes to course enrollment are minimal and random groups are in effect. Q_{ic} is the measure of student i's own interaction that targets other students in weeks 3-10. X_{ic} and θ_c are the same as in Equation 2.1. As discussed in previous sections, causal estimates of peer effects are challenged by the reflexive nature of peer interactions. Interaction that student i receives from peers is not predetermined: it is affected by how much interaction these peers receive from their peers, who might include student i herself. To address this issue, we use instrumental variable approach to carve out exogenous variations from the endogenous peer interaction.

We first look into the quantity of peer interaction (RQ 2), where P_{ic} is the total number of replies that student i receives through weeks 3-10:

$$P_{ic} = \sum_{w=3}^{10} \sum_{j \neq i} |S_{jiwc}| \tag{2.3}$$

where S_{jiwc} is the set of posts that are authored by student j in reply to student i during week w of class c, and $|S_{jiwc}|$ is the size of this set. As suggested by Table 2.2, the variation of this measure results primarily from students selecting posts to reply to within their own group. While this selection can be influenced by ongoing peer interactions, which adds to the endogeneity of P_{ic}, students' intrinsic characteristics may also play a significant role given that most students might not have sufficient knowledge about their group members outside of the online course space. For example, students might be more likely to respond to same-sex classmates. Because these characteristics are predetermined and the group assignment is random, we use the extent to which groupmates are initially different along these dimensions from the focal student, which is exogenous to the peer interaction processes, to instrument the number of replies this student received. Specifically, we calculate for each covariate X_{icm} (as in Equation 2.2) the average difference (AD) between student i and her group members:

$$AD_{icm} = \begin{cases} \frac{\sum_{G(j)=G(i), j \neq i} (X_{jcs} - X_{ics})}{|G(i)|-1} & \text{for numerical } X_{icm} \\ \frac{\sum_{G(j)=G(i), j \neq i} \mathbf{1}(X_{jcs} = X_{ics})}{|G(i)|-1} & \text{for categorical } X_{icm} \end{cases} \tag{2.4}$$

where $G(i)$ indicates the group that student i is assigned to and $|G(i)|$ is the group size. Using this equation over all covariates m, we get the vector AD_{ic} which instruments the quantity of peer interaction. Here the underlying assumption is that, on each student characteristic included, students who are more similar will be more likely to interact, or the opposite. As a result, a student who is in the "majority" (or "minority") of the group would be expected to receive higher (or lower) volumes of peer responses. Likewise, Q_{ic} in this specification is the total number of posts that student i authors through weeks 3-10 and is instrumented in a similar manner.

When it comes to the quality of peer interaction (RQ 3), we use a slightly different and more complicated measure. For each reply that student i receives over weeks 3-10, we measure its quality using the discussion score this reply's author receives for the corresponding week. Then we average this measure across all the replies to student i:

$$P_{ic} = \frac{\sum_{w=3}^{10} \sum_{j \neq i} d_{jwc} \, |S_{jiwc}|}{\sum_{w=3}^{10} \sum_{j \neq i} |S_{jiwc}|} \tag{2.5}$$

where d_{jwc} is the discussion points that student j receives for week w of class c and the dominator is the sum of this response quality metric weighted by the number of times that j replies to i. The denominator is the same as Equation 2.3, which standardizes the weighted sum in the numerator and therefore rules out the effect of quantity of peer responses. Note that the quality of student j's reply is recursively a function of the quality of others' replies to j, thus inducing endogeneity. We follow the approach described in (Bettinger et al., 2016) and capitalize on the sequential nature of our timestamped discussion data to construct instrumental variables for P_{ic}. Intuitively, the instrument is student j's inherent propensity to author high-quality posts, which is exogenous but associated with the actual discussion points d_{jwc}. It comes from the following equation:

$$d_{jwc} = \delta d_{j,w-1,c} + \rho \frac{\sum_{k \neq j} d_{kwc} \, |S_{kjwc}|}{\sum_{k \neq j} |S_{kjwc}|} + \mu_{jc} + \eta_{jwc} \tag{2.6}$$

where the second term on the right-hand side is measuring the average quality of peer responses that student j's receives in week w of class c. In this specification, the μ_{jc} term captures the exogenous variation in response quality that is invariant to the dynamics of peer interaction. We employ the two-stage least-squares first-differenced estimator (FD2SLS) (Anderson and Hsiao, 1981) for Equation 2.6, using data from the entire term, and get the estimated coefficient, $\hat{\delta}$ and $\hat{\rho}$. Then we have:

$$\widehat{\mu_{jc}} = \frac{1}{9}\sum_{w=2}^{10}(d_{jwc} - (\widehat{\hat{\delta}} d_{j,w-1,c} + \hat{\rho}\frac{\sum_{k\neq j} d_{kwc}\,|S_{kjwc}|}{\sum_{k\neq j}|S_{kjwc}|})) \tag{2.7}$$

where $\widehat{\mu_{jc}}$ is used to instrument d_{jwc} for all w. Substituting this for d_{jwc} in Equation 2.5, we get the final instrument variable for P_{ic}, the average quality of peer responses to student i. Q_{ic}, in this case the weighted average of student i's discussion scores across weeks 3-10, is instrumented using $\widehat{\mu_{ic}}$ computed by Equation 2.7.

2.5 Empirical Results

2.5.1 RQ1: Who authored more initial posts and received more responses?

Equation 2.1 examines who posted more and who received more responses across the five classes, and the results are presented in Table 2.4. The first two columns show that younger and female students were more active in authoring posts, whether we look at the entire quarter or the period after students were randomly assigned into groups. Underrepresented minorities contributed significantly less to the discussion forum, but family background or transfer status had no relationship with posting behavior. Academically, students with higher cumulative GPAs posted more but those with higher SAT scores posted less. Finally, those who posted earlier and wrote more in the first two weeks, a signal of more serious attitudes towards coursework, showed a positive correlation with the number of posts.

Table 2.4: Factors associated with authoring posts and receiving peer responses

	# of posts authored		# of replies received	
	Full term (1)	W3-W10 (2)	Full term (3)	W3-W10 (4)
Personal background				
Age	-0.154*	-0.136*	-0.0922	-0.165*
	(0.079)	(0.0736)	(0.122)	(0.0917)
Female	0.383*	0.339*	1.28***	0.987***
	(0.216)	(0.19)	(0.366)	(0.265)
Underrepresented minority	-0.758***	-0.604***	-1.39***	-0.955***
	(0.235)	(0.209)	(0.371)	(0.272)
Low-income family	0.175	0.151	0.0162	0.214
	(0.189)	(0.168)	(0.363)	(0.27)
First generation student	-0.0924	-0.0866	0.196	-0.0374
	(0.19)	(0.171)	(0.379)	(0.275)
Transfer student	0.04	-0.0506	-0.633	0.107
	(0.326)	(0.301)	(0.636)	(0.465)
SAT total score	-0.00191***	-0.00148***	-0.00367***	-0.00269***
	(0.000573)	(0.000509)	(0.000965)	(0.000636)
Cumulative college GPA	1.05***	0.89***	1.6***	1.19***
	(0.211)	(0.186)	(0.381)	(0.25)
Current units attempted	0.0591*	0.0425	-0.08	-0.0707
	(0.0327)	(0.0286)	(0.0645)	(0.0453)
Course behavior				
Total # of posts (w1-w2)			0.658**	0.601***
			(0.264)	(0.193)
Avg # of words per post (w1-w2)	0.0204***	0.0203***	0.0109	0.0115
	(0.00708)	(0.00607)	(0.0102)	(0.00738)
Avg posting time before deadline	0.0048**	0.00377**	0.0251***	0.0118***
(w1-w2; in hours)	(0.00213)	(0.0017)	(0.00792)	(0.00411)
Class FE	Yes	Yes	Yes	Yes
N	1,028	1,028	1,028	1,028
R^2	0.131	0.125	0.198	0.161

Standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01

When scrutinizing peer responses that students received (the last two columns), the effects of most variables are qualitatively similar to those on the volume of posts. Female students received more peer responses, while under-represented minorities again garnered less attention from their classmates. SAT score and college GPA still correlated with a student's "attractiveness" in opposite directions. Post length early in the course did not predict the number of replies received, but the volume and earliness of posting did. All these results from Table 3 combined suggest that in our college context, peer interaction behavior differed systematically across subgroups of students.

2.5.2 RQ2: What is the effect of the quantity of received replies on course performance?

We further investigate how the quantity of peer interaction, measured by the number of peer responses (Equation 2.3), affected individual student's course outcomes. Following Equation 2.4, we instrument this quantity measure using the average peer difference in a series of student characteristics. In practice, we only include those characteristics that, at the individual student level, significantly predicted the number of peer responses (Column 4, Table 2.4), including age, gender, ethnicity[6], SAT total score, college GPA cumulative, number of posts and timing of posting in the first two weeks. This selection aims to minimize the possibility of including weak instruments. Moreover, we divide these instrument variables (IVs) into two groups based on the nature of student attributes on which they are built. Background IVs include the average peer difference in pre-class attributes (age, gender, ethnicity, SAT total score and college GPA cumulative) and behavioral IVs refer to the average peer difference in students' early-course activities (number and timing of posts in the early weeks). A hypothesis is that the latter group might be stronger IVs because

[6] Here the raw categories (as in Table 2.2) instead of the combined "underrepresented minority status" are used to calculate the average peer difference.

they were directly observable in the online course space and therefore immediately shaped students' behavior. For example, an early poster within a group, who had a larger average peer difference in the timing of posting, may intuitively get more responses.

Based on the discussions above, Columns (1)-(4) of Table 2.5 reports the estimates of peer interaction from four separate two-stage least-squares (2SLS) regressions. The first three specifications compare the inclusion of background IVs, behavioral IVs and the combination of both. Regardless of the specific set of IVs being used, the estimates suggest that receiving more replies from peers significantly improved focal student's performance in the course. Consistent with the foregoing hypothesis, behavioral IVs are stronger instruments of the quantity of peer responses than either background IVs alone or the combination of both, reflected in the largest F statistic of the first-stage regressions. As such, we only use behavioral IVs in our final specification (Column (4)) and control for individual background characteristics. Adding these control variables reduces the first-stage F statistic but the IVs are still strong. As a comparison, Column (5) reports the Ordinary Least Squares (OLS) estimates without using IVs, suggesting a smaller observed association than the IV estimate. These two columns combined reveal that the quantity of peer interaction received did affect individual learning outcomes, and that on average, receiving an additional response from other students increased the final course score by 1.14 points (out of 100), or one ninth of a letter grade.

2.5.3 RQ3: What is the effect of the quality of received replies and course performance?

The previous subsection suggests that the quantity of peer responses matters. What about the quality? Using Equations 2.6 and 2.7 to instrument the quality measure, i.e. average peer discussion scores (Equations 2.5), we illustrate the 2SLS estimation results in Columns

Table 2.5: Effects of quantity of peer responses received on final course score (out of 100)

	(1)	(2)	(3)	(4)	(5)
# of replies received (w3-w10)	-0.291	1.29***	0.87***	1.14***	0.169**
	(0.567)	(0.349)	(0.249)	(0.362)	(0.0681)
# of posts authored (w3-w10)	5.37***	2.66***	3.16***	2.84***	3.01***
	(0.99)	(0.393)	(0.335)	(0.372)	(0.104)
Age				-0.194	-0.346**
				(0.196)	(0.173)
Female				-1.39**	-0.374
				(0.696)	(0.571)
Underrepresented minority				1.19	0.0845
				(0.729)	(0.596)
Low-income family				-0.349	-0.0702
				(0.595)	(0.54)
First generation student				-0.287	-0.283
				(0.605)	(0.555)
Transfer student				0.834	0.997
				(1.19)	(1.09)
SAT total score				0.00984***	0.00703***
				(0.00179)	(0.00145)
College GPA cumulative				0.766	2.18***
				(0.685)	(0.492)
Current units attempted				0.114	0.0535
				(0.112)	(0.0983)
Class FE	Yes	Yes	Yes	Yes	Yes
First-stage F	1.65	24.9	7.79	21.5	-
Background IV	Yes	No	Yes	No	No
Behavioral IV	No	Yes	Yes	Yes	No
N	1,054	1,090	1,053	1,028	1,029

Standard errors in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$

Table 2.6: Effects of quality of peer responses received on final course score (out of 100)

	(1)	(2)	(3)
Avg peer discussion points (out of 10) (w3-w10)	0.022	0.0513	-0.0798
	(0.11)	(0.11)	(0.107)
Avg discussion points (out of 10) (w3-w10)	5.83***	5.24***	5.2***
	(0.17)	(0.184)	(0.183)
Age		-0.175	-0.174
		(0.158)	(0.159)
Female		-0.643	-0.607
		(0.517)	(0.521)
Underrepresented minority		0.0249	0.0153
		(0.538)	(0.542)
Low-income family		-0.0151	-0.0138
		(0.489)	(0.493)
First generation student		-0.412	-0.436
		(0.503)	(0.507)
Transfer student		-0.31	-0.295
		(0.991)	(0.999)
SAT total score		0.00525***	0.00519***
		(0.0013)	(0.00131)
College GPA cumulative		1.44***	1.46***
		(0.452)	(0.455)
Current units attempted		0.0813	0.084
		(0.0891)	(0.0898)
Class FE	Yes	Yes	Yes
First-stage F	6684	6492	-
N	1,076	1,015	1,015

Standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01

(1)-(2) of Table 2.6. It is somewhat surprising to see that, although the instruments were extremely highly correlated with the quality measure, the estimated effect of the latter on the course outcome was not significantly different from zero. Controlling for individual characteristics only slightly reduced the first-stage F-statistic, and the estimated main effect was even closer to zero. Again, the last column (3) reports the OLS estimates for comparison, and it is obvious that simple correlation did not lend support to a significant effect either. In other words, overall, the quality of peer interaction had no effect on individual student's learning outcomes.

2.6 Discussion and Conclusion

Drawing on fine-grained records of discussion forum posts in five fully online classes, we separately estimated the causal effects of the quantity and quality of received peer responses on individual course outcomes. To better contextualize those insights, we started by analyzing the systematic differences, if any, in posting behavior across student subpopulations.

We found that students' demographic characteristics were associated with their posting records. For example, disadvantaged students seemed to be marginalized and more "invisible" in the discussion forum evidenced by fewer posts they authored as well as fewer responses they received from peers. This held true mainly for underrepresented minorities and low achievers in our research contexts but might apply to other disadvantaged groups as well in alternative course contexts, such as female students in a dominantly male classroom (not the case in our context). These correlational insights informed us to control for student background characteristics in the following estimation of peer effects. When comparing the quantity and quality of peer interaction, we found that receiving more peer responses generally had a strong positive effect on individual students' learning outcome, but how well these responses were written did not matter. In our hypothesis, the quantity of peer responses can potentially realize the social benefits of peer interaction, e.g., sustaining psychological connections to the learning community (Kreijns et al., 2013), whereas the quality of those responses should illustrate the cognitive functions of peer interaction (Stahl et al., 2014). Following this framing, our results suggest that forum discussions afford to increase social presence and therefore motivation and engagement, but they cannot effectively bolster higher-order knowledge construction processes. Reflecting on the specific course context, we assume that the social presence mechanism might work because students received notifications when others responded to them in the discussion forum. The estimated effect of receiving on additional response leading to one ninth of a letter grade is also consistent with the similar study that we closely follow (Bettinger et al., 2016). The failure of cognitive mechanism to work, on the other hand, may actually be explained by the fact that most students did not post more than what was required (as seen in Table 2.2) – they might merely "passively" participate in the discussion forum and might not necessarily read and reflect on the peer responses received, so the cognitive benefits were absent. Therefore, the implication for instructors is that standard required discussion assignments may not necessarily take full advantage of peer interaction in the online world. One simple strategy to address this might be getting themselves involved in students' discussions (Jaggars and Xu, 2016). Moreover, with emerging learning technologies, there are more possibilities of fulfilling the potential of peer intelligence (Jayaprakash et al., 2017; Kent et al., 2016), which require online instructors to put more effort into group-based instructional designs. For example, students can be encouraged to curate, socialize around and remix their own multimedia artifacts, in which case they are can form a more connected learning community, augment instructional materials and learn from each other through intensive peer interaction.

The current study has several limitations and directions of future work. First, the instructor was lenient in practice and most students got similarly high scores on their discussion posts (see Table 2.2). This limited variation contributes to the extremely large first-stage F-statistic in Table 2.6 and may mask the real effects of the quality of peer responses. To address this concern, we will use natural language processing (NLP) techniques to capture various linguistic features of the discussion posts. Second, we only look at the effect of peer interaction on the immediate course grade, but there might alternative outcome measures that benefit from the peer community. We will be investigating longer-term and/or socioemotional outcomes (e.g., sense of belonging, motivation, self-efficacy) as well to comprehensively understand the affordances of peer interaction. Third, peer interaction may function differently across student subpopulations, and a systematic heterogeneity analysis of the estimated main effects is our immediate next step. Fourth, although we include five different classes, they are homogeneous in terms of subject matter and course design. To generalize our current findings, future work will examine other courses with substantially different characteristics but similar discussion arrangements.

Chapter 3

Disruption and Resilience of Academic Engagement in the Pandemic: A Large-Scale Longitudinal Analysis of Digital Learning Behavior

3.1 Background

Since early 2020, the COVID-19 pandemic has significantly changed the landscape of higher education. The initial outbreak of the pandemic led to an emergency shift to online learning. The following academic terms continued to witness unprecedentedly prevalent fully online college experience across the globe. Even after campuses reopened and students came back, the fluctuation of local caseload still led to much more flexible instructional modalities than before. The increased share of virtual learning experience during the pandemic came with common challenges associated with online learning such as reduced psychological connection and higher requirements of self-regulation (Xu and Xu, 2020). In addition, the lasting impacts of COVID-19 on the economy and public health exposed students to other obstacles on their way to meet their educational goals, such as increased family responsibilities, uncertain job prospects, and health-related anxieties. Even worse, these negative impacts were not evenly distributed across students from different backgrounds and would exacerbate existing inequities in the education system. Existing research has found that students from marginalized groups suffered more challenges in academic progress, physical and mental health, financial security, and job prospects, compared to their peers (Aucejo et al., 2020; Cao et al., 2020; Huckins et al., 2020; von Keyserlingk et al., 2021; Means and Neisler, 2020; Rodriguez-Planas, 2022; U.S. Department of Education Office for Civil Rights, 2021). For example, during the initial disruption in Spring 2020, students from racial minority groups reported more challenges in finding an appropriate physical environment for online learning and in feeling motivated to participate in classes (Means and Neisler, 2020); low-performing students leveraged flexible grading policies (e.g., pass/no pass, incompletes) more than their high-income peers and had more concerns about their financial aid (Rodriguez-Planas, 2022); LGBTQ+ students struggled much more with their mental health and well-being (U.S. Department of Education Office for Civil Rights, 2021).

Because the pandemic has brought unexpected challenges to the education system, these empirical findings can provide important insights for instructors and policymakers to make up for the learning loss and support more equitable educational experience in the post-pandemic era. However, due to the recency of COVID-19, these insights are still insufficient. Most existing studies collect students' self-reports of their experience through surveys or examine academic outcomes via institutional records. While these data sources can collect students' subjective feelings and static states at a few critical time points (e.g., end of term), they do not capture the dynamics of students' day-to-day educational experience, which is highly relevant for policymaking in the context of a rapidly evolving pandemic. In addition, existing research largely focuses on the initial disruption when the COVID-19 broke out, but the longer-term impact of the pandemic on (higher) education is less clear. In this context, the current study uses large-scale, fine-grained behavioral trace data from learning management systems (LMS) at a large minority-serving institution to understand how college students' academic engagement changed through the pandemic. Importantly, the analyses focus on the experience of different socio-demographic groups over a wide time window of two years after the outbreak. As prior research suggests, academic engagement provides an important lens through which to understand educational outcomes and achievement gaps, and the granularity of behavioral trace data allows for capturing engagement in a scalable and authentic manner (Fischer et al., 2020; McCormick et al., 2013). This becomes particularly useful for analyzing student during the pandemic because they spent less time on campus than before and their experience was less directly observable by instructors, counselors, and other educational practitioners.

Specifically, this study aims to answer the following research questions:

1. How do college students' academic engagement patterns change at different stages of the COVID-19 pandemic?
2. How do these changes in academic engagement patterns vary across socio-demographic student groups?

As discussed above, the pandemic disturbed students' academic experience in multiple ways such as forcing online learning and increasing stress levels. In the meantime, vaccines became available, and institutions made systematic effort to improve online instruction and support students' wellbeing. While this study is not able to separate these different factors, the longitudinal analyses can unveil how the factors combined might have modified engagement patterns in different ways over time.

This study is expected to make a few contributions to existing research and practice. First, it presents one of the first longitudinal analyses of students' academic engagement and educational experience until more than two years into the pandemic, thereby capturing how students dynamically respond to the changing policy and public health conditions. Second, it leverages campus-wide real-time digital behavior traces data to objectively measure (online) engagement patterns with a much higher level of granularity than in existing research, which can provide more actionable insights to instructors and policymakers. Third, beyond the COVID-19 context, this study illustrates "big data" approaches to understanding micro-level educational experience and macro-level inequality on a large scale. While educational data mining and learning analytics researchers have pioneered such approaches in the past decade (Romero and Ventura, 2020), they have mostly taken a learning science perspective and rarely performed longitudinal, campus-wide analyses which can more directly inform macro-level policymaking.

3.2 Material and Methods

3.2.1 Research Context

This study focuses on student experience at a large public four-year institution in the United States. The institution is officially designated as a Hispanic-Serving Institutions (HSI) and an Asian American and Native American Pacific Islander-Serving Institution (AANAPISI), suggesting that a significant share of the student population comes from marginalized racial groups. Each year the institution enrolls around 7,000 undergraduate freshmen and transfer students.

The institution runs on a quarter system which divides the academic year (excluding summer) into three ten-week quarters: Fall, Winter, and Spring. The COVID-19 pandemic hit the campus during the last week of instruction in Winter 2020 (mid-March), when the administration made an emergency announcement to halt in-person final exams, move Spring 2020 instruction completely online (except for courses that had to take place in person such as hands-on lab sessions in natural sciences), and encourage students to return to their off-campus residence. To help instructors better prepare for online teaching, the institution offered extensive professional development opportunities in Summer 2020. During the 2020-21 Academic year, students returned to on-campus housing with limited capacity, but most undergraduate courses were still fully remote. Fall 2021 marked the full return of on-campus educational experience after a decent proportion of faculty, staff and students were vaccinated. Throughout the Academic Year 2021-22, instruction was mostly in-person in campus classrooms, although temporary transition to remote instruction was in order when COVID-19's Omicron variant first became prevalent in the local community. As of Spring 2022, various instructional policies (e.g., modality, grading) had returned to what they looked like before the pandemic.

Given this context, this study examines students' academic experience throughout four phases:

- Phase 0: Pre-pandemic (by the end of Winter 2020)
- Phase 1: Emergency transition to online learning (Spring 2020 and Summer 2020)
- Phase 2: Routinized online learning (Fall 2020 to Summer 2021)
- Phase 3: In-person learning with the pandemic (Fall 2021 to Summer 2022)

The institution formally deployed Canvas as the standard learning management system (LMS) across the campus in 2016 and provided extensive training and technical support to help instructional staff adopt the system ever since. Before the pandemic started, a large share of instructors had already managed their courses in Canvas to some extent, regardless of the course modality. During the pandemic, Canvas became the major system to manage instructional content and logistics.

3.2.2 Data Sources and Measures

This study primarily takes advantage of campus-wide Canvas log data, which tracks every single action a student has ever taken in the system, as well as the metadata of these actions (time, content, associated Canvas content, etc.), since the system was deployed at the institution. Another important data source is the institution's administrative data, which documents students' background information such as demographics and prior academic achievement. Before research access, the identifiable student information (names, student IDs) in both data sources has been replaced with random IDs following FERPA requirements and this deidentification process has been approved by the IRB. Across data sources, a given student will always be assigned the same random ID, so that their Canvas data can be linked to their background information. The analyses include Canvas data generated by all full-time undergraduate students between Fall 2016 and Spring 2022, save for summer quarters due to their optional nature for students.

The richness of Canvas data makes it possible to characterize various aspects of student engagement. Based on learning analytics research in the past decade and the COVID-19 context, the current analysis constructs six measures of academic engagement shown in Table 3.1. The first three measures capture overall engagement (i.e., all types of engagement) and represent the most commonly used behavioral variables in prior research. By contrast, the latter three measures reflect more nuances of students' study habits. Because students' online behavior is largely shaped by instructional conditions, all the measures are computed at the student-by-course level.

Table 3.1: List of engagement measures extracted from Canvas data

Category	Measure	Definition/Notes	References
Overall engagement	Number of study sessions	Study session is an extended period with frequent actions (no lapse longer than 30 minutes).	(Cicchinelli et al., 2018; Conijn et al., 2017)
	Total time online	Sum of time lapse (in minutes) of all user actions	(Cicchinelli et al., 2018; Conijn et al., 2017)
	Number of actions	Including user action of any kind in the system	(Cicchinelli et al., 2018; Conijn et al., 2017)
Consistency & regularity	Average session duration	Average duration (in minutes) of study sessions	(Cicchinelli et al., 2018; Conijn et al., 2017)
	Regularity of session duration	Standard deviation of study session duration (in minutes)	(Conijn et al., 2017)
	Share of late-night study time	Proportion of time online between midnight and 6am	(Motz et al., 2019)

3.2.3 Statistical Analysis

To understand the changes in engagement patterns of the entire student population over time (RQ1), a fixed-effects regression model is utilized:

$$y_{ict} = \alpha + \sum_{k=1}^{3} \beta_k Phase_{tk} + \mu_i + \lambda_{s(t)} + \epsilon_{ict} \tag{3.1}$$

where y_{ict} is a measure of student i's engagement in course c offered in term t and $Phase_{tk}$ is a vector of binary variables that indicate three phases respectively. Additionally, student fixed effects μ_i and season fixed effects $\lambda_{s(t)}$ are included. With this specification, the estimated coefficients β_k can be interpreted as the expected change in academic engagement in Phase k compared to the same students and the same quarters in previous years.

To capture the more nuanced dynamics of academic engagement through the pandemic, an alternative term-by-term event study model is specified:

$$y_{ict} = \alpha + \sum_{q=-10, q\neq 0}^{7} \gamma_q D_t^q + \mu_i + \epsilon_{ict} \tag{3.2}$$

where D_t^q is a dummy variable which indicates the $|q|$th term after (for $q > 0$) or before (for $q < 0$) the initial hit of the pandemic in Winter 2020 (where $q = 0$). The coefficients γ_q thus estimate the post-pandemic (for $q > 0$) or pre-pandemic (for $q < 0$) trends by term compared to Winter 2020.

To depict how these trends differ across sociodemographic groups (RQ2), four sociodemographic variables from the administrative data are used to define student subpopulations: gender, race/ethnicity, parental education, family income. Each variable defines two comparative groups, and the less disadvantaged group is treated as the reference group in the analyses:

- Male (reference) vs. other genders

- Underrepresented racial minority (URM) vs. non-URM (reference)

- First-generation vs. continuing generation college students (reference)

- Low-income vs. high-income family (reference)

Equations 3.1 and 3.2 are then modified to include the grouping information:

$$y_{ict} = \alpha + \sum_{k=1}^{3} (\beta_k Phase_{tk} + \delta_k Group_i Phase_{tk}) + \kappa Group_i + \lambda_{s(t)} + \epsilon_{ict} \tag{3.3}$$

$$y_{ict} = \alpha + \sum_{q=-10, q\neq 0}^{7} (\gamma_q D_t^q + \theta_q Group_i D_t^q) + \kappa Group_i + \epsilon_{ict} \tag{3.4}$$

where $Group_i$ is an indicator of non-reference group in one of the four group pairs above, and its coefficient κ estimates the group difference in engagement in the baseline phase/term. The interaction terms between $Group_i$ and $Phase_{tk}$ or D_t^q capture the additional change in the group difference in the corresponding phase/term.

The absolute levels and variations of academic engagement highly depend on the nature and design of individual courses, and students from different sociodemographic groups may self-select into different bundles of courses. Therefore, the engagement measures are standardized within courses using z-scores for subgroup analyses. Substituting these standardized measures for their raw values in Equations 3.3 and 3.4, the alternative specifications become:

$$\widetilde{y_{ict}} = \alpha + \sum_{k=1}^{3} \delta_k Group_i Phase_{tk} + \kappa Group_i + \epsilon_{ict} \tag{3.5}$$

$$\widetilde{y_{ict}} = \alpha + \sum_{q=-10, q\neq 0}^{7} \theta_q Group_i D_t^q + \kappa Group_i + \epsilon_{ict} \tag{3.6}$$

where $\widetilde{y_{ict}}$ is a standardized engagement measure. Time-related fixed effects are removed in these specifications because the within-course standardization removes any between-term differences in average engagement measures. In this case, the coefficients κ, δ_k and θ_q have the meanings as before, except that the estimated group differences in engagement are all measured relative to class averages instead of by raw values.

3.3 Results

3.3.1 Data Coverage

While the institution adopts Canvas as part of their technical infrastructure and supports instructors in effectively using the system, the actual usage is at the discretion of individual instructors. Therefore, engagement measures derived from Canvas, while unobtrusive and scalable, only partially capture students' overall engagement, and it is important to be aware of the coverage of this data source in empirical analyses. Toward this end, Table 3.2 compares on a term-by-term basis the number of distinct students, course sections and student-by-section records in Canvas data to those in the administrative data, which cover the full population of students and courses. In line with the timeline of Canvas adoption at the institution, there has been an increasing number of students and courses that used Canvas since 2016. Figure 3.1 presents a visual summary of this trend by plotting the percentages over time. There was a notable leap in Canvas coverage in 2018 and 2019, and before the outbreak of COVID-19, almost all students had left some traces of engagement across over 75% of their enrolled courses. In addition, the hit of COVID-19 moved classes online and further elevated this coverage. Throughout the pandemic, around 90% of student-by-section enrollments left some behavioral trace data, even so during the 2021-22 academic year when most instruction already returned to in-person. Note that in terms of distinct course sections, the coverage seems much lower, but that is mostly because the administrative data includes sections that are not academic courses with mostly small enrollments, such

Table 3.2: Count of distinct students, course sections, and student-by-section records in Canvas data and administrative data, by academic term

	Course section		Student		Student-by-section	
Term	Admin	Canvas	Admin	Canvas	Admin	Canvas
Fall 2016	4,838	922	27,200	19,383	168,041	35,483
Winter 2017	4,911	929	26,498	20,372	164,110	40,996
Spring 2017	4,751	915	25,480	20,005	154,017	41,755
Fall 2017	5,028	1,335	29,258	25,723	179,239	58,511
Winter 2018	5,189	1,385	28,426	25,886	175,709	65,494
Spring 2018	4,921	1,354	27,319	24,634	166,046	63,837
Fall 2018	4,949	1,756	29,793	28,872	184,033	93,105
Winter 2019	5,648	2,469	29,719	29,135	182,734	131,959
Spring 2019	5,297	2,252	28,300	27,858	168,636	125,737
Fall 2019	5,367	2,406	31,312	30,563	187,722	141,106
Winter 2020	5,538	2,641	30,264	29,704	183,702	141,344
Spring 2020	5,164	2,894	28,845	28,484	170,832	148,020
Fall 2020	5,288	3,103	30,383	30,039	182,885	160,720
Winter 2021	5,237	2,940	28,904	28,572	173,178	151,163
Spring 2021	4,967	2,841	27,284	26,992	160,287	146,409
Fall 2021	5,284	2,922	30,041	29,826	177,454	152,709
Winter 2022	5,266	3,004	28,685	28,556	169,389	150,302
Spring 2022	4,932	2,766	27,385	27,060	155,113	139,489

individual studies. These numbers justify Canvas data as a useful source of information for researchers and practitioners to understand students' academic experience across campus and over time.

Because this study is focused on equity consequences of the pandemic, it is also important to know if the analytical sample includes students from marginalized groups. Figure 3.2 depicts the proportion of students in Canvas data who belong to the four sociodemographic groups defined in the previous section. While the data does not cover the full student population on campus before 2019 (as shown in Figure 3.1), the demographic composition is consistent over time, suggesting that the early Canvas data may still well represent campus-wide student experience. Overall, there is a decent share of students from backgrounds that place them

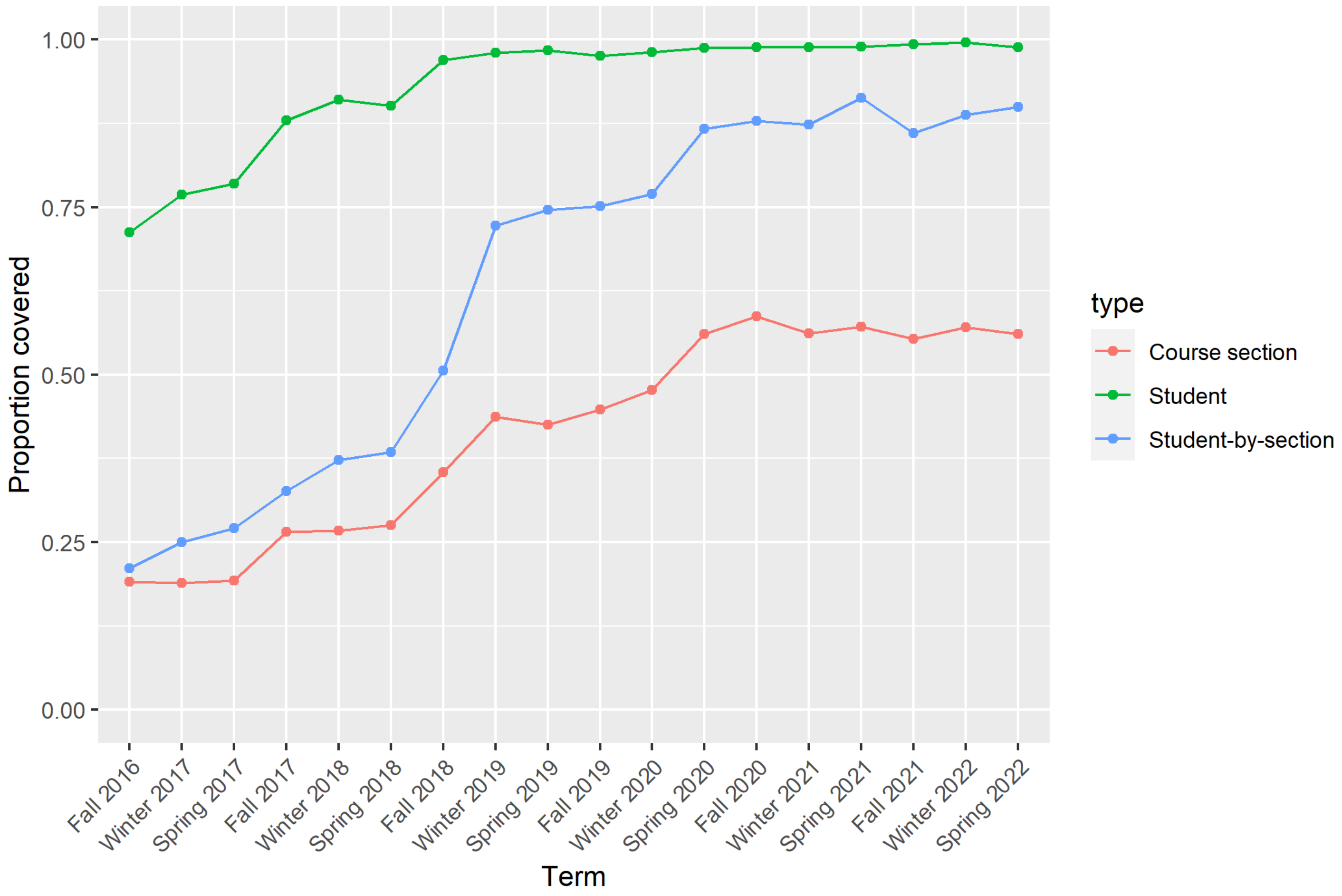


Figure 3.1: Proportion of distinct students, course sections, and student-by-section records covered by Canvas data, by academic term

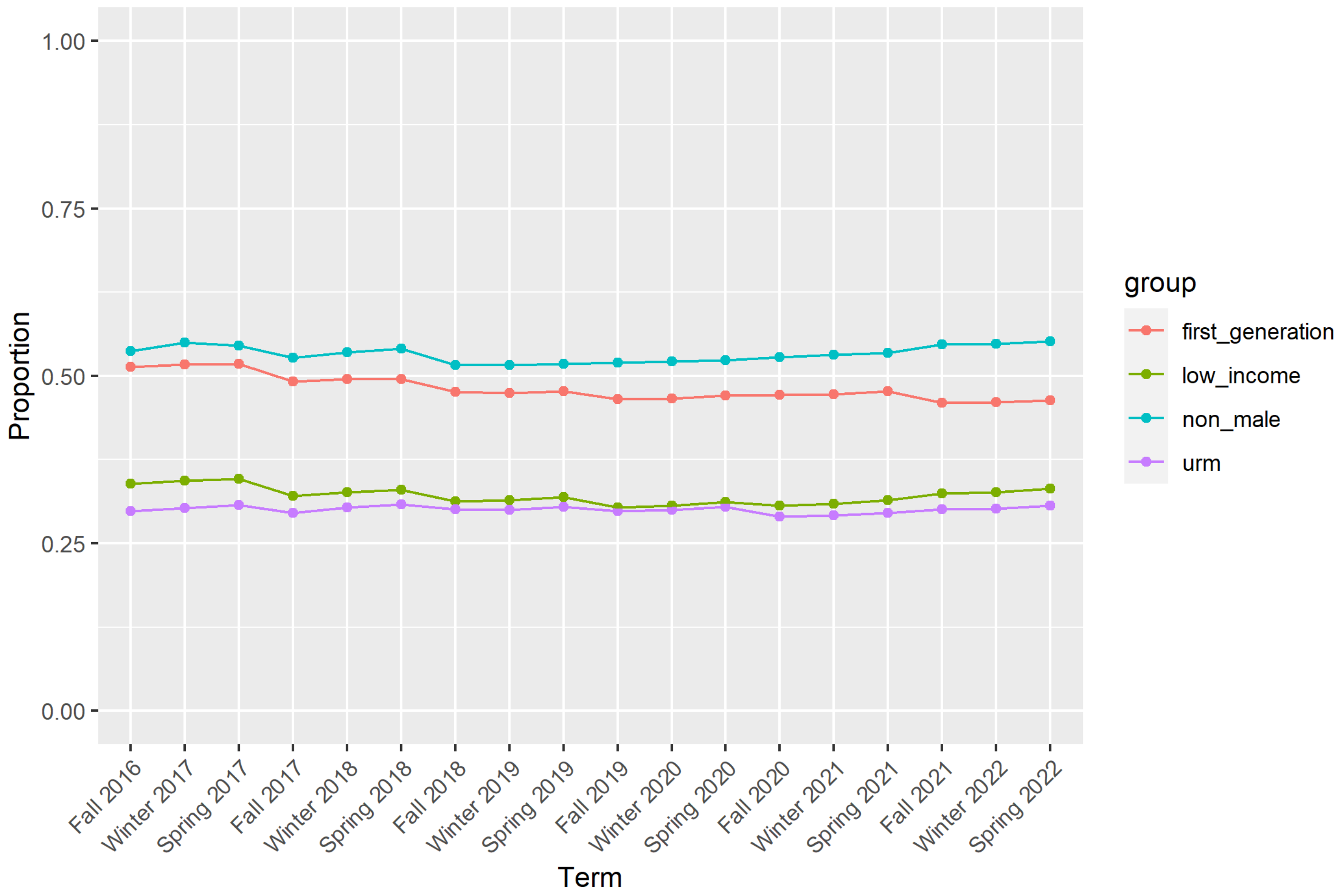


Figure 3.2: Proportion of students from marginalized groups covered by Canvas data, by academic term

at a comparative disadvantage in higher education and expose them to greater challenges during the pandemic.

3.3.2 Population-Level Changes in Engagement Patterns

Figure 3.3 depicts the term-by-term averages of the six engagement measures, with standard deviations in the shadow. The three measures in the upper row capture overall engagement, which remained consistent before the pandemic, substantially rose during the initial hit of the pandemic, and then slightly dropped to a higher consistent level than before. Because the averages are computed at the student-by-course level, these trends are independent of the increasing Canvas data coverage depicted in Figure 3.1. These fluctuations are not surprising given that the pandemic initially forced instructors to teach online and leverage the LMS much more than in-person, and that instructors' increasing knowledge about the system kept their usage at a decent level. The other three measures in the lower row reflect more nuances of students' engagement patterns, i.e., consistency and regularity. Different from overall engagement, these three measures showed slight to moderate increases after the initial COVID outbreak but later dropped by larger amounts than these initial increases. In addition, the standard deviations of all six measures are sizeable compared to their mean values, suggesting the high variability of behavioral patterns not only across students but also across course contexts.

To more formally depict these changes, Table 3.3 presents the results from Equation (1) with a four-phase specification, where each column examines one engagement measure, and the coefficients estimate the average deviation in each phase from the pre-pandemic average. Figure 3.4 illustrates the estimated term-by-term changes in engagement compared to the onset of COVID-19. These two sets of model estimates mostly reaffirm the patterns in Figure 3. Specifically, during the initial pandemic hit (Phase 1), students engaged with Canvas significantly more frequently and stayed much longer each time they engage, leading their total time online and number of activities to almost double than right before the pandemic. They also engaged less regularly than before, reflected by increased variations in session duration. As instructors and students got more used to fully remote learning experience in the following few terms (Phase 2), students' overall engagement gradually went down but still engaged more frequently with more time in total than if not in the pandemic. However, individual online sessions during this phase became increasingly shorter and more regular in length, and by the end of this phase they had been more so than pre-pandemic terms. After instruction returned to campus (Phase 3), overall engagement in the system continued to decrease and in the last term of this phase, students had slightly fewer sessions and spent less time in total than before the pandemic, and individual sessions were also shorter and less variable. Finally, there were fluctuations in the share of late-night study time before and throughout the pandemic, but in the last few terms of the analytical window, this share went down to its lowest point.

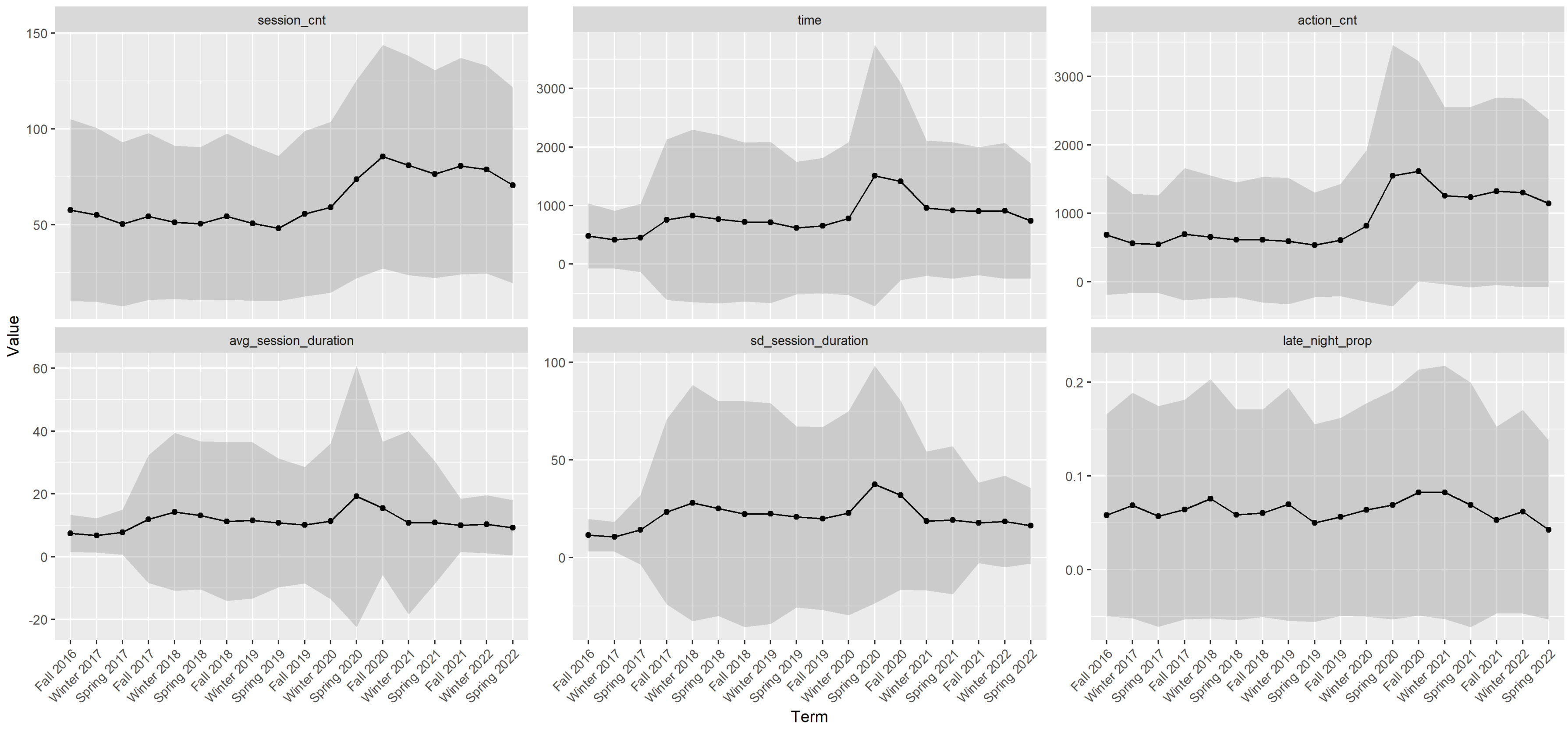
3.3.3 Sociodemographic Differences in Engagement Patterns

All the foregoing trends are further decomposed along sociodemographic lines, or more specifically, the four grouping variables. Table 3.4 presents the estimates from Equation 3.3, where each engagement measure is regressed with the four-phase specification interacted with each of the grouping variables. These results first unveil the baseline group differences in the pre-pandemic terms via the estimated coefficients in the "Group" row. Overall, racial minorities, first-generation college students and disadvantaged gender groups had slightly lower levels of engagement than their peers, whereas low-income students had more mixed patterns compared to their counterparts.

Second, the estimated coefficients of the interaction terms depict how changes in engagement differ across these groups during the pandemic. During the initial emergency transition, these marginalized groups had similar or smaller increases in engagement than their peers, suggesting widened engagement gaps than before. However, during the routinized online learning phase when engagement levels were still higher than before the pandemic, marginalized groups had even larger deviations such that the pre-pandemic engagement gaps between groups were closed and even reversed for the count of study sessions and total time online. This pattern became more prominent in the new in-person learning phase, where the marginalized groups had a significant bump in engagement compared to their peers, which largely reversed the pre-pandemic engagement gaps.

These estimates are further broken down on a term-by-term basis with Equation 3.4 and the results are depicted in Figure 3.5. Each point (with error bars) plots the expected difference between each pair of groups (e.g., URM vs. non-URM) in their change of engagement in the given term. The patterns in the plots reaffirm that marginalized student groups experienced less positive changes in the early stage of the pandemic but more positive changes later on. Notably, the share of late-night study time has a more consistent pattern than other engagement measures. All of the four marginalized groups spent less time studying late at night before the pandemic, and throughout the pandemic they were also less likely to increase late-night study time than their peers.

Finally, Table 3.5 and Figure 3.6 depict the estimates from Equations 3.6 and 3.6, which are standardized versions of the analyses above that control for both contextual differences across courses and different course enrollments between student groups. With this standardization, the estimates should be interpreted as a given group's baseline engagement level or the temporal change in engagement relative to their classmates who come from the corresponding reference group. The results mostly align with the findings from Table 3.4 and Figure 3.5, except that the baseline group differences are less consistent. Specifically, URM and low-income students have higher baseline engagement levels than their peers before the pandemic, whereas gender minorities are less engaged than male students in the same period. These baseline differences slightly alter the interpretation of interaction terms for URM and low-income students, compared to non-standardized results in Table 3.4 and Figure 3.5. The less positive changes among these two groups in the initial stage of the pandemic mean shrunk baseline premiums (instead of widened gaps), and the more positive changes in later stages translate into their increasing lead in engagement (instead of closed gaps).



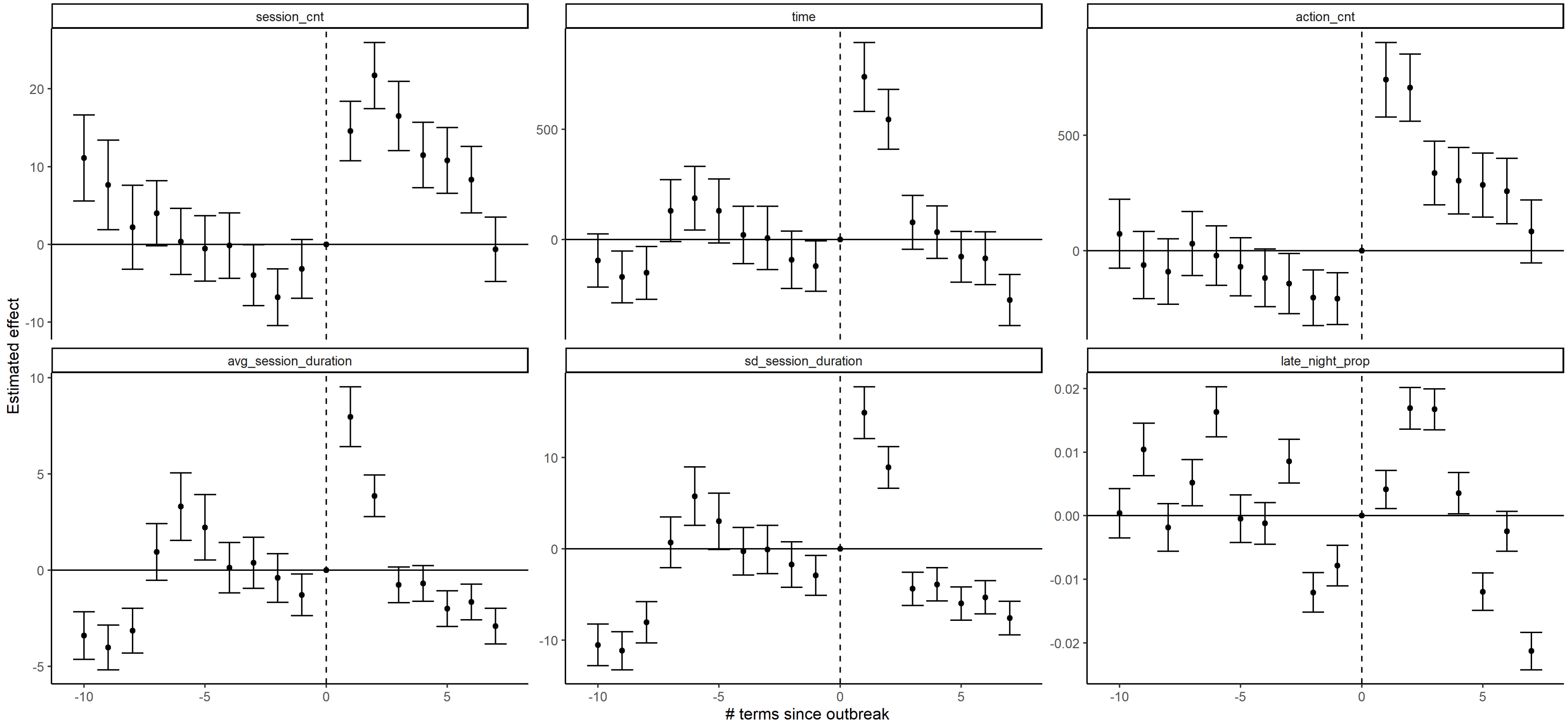
Notes: Standard deviations in shadow

Figure 3.3: Average engagement patterns per student per course, by academic term

Table 3.3: Estimated changes in engagement patterns through three phases of the pandemic

	# sessions	Time	# actions	Avg. session duration	SD session duration	Share of late study time
	(1)	(2)	(3)	(4)	(5)	(6)
phase = 1	21.7***	881.5***	967.2***	8.86***	17.7***	0.016***
	(1.55)	(67.6)	(68.5)	(0.729)	(1.29)	(0.001)
phase = 2	19.7***	280.2***	592.4***	1.22***	1.74***	0.017***
	(1.16)	(33.0)	(35.5)	(0.281)	(0.569)	(0.0009)
phase = 3	9.52***	-83.4***	353.3***	-1.76***	-4.85***	-0.007***
	(1.25)	(32.2)	(37.8)	(0.276)	(0.537)	(0.0009)
student FE	Yes	Yes	Yes	Yes	Yes	Yes
season FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,703,712	1,703,710	1,703,712	1,703,710	1,655,372	1,698,581
R^2	0.279	0.166	0.209	0.102	0.140	0.226

Notes: Course-clustered standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01



Notes: Each subplot illustrates results from a separate event study model with student fixed effects. Each point estimates the expected change in engagement from before the pandemic. Error bars are 95% confidence intervals calculated with course-clustered standard errors.

Figure 3.4: Estimated term-by-term changes in engagement patterns

Table 3.4: Estimated changes in engagement patterns across sociodemographic groups through three phases of the pandemic

Group:	URM	First generation	Low income	Non-male
Outcome: # sessions				
phase = 1	24.9*** (1.76)	25.1*** (1.74)	25.0*** (1.64)	25.1*** (1.83)
phase = 2	27.7*** (1.17)	27.6*** (1.16)	28.1*** (1.10)	28.1*** (1.26)
phase = 3	23.4*** (1.11)	23.3*** (1.07)	24.3*** (1.02)	24.1*** (1.14)
Group	-0.594*** (0.182)	-1.09*** (0.211)	1.34*** (0.348)	-3.38*** (0.274)
Group × phase = 1	0.910* (0.523)	0.944 (0.733)	0.899 (0.990)	1.73* (0.889)
Group × phase = 2	0.763** (0.388)	1.91*** (0.479)	-0.024 (0.764)	0.177 (0.631)
Group × phase = 3	0.771* (0.397)	1.67*** (0.417)	-0.730 (0.736)	-0.522 (0.624)
Outcome: Time				
phase = 1	962.6*** (75.5)	947.3*** (74.2)	986.9*** (87.2)	973.6*** (80.5)
phase = 2	414.6*** (30.4)	413.7*** (29.7)	388.4*** (28.6)	414.5*** (32.3)
phase = 3	157.1*** (24.1)	162.2*** (23.0)	125.6*** (21.5)	147.7*** (24.3)
Group	-27.9*** (5.14)	-15.9*** (5.57)	-11.7 (10.9)	-89.6*** (7.48)
Group × phase = 1	-51.0** (24.7)	-23.3 (30.5)	-83.4 (61.2)	-107.0** (49.9)
Group × phase = 2	31.4*** (10.1)	53.5*** (11.1)	78.3*** (19.9)	56.8*** (14.6)
Group × phase = 3	47.7*** (8.60)	55.4*** (9.43)	99.7*** (16.6)	113.2*** (14.1)

Outcome: # actions					
phase = 1	1,018.8*** (75.5)	1,017.3*** (74.6)	1,029.7*** (82.2)		1,048.1*** (81.6)
phase = 2	721.3*** (34.8)	723.7*** (33.9)	687.4*** (32.4)		743.2*** (37.0)
phase = 3	599.9*** (31.3)	609.3*** (30.1)	562.3*** (29.3)		618.7*** (31.9)
Group	-7.72* (4.37)	-8.18* (4.64)	19.5** (8.72)		-58.0*** (6.32)
Group × phase = 1	7.25 (20.5)	18.4 (26.7)	-8.78 (55.5)		-74.4* (44.6)
Group × phase = 2	58.2*** (11.3)	83.2*** (12.3)	115.2*** (22.1)		26.8 (17.0)
Group × phase = 3	70.0*** (11.1)	71.2*** (13.2)	127.0*** (20.6)		53.4*** (20.7)
Outcome: Avg. session duration					
phase = 1	9.34*** (0.803)	8.94*** (0.772)	10.0*** (0.997)		9.24*** (0.849)
phase = 2	1.43*** (0.259)	1.47*** (0.249)	0.972*** (0.285)		1.30*** (0.262)
phase = 3	-1.35*** (0.209)	-1.18*** (0.204)	-1.97*** (0.220)		-1.49*** (0.211)
Group	-0.370*** (0.072)	-0.088 (0.073)	-0.922*** (0.127)		-0.815*** (0.099)
Group × phase = 1	-0.952*** (0.325)	-0.113 (0.406)	-2.04*** (0.717)		-1.27** (0.552)
Group × phase = 2	0.237* (0.123)	0.222* (0.116)	1.07*** (0.201)		0.833*** (0.152)
Group × phase = 3	0.653*** (0.086)	0.451*** (0.091)	1.73*** (0.150)		1.54*** (0.130)
Outcome: SD session duration					
phase = 1	18.7*** (1.46)	18.3*** (1.42)	20.7*** (1.79)		18.8*** (1.53)
phase = 2	1.92*** (0.537)	2.13*** (0.515)	0.921 (0.623)		1.72*** (0.548)

phase = 3	-4.59*** (0.413)	-4.08*** (0.397)	-6.34*** (0.459)	-4.85*** (0.416)
Group	-1.71*** (0.155)	-1.10*** (0.159)	-4.09*** (0.292)	-3.22*** (0.213)
Group × phase = 1	-1.77*** (0.571)	-1.04* (0.615)	-5.05*** (1.26)	-2.88*** (0.980)
Group × phase = 2	1.01*** (0.243)	0.885*** (0.244)	2.80*** (0.441)	2.39*** (0.322)
Group × phase = 3	2.16*** (0.183)	1.65*** (0.190)	5.18*** (0.328)	4.40*** (0.261)
Outcome: Share of late study time				
phase = 1	0.020*** (0.001)	0.020*** (0.001)	0.016*** (0.001)	0.020*** (0.001)
phase = 2	0.024*** (0.001)	0.026*** (0.001)	0.018*** (0.001)	0.025*** (0.001)
phase = 3	-0.008*** (0.0007)	-0.006*** (0.0007)	-0.010*** (0.0007)	-0.007*** (0.0007)
Group	-0.0002 (0.0003)	0.003*** (0.0003)	-0.008*** (0.0004)	-0.012*** (0.0004)
Group × phase = 1	-0.006*** (0.0008)	-0.011*** (0.001)	0.0008 (0.001)	-0.014*** (0.001)
Group × phase = 2	-0.014*** (0.0007)	-0.029*** (0.001)	-0.0009 (0.0007)	-0.030*** (0.001)
Group × phase = 3	-9.23×10^{-5} (0.0005)	-0.004*** (0.0006)	0.004*** (0.0006)	-0.002*** (0.0006)

Each combination of grouping variable and engagement measure comes from a separate regression model with season fixed effects.

“Group” is a binary indicator of the student group in the corresponding column header.

Course-clustered standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01

Table 3.5: Estimated changes in engagement patterns (standardized) across sociodemographic groups through three phases of the pandemic

Group:	URM	First generation	Low income	Non-male
Outcome: # sessions				
Group	0.019*** (0.002)	-0.005** (0.002)	0.068*** (0.002)	-0.042*** (0.002)
Group × phase = 1	0.007 (0.004)	0.004 (0.005)	0.013*** (0.004)	0.024*** (0.006)
Group × phase = 2	0.005* (0.003)	0.020*** (0.004)	-0.014*** (0.003)	0.016*** (0.004)
Group × phase = 3	0.005* (0.003)	0.021*** (0.003)	-0.013*** (0.003)	0.011*** (0.004)
Outcome: Time				
Group	0.019*** (0.002)	0.014*** (0.002)	0.053*** (0.002)	-0.017*** (0.002)
Group × phase = 1	-0.005 (0.004)	-0.006 (0.005)	0.003 (0.004)	0.0008 (0.006)
Group × phase = 2	0.001 (0.003)	0.009** (0.004)	0.012*** (0.003)	0.035*** (0.004)
Group × phase = 3	0.009*** (0.003)	0.014*** (0.003)	0.018*** (0.003)	0.043*** (0.004)
Outcome: # actions				
Group	0.023*** (0.002)	-0.005** (0.002)	0.044*** (0.002)	-0.039*** (0.002)
Group × phase = 1	0.003 (0.004)	0.007 (0.005)	0.002 (0.004)	-0.005 (0.006)
Group × phase = 2	0.013*** (0.003)	0.028*** (0.004)	0.013*** (0.003)	0.034*** (0.004)
Group × phase = 3	0.017*** (0.003)	0.029*** (0.003)	0.022*** (0.003)	0.032*** (0.004)

Outcome: Avg. session duration					
Group	0.011*** (0.002)	0.019*** (0.002)	0.012*** (0.002)		0.014*** (0.002)
Group × phase = 1	-0.015*** (0.004)	-0.011** (0.005)	-0.010** (0.004)		-0.016*** (0.006)
Group × phase = 2	-0.005* (0.003)	-0.006* (0.004)	0.024*** (0.003)		0.031*** (0.004)
Group × phase = 3	0.008*** (0.003)	0.002 (0.003)	0.035*** (0.003)		0.050*** (0.004)
Outcome: SD session duration					
Group	0.001 (0.002)	0.007*** (0.002)	-0.026*** (0.002)		-0.013*** (0.002)
Group × phase = 1	-0.010** (0.004)	-0.007 (0.006)	-0.020*** (0.004)		-0.014** (0.006)
Group × phase = 2	0.001 (0.003)	0.005 (0.004)	0.020*** (0.003)		0.043*** (0.004)
Group × phase = 3	0.012*** (0.003)	0.013*** (0.004)	0.036*** (0.003)		0.065*** (0.004)
Outcome: Share of late study time					
Group	-0.003 (0.002)	0.024*** (0.002)	-0.051*** (0.002)		-0.105*** (0.002)
Group × phase = 1	-0.027*** (0.004)	-0.052*** (0.005)	0.012*** (0.004)		-0.068*** (0.006)
Group × phase = 2	-0.047*** (0.003)	-0.117*** (0.004)	0.018*** (0.003)		-0.119*** (0.004)
Group × phase = 3	-0.003 (0.003)	-0.019*** (0.003)	0.015*** (0.003)		-0.019*** (0.004)

Each combination of grouping variable and engagement measure comes from a separate regression model.

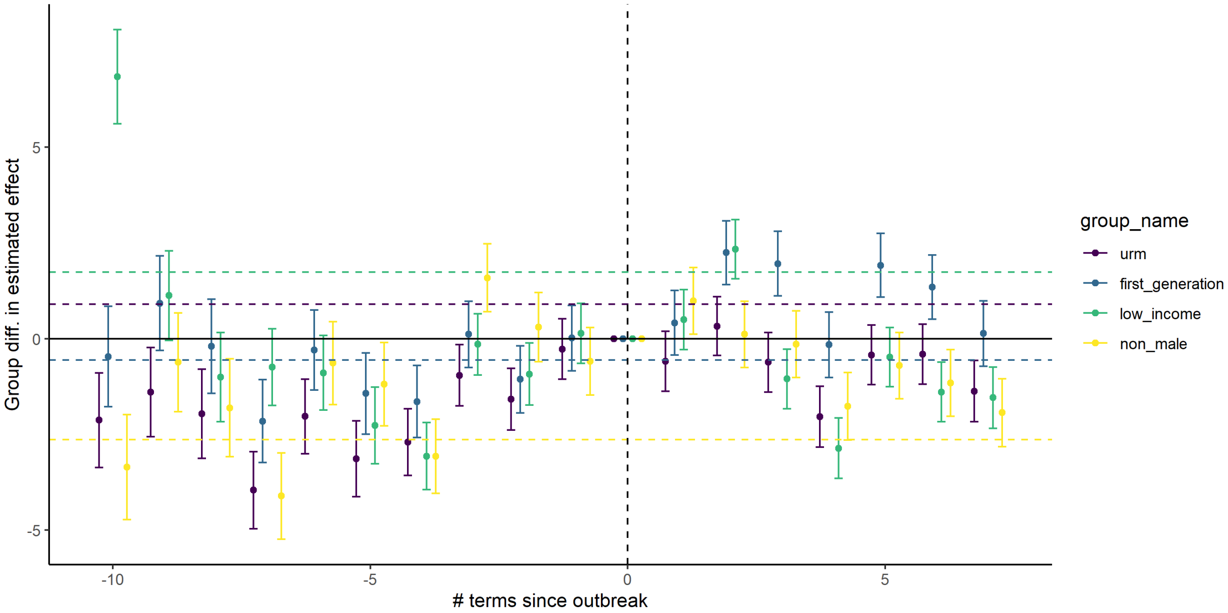
“Group” is a binary indicator of the student group in the corresponding column header.

All outcome variables are z-standardized within classes. Standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01

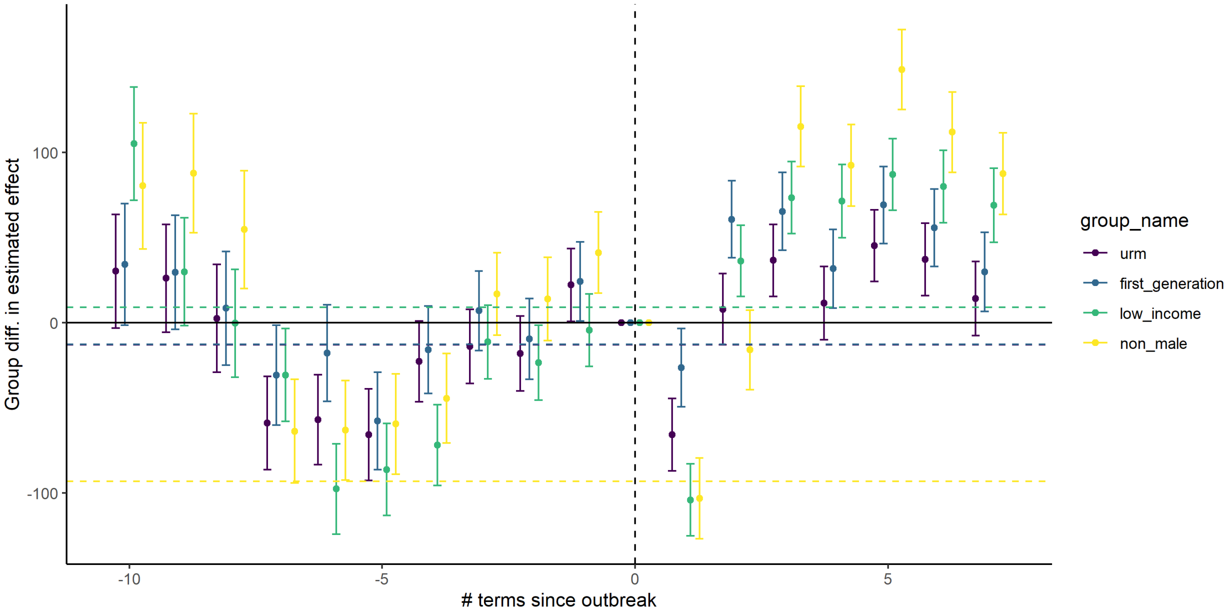
3.4 Discussion and Conclusion

This study provides one of the first and largest empirical analyses of how college students' academic engagement changes, and how students from different sociodemographic groups change differently, through two years of the COVID-19 pandemic. Based on campus-wide behavioral trace data from Canvas LMS at a minority serving institution, six behavioral measures of overall engagement and consistency and regularity of engagement are constructed and computed for around 30,000 students, 2,000 course sections, and 150,000 course enrollments per academic term over six years. A mixture of fixed effects regression models and event study models are used to unveil a handful of prominent patterns. First, student (online) engagement exhibited substantial fluctuations during the pandemic. The overall amount of engagement almost doubled during the initial outbreak, but in later terms students showed gradually decreasing engagement levels and eventually became slightly more dormant than before the pandemic. Second, these trends of engagement patterns are not evenly distributed across different sociodemographic groups. Marginalized groups, including racial and gender minorities, first-generation college students, and students from low-income families, experienced less stark fluctuations in engagement patterns over time compared to their peers. Specifically, the overall engagement of marginalized student groups did not increase as much as their counterparts during the initial outbreak, but these students also did not undergo as sharp decreases in engagement in later stages, and their final engagement levels were still higher than before the pandemic.

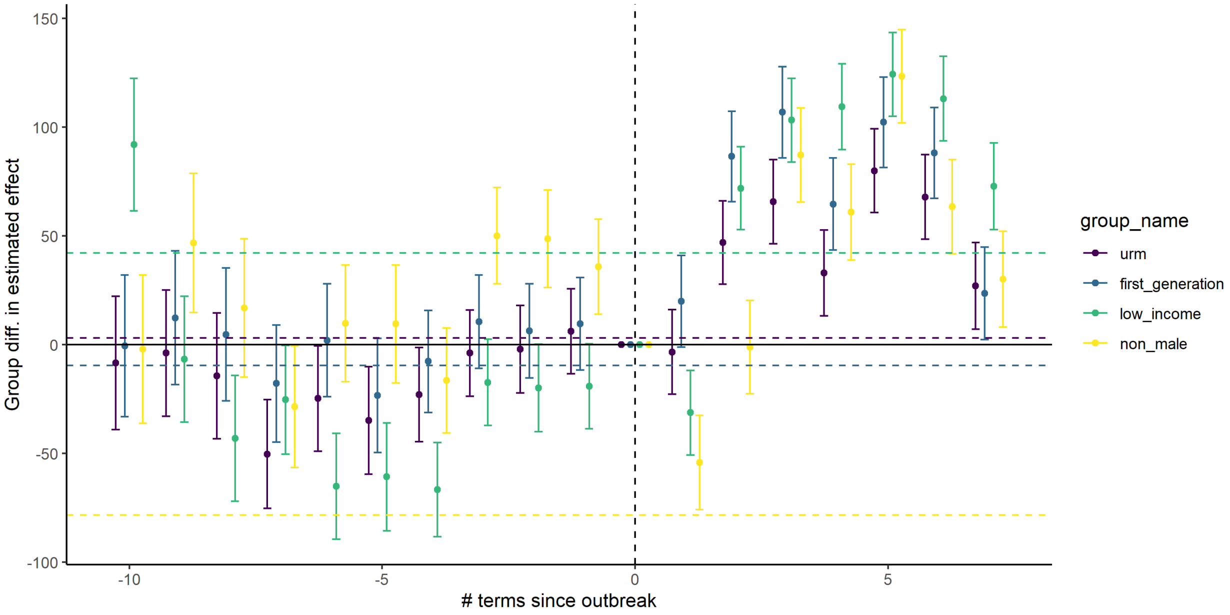
The population-level engagement trend aligns with the changing history of institutional policies throughout the pandemic, but it is somewhat surprising that engagement ended at lower-than-before levels within the analytical window, because even though instruction already returned to in-person in the final academic year, instructors should have been much more used to managing their courses with LMS than before. On the other hand, the sociodemographic breakdown of the time trend unveils that the sharp initial increase and later



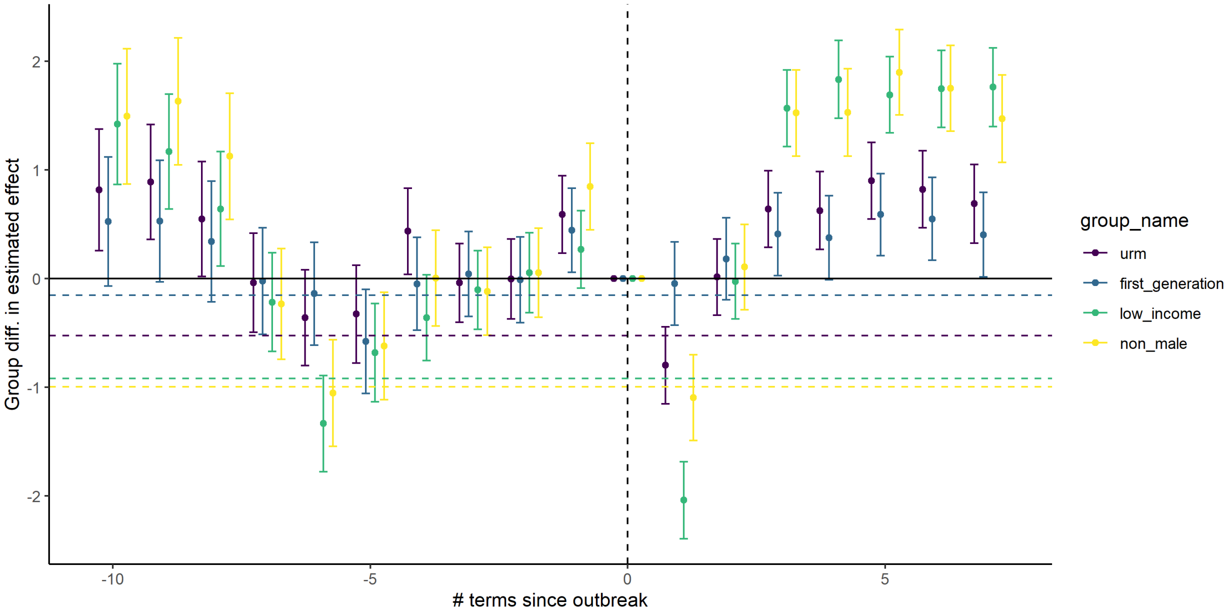
(a) Outcome: # sessions



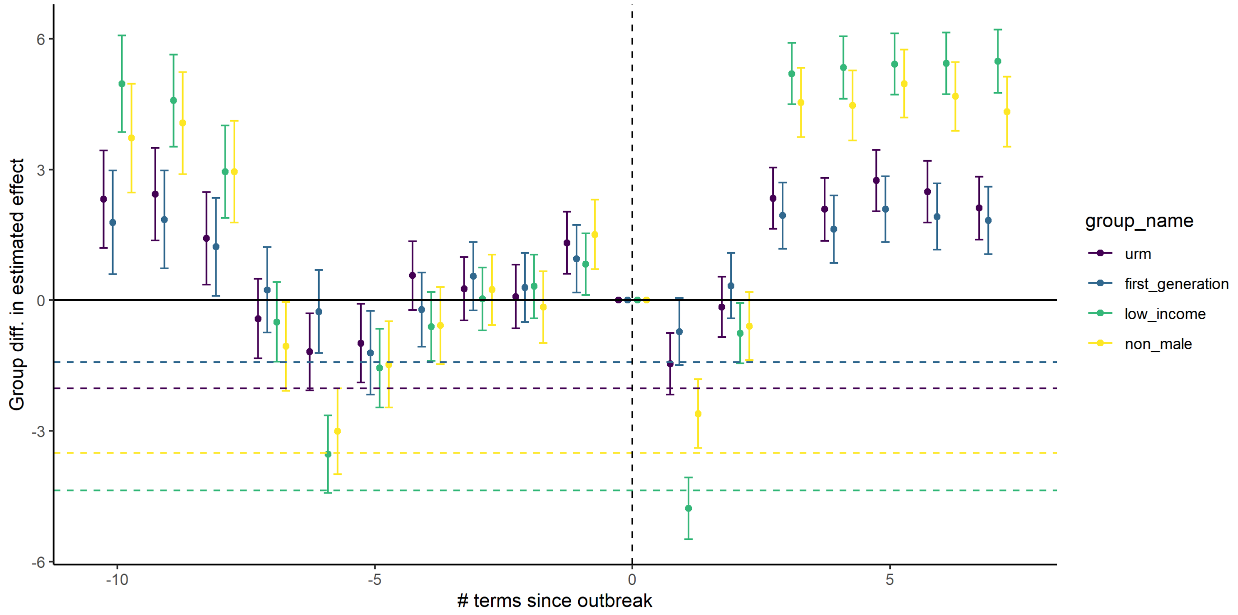
(b) Outcome: Time



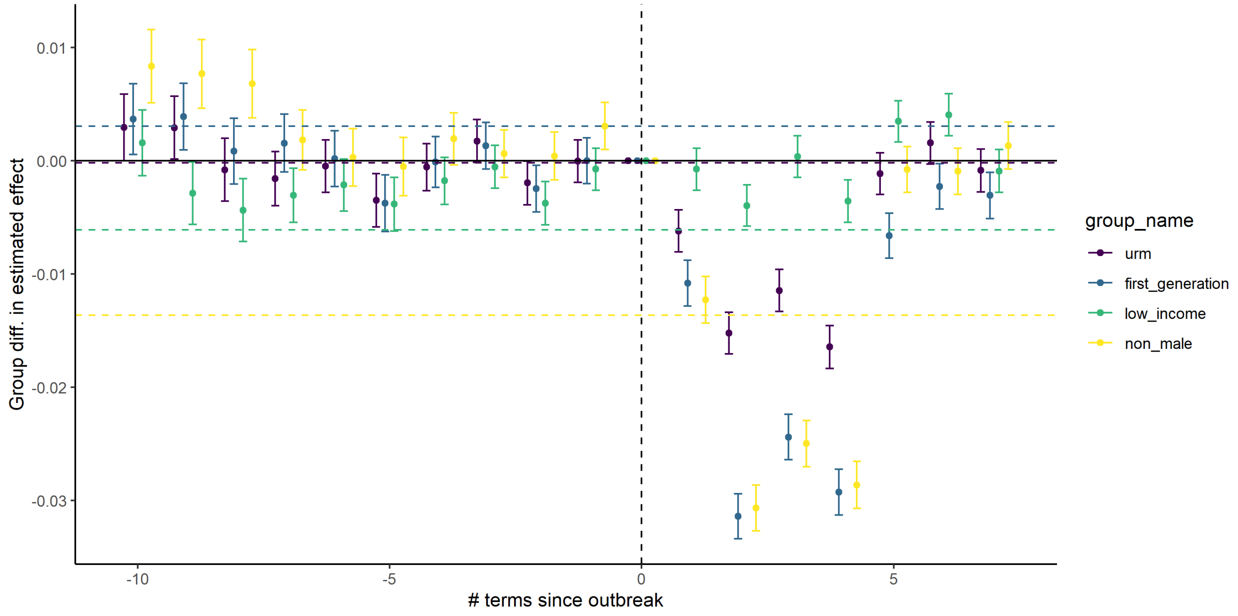
(c) Outcome: # actions



(d) Outcome: Avg. session duration



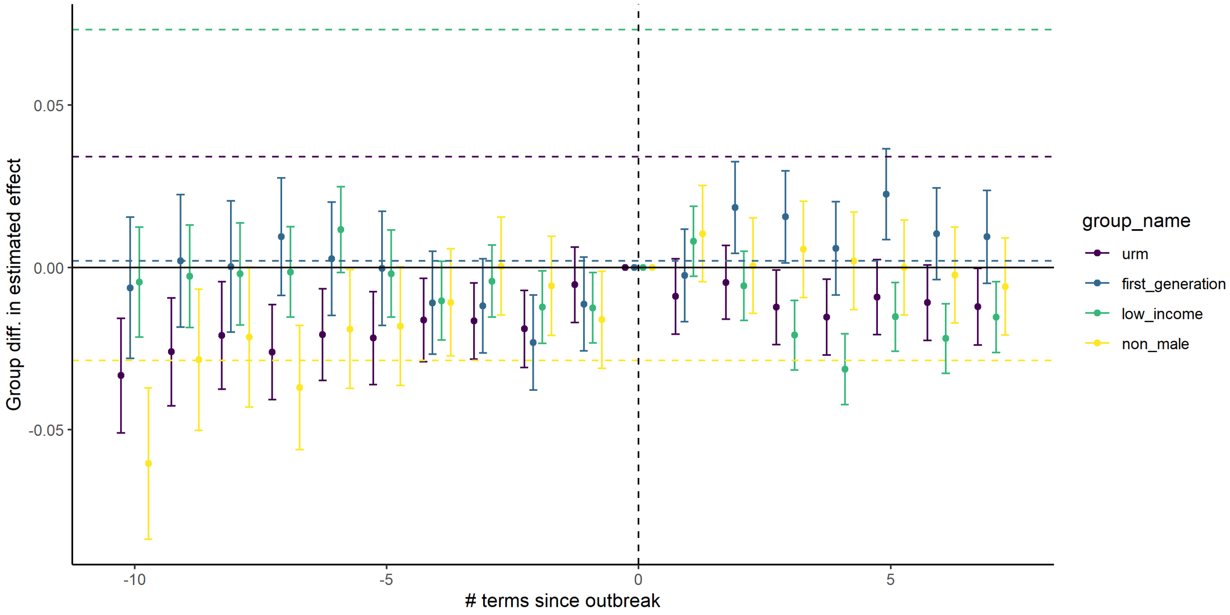
(e) Outcome: SD session duration



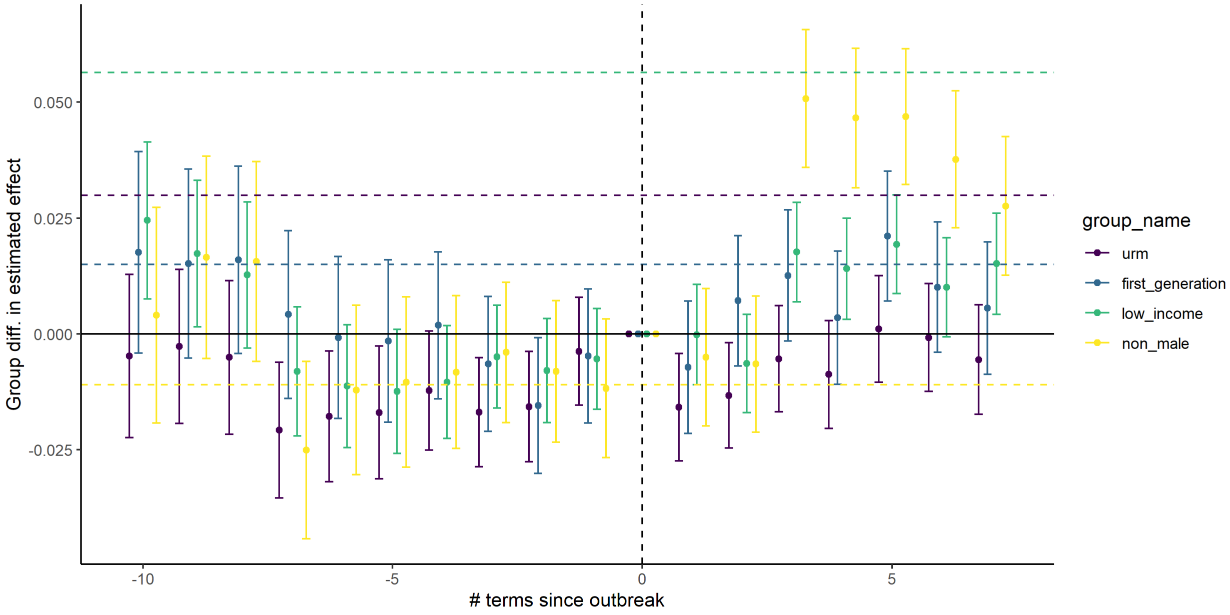
(f) Outcome: Share of late study time

Notes: Each color in each subplot presents results from a separate event study model. Each point estimates the group difference (listed - reference) in the expected change in engagement from before the pandemic. Dashed horizontal lines are the point estimates of baseline group differences. Error bars are 95% confidence intervals calculated with course-clustered standard errors.

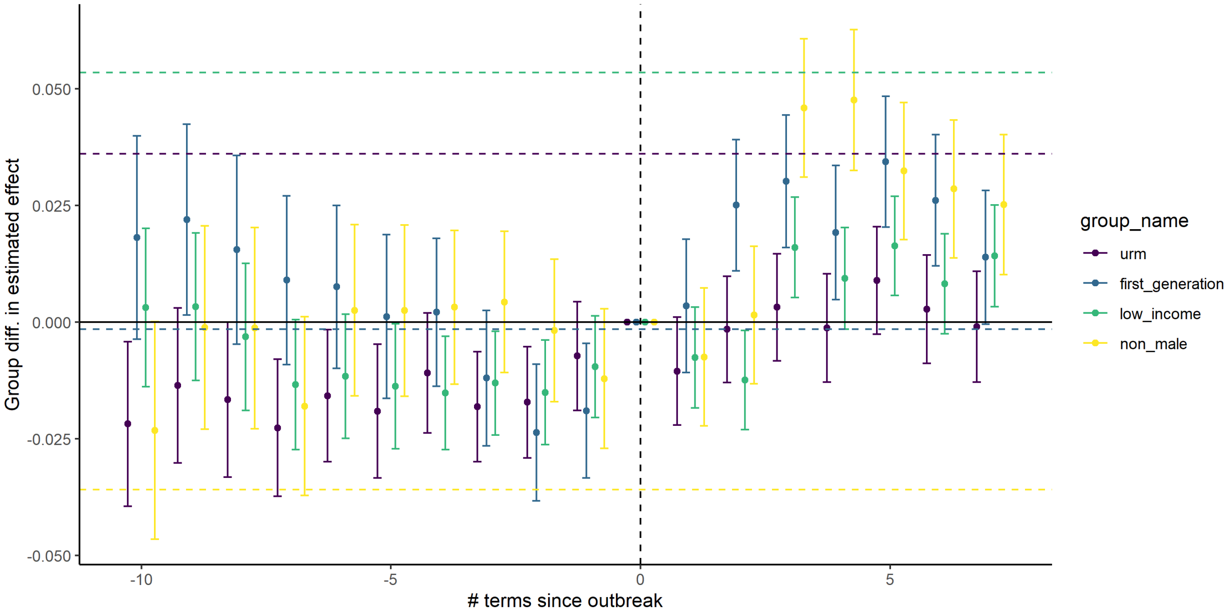
Figure 3.5: Estimated sociodemographic difference in term-by-term changes in engagement patterns



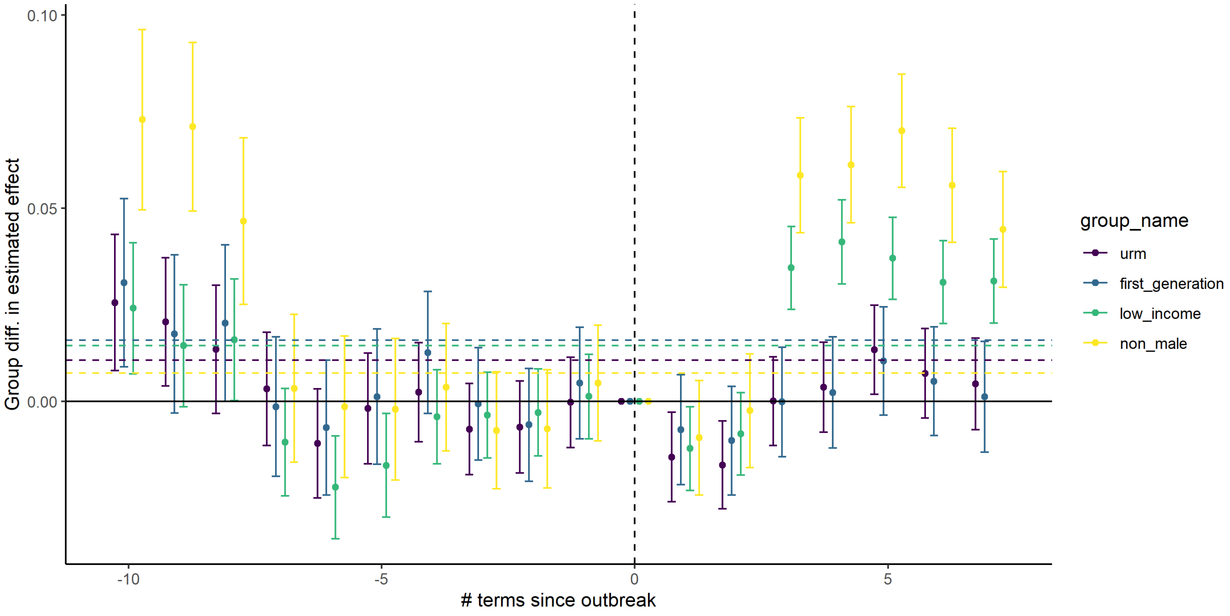
(a) Outcome: # sessions



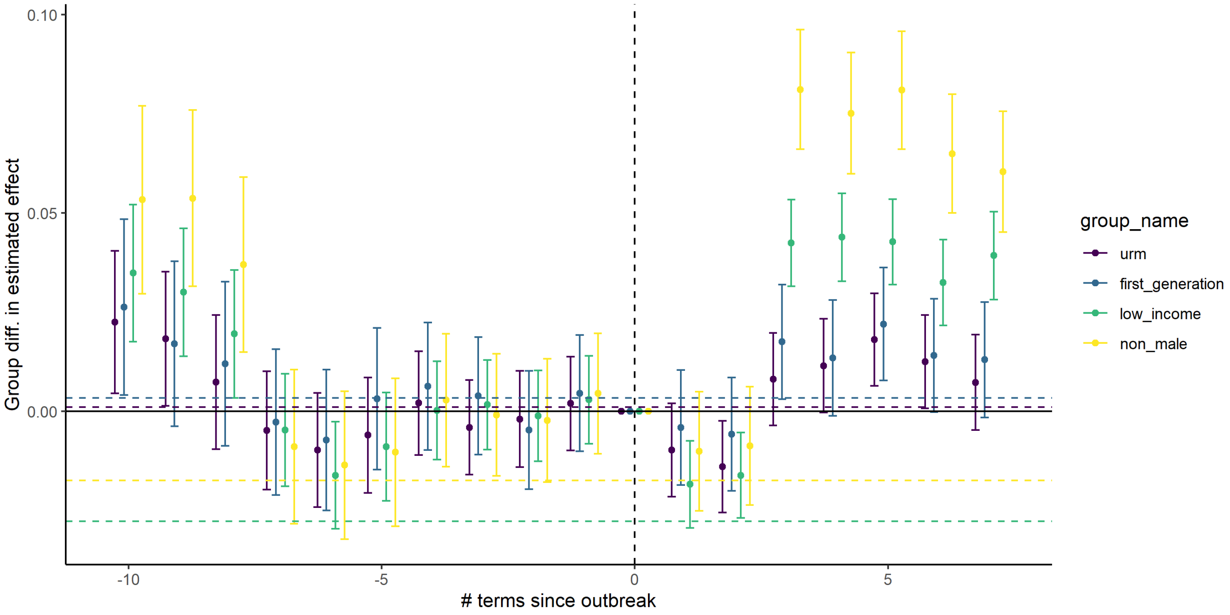
(b) Outcome: Time



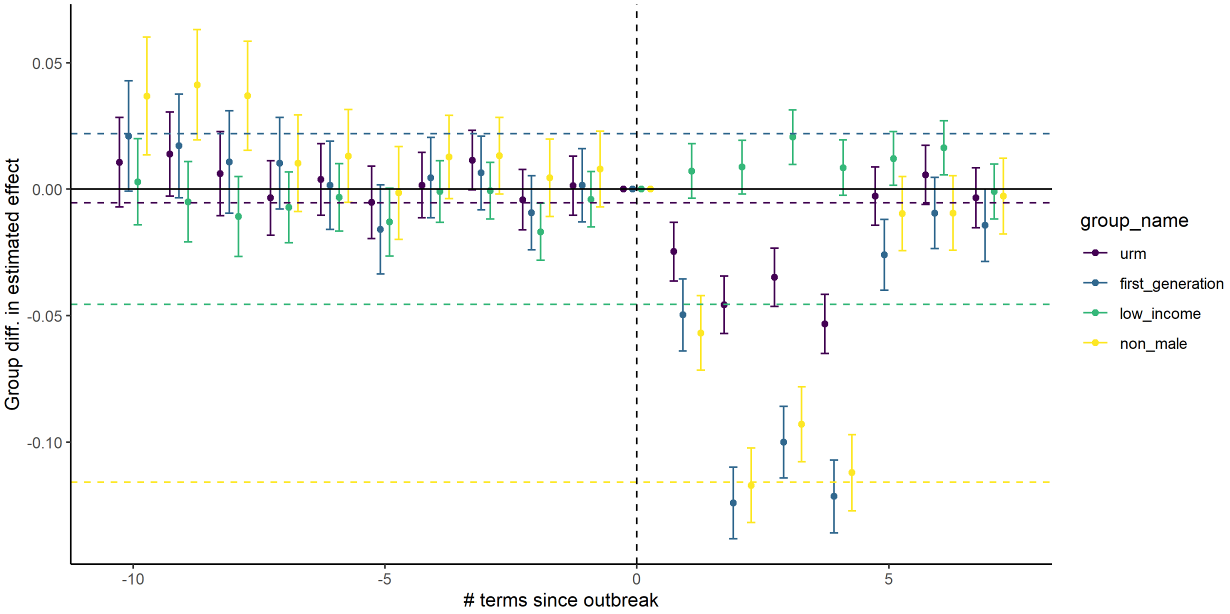
(c) Outcome: # actions



(d) Outcome: Avg. session duration



(e) Outcome: SD session duration



(f) Outcome: Share of late study time

Notes: Each color in each subplot presents results from a separate event study model. Each point estimates the group difference (listed - reference) in the expected change in engagement from before the pandemic. Dashed horizontal lines are the point estimates of baseline group differences. All outcome variables are z-standardized within classes. Error bars are 95% confidence intervals.

Figure 3.6: Estimated sociodemographic difference in term-by-term changes in engagement patterns (standardized)

decline in engagement were both more concentrated on advantaged student groups. A few studies that examine online learning engagement in non-institutional contexts have mixed findings about demographic differences in engagement during the initial disruption stage of the pandemic (Kizilcec et al., 2021; Bacher-Hicks et al., 2021). Findings from this study, however, suggest that the pandemic might have initially undermined educational equity but made contribution to equitable academic experience in the longer term by way of more flexible instructional modalities and more accommodations for diverse learning needs.

This study highlights the affordances of large-scale behavioral trace data for both researchers and practitioners to better understand the dynamics of student experience and educational inequalities at a granular level, both in the pandemic context and beyond. Importantly, because these data are automatically generated by students in their everyday life, no additional data collection is required which might burden students especially in the already disruptive pandemic. Also, most institutions have this type of data with similar or even identical formats, so the current analyses can be easily applied to different institutional contexts with minimal cost, which is especially desirable for low-resourced institutions in stronger need of data-driven insights and support.

This study is only a first step to examine fine-grained academic engagement during the pandemic, and a few important limitations must be recognized. First, the changes in behavior-based engagement measures result from a mixture of institutional, instructional and individual students' responses to the pandemic. Therefore, the estimates from population-level analyses should not be interpreted as pandemic-incurred disturbance on students alone, while the subgroup analyses are less subject to this complication because each student group was compared to their reference group under the same instructional conditions. Second, the behavioral trace data does not capture the physical and psychological contexts of students' actions, so the observed patterns should not be overinterpreted beyond the behavioral level. For example, a large amount of time online might be a signal of strong motivation and hard work but can also indicate academic struggles. Given these inherent challenges, the current findings do not lead to a firm conclusion about the equity implications of the pandemic over the two years. Toward more conclusive insights, future work will include explicitly modeling instructional conditions, constructing more nuanced behavioral measures, and triangulating behavioral measures with survey data collected from the same population.

Chapter 4

Towards Accurate and Fair Prediction of College Success: Evaluating Different Sources of Student Data

4.1 Introduction

The most common application of learning analytics in higher education is using predictive modeling to understand critical factors contributing to student success, or to identify students who need support in a timely manner. Predictive analytics have been used within a course (Arnold and Pistilli, 2012) or while using tutoring software (Xie et al., 2017). They have also been used to optimize student success in the longer term, for example to predict graduation rates (Aulck et al., 2019) or to make course recommendations (Pardos et al., 2019). Different data sources can be used to build these predictive models, with varying trade-offs. For example, when making predictions at the course level, log data from learning management systems (LMS) are often used. These systems allow for automated and scalable recording of hundreds of learner actions in every single minute, but they require robust and efficient data management systems. When making longer-term predictions, on the other hand, institutions can use data typically stored in student information systems (SIS), including prior academic history, standardized test scores, and demographic information. While this data source might be readily available to college administrators, it might be more difficult to access, due to ethical concerns or logistic barriers, for individual instructors or researchers trying to build such models for particular use cases. In some cases, both data sources are further combined with assessments or surveys that measure students' metacognitive abilities or other non-cognitive attributes that might predict college success (Whitmer et al., 2019). However, collecting and managing these data is often costly for institutions if they are not already doing so. Given all these trade-offs, it is necessary to examine the utility of different student data sources for building predictive analytics-based solutions to guide instructors, administrators and education policy makers on the costs and benefits of utilizing different data sources.

To date, research that systematically compares data sources and predictions is underrepresented in the literature (Fischer et al., 2020). To respond to this call for research, this study evaluates the usefulness of three common student data sources for two representative prediction tasks. These three data sources, including institutional data, LMS data, and survey data, are all widely used across research settings and have been shown to predict various measures of college success. Given the different use cases of short-term and long-term predictions as discussed above, we construct two success measures: individual course grades (short-term success) and yearly average GPA (long-term success). The usefulness of each data source is determined by its contribution to overall prediction accuracy and to prediction fairness across student subpopulations. The focus on fairness arises from the concern that predictive models trained on the entire student population may perform systematically worse on selected subpopulations than other others, which may have unintended negative effects for vulnerable students (Barocas et al., 2019). For instance, if models are less confident in identifying struggling students among an already underrepresented group, this bias may eventually amplify existing achievement gaps.

In short, our research aims to identify what combinations of student data (a) more accurately predict different success measures; and (b) more fairly predict these measures. The remainder of this paper is organized as follows: Section 5.2 summarizes the related work on college success prediction and fairness of predictive models; Section 4.3 describes the data and methods we use to construct and evaluate prediction models; Section 4.4 presents the results from various predictions; Section 4.5 reflects on the findings and discusses the practical implications for stakeholders; Section 4.6 concludes the study with limitations and future work.

4.2 Related Work

4.2.1 Predicting College Success Using Student Characteristics

Although college is a complicated ecosystem with numerous factors shaping student outcomes, prior research has identified several groups of student characteristics across institutional data, LMS data, and survey data that consistently predict commonly used measures of success.

Personal Background - Institutional Data

Student success in higher education is often stratified by students' demographic, socioeconomic and academic background prior to college experience. For example, college graduation rates substantially differ by students' race/ethnicity. National data indicates that Hispanic students are 15% less likely to graduate college within six years than their white counterparts, and this gap is 25% between black and white students (Shapiro et al., 2019). Such inequalities are particularly pronounced in STEM fields, where even more underrepresented students drop out of their college careers (Allen-Ramdial and Campbell, 2014). Also, student performance prior to entering college (e.g., on standardized tests) has often been found to strongly predict college performance across different subpopulations (Bettinger et al., 2013). These overall trends suggest that what happens before college remains predictive of student success in higher education settings. Of course, this could be due to a variety of factors, such student background being correlated with patterns of historical and institutionalized oppression as well as other barriers that students from different backgrounds might face both before and during college.

Learning Behavior - LMS Data

In contrast to latent psychological states, learning behavior is a more extrinsic and observable predictor of academic success (Beattie et al., 2019). Behavioral patterns capture variations in college experience that may be orthogonal to students' incoming characteristics, allowing for insights into the mechanism of academic success at a day-to-day granularity. With the prevalence of digital learning platforms, learning behavior can be authentically recorded in the form of clickstream data. These time-stamped data record learner's interactions with LMSs. This allows researchers to create measures that look into the "black box" of study behaviors (Baker et al., 2020). For example, how students allocate their study time is a consistent predictor of performance. Those who have more regular engagement patterns and who space out their study effort (instead of cramming) are more likely to be high-achieving (Park et al., 2018). Similarly, students who strategically regulate their learning effort (e.g., starting from exercise-oriented tactics and moving to other tactics based on encountered challenges) perform equally well but with less effort, compared to simply hard-working students (Matcha et al., 2019).

Non-Cognitive Abilities - Survey Data

There is emerging evidence that non-cognitive factors, such as personality traits, task values and self-efficacy, are associated with positive academic outcomes even after controlling for cognitive factors measured by intelligence tests as well as various background characteristics (Beattie et al., 2018). Among these factors, researchers seem to have reached consensus that self-regulated learning skills are essential because unlike in K-12 schooling, college students have the flexibility as well as responsibility to actively and constantly monitor, reflect on, and adjust their motivation, cognition, and study behavior (Wolters, 1998). To better describe and measure a student's ability to regulate their learning process, (Pintrich and De Groot, 1990) divided it into three subcomponents with two cognitive components (the use of cognitive strategies and the use of metacognitive strategies) and one non-cognitive component (resource management, including skills of time and study environment management, effort regulation, peer learning, and help seeking). A systematic literature review focused on online learning contexts found consistent evidence that resource management skills, especially time management skills and effort regulation skills, are predictive of performance (Broadbent and Poon, 2015). While new technologies are creating novel measurement tools for these intangible qualities, the "ground truth" mostly comes from validated surveys.

4.2.2 Comparison of Different Data Sources

Previous work has examined combining various data sources for predictive analytics in higher education. For example, Arnold and Pistilli (2012) combined institutional data, course performance data and LMS data to predict students' within-course success. However, there has been little work comparing the impact of various data sources on student success. Aulck et al. (2019) compared the impact of different types of institutional variables, including demographic variables, prior academic achievement, student majors, and academic achievement in college courses on predicting graduation and re-enrollment rates. Wolff et al. (2013) compared the impact of virtual learning environment (VLE) data, course assessment data, and a demographic variable on predicting whether a student's performance will drop in a course and whether a student will pass or fail a course. They generally found that using VLE data in conjunction with assessment data was seemingly better than using either alone. In what is perhaps the closest study to ours, Whitmer et al. (2019) compared the impact of learning behavioral features, student background, and non-cognitive features measured by a socio-emotional skill assessment on predicting within-course success. Our study differs from theirs in that we look at long-term outcomes as well as short-term outcomes, we analyze the fairness of predictive models, and we fit models that span across several courses.

4.2.3 Fairness of Predictive Analytics in Education

In recent years, the fairness and biases of machine learning algorithms and systems have developed into a focused research area in the general machine learning research community[1]. Research efforts encompass developing statistical measures of fairness, evaluating existing algorithms/systems, and correcting for biases in algorithmic pipelines, among others. As fairness is a concept rooted in a variety of disciplines, it has been a consensus that there is no single "correct" definition of fairness. Rather, what is fair is highly dependent on the specific application scenarios (**?**). As such, contextualizing the fairness research in different fields is critical to improving real-world applications.

In earlier education research, there has been a focus on heterogeneous effects across student subpopulations in the contexts of testing (Thorndike, 1971), observational studies (Xu and Jaggars, 2014) and program evaluation (Schippers et al., 2015). These earlier perspectives resonate with the current theme of fairness, but as the adoption of predictive analytics systems in education for high-stakes purposes has a comparatively shorter history, formalized

[1] https://facctconference.org/

research on fairness in such contexts has been somewhat limited. Among the handful of empirical papers that have directly evaluated this aspect of predictive analytics in education, Doroudi and Brunskill (2019) showed through a simulation study that misspecified student models in intelligent tutoring systems could leave “slow” learners at lower mastery levels than “faster” learners; Gardner et al. (2019) examined the ROC curves from MOOC dropout prediction models, and identified significant gaps between gender groups through slicing analysis; and Hutt et al. (2019) used college application materials to predict on-time graduation and, employing the same slicing analysis, concluded that their model could make fair predictions across five sociodemographic groups.

As Barocas et al. (2019) points out, while the biases of predictive systems may be attributed to unfair algorithms, they can also arise from biased data which “reflect historical prejudices against certain social groups, prevailing cultural stereotypes, and existing demographic inequalities”. Therefore, unlike the previous studies described above, this paper examines fairness as an attribute of *data sources* rather than of *algorithms*. We look at fairness with respect to between-groups differences in three metrics: accuracy, false positive rate, and false negative rate. These metrics are among the many fairness metrics that have been proposed in the literature (Barocas et al., 2019). For example, having an equal false negative rate between subgroups has been called “equality of opportunity” in the context of giving everyone an equal opportunity to receive a positive intervention (e.g., being part of the university’s honor roll for having a high GPA) (Hardt et al., 2016).

4.3 Data and Methods

4.3.1 Data Sources

Following Section 4.2.1, this study compares the three widely available data sources in higher education settings: institutional data, Canvas LMS log data, and survey data. Specifically, we drew the sample of all students who enrolled and received final grades in ten fully online, introductory STEM courses taught from 2016 to 2018 at a large, public research university in the United States. Six of the courses were in public health while the remaining four were distributed across biology, chemistry and physics. These courses were the subject of a large research project, where our research team administered a series of standard survey questions about students' motivation, self-regulation and other psychological constructs before, during and/or after each course. Therefore, we had valid survey data across multiple courses. Also, looking at online courses ensured that LMS data can provide holistic representations of learning behavior. A total of 2,244 students were in the original dataset, and after data cleaning as described below in Section 4.3.2, the final sample size was 2,093. Traditionally underrepresented groups in STEM fields made up a large portion of the sample: 72% were female, 48% came from low-income families, 54% were first generation college students, 33% were underrepresented minorities (URM)[2], and 13% were transfer students.

4.3.2 Features and Outcomes

From each of the three data sources, we constructed a separate feature set in line with the literature. Table 5.1 gives a summary of these features. *Institutional features* included student demographics and academic achievement prior to college. *Click features* were derived from the LMS data and only included general measures of behavioral engagement to accommodate

[2] This includes African American, Hispanic, and Native American students.

Table 4.1: Features derived from the three data sources

Institutional	Click	Survey
Female	Total clicks	Effort regulation
Transfer	Total clicks by category	Time management
Low income	Total time	Environment management
First-gen	Total time by category	Self-efficacy
URM	(All above for the first 5 weeks)	
SAT total score		
High school GPA		

Table 4.2: Details of survey features

Feature	Items (5-point Likert scale)
Effort regulation	I often feel so lazy or bored when I study that I quit before I finish what I planned to do (reverse coded). I work hard to do well in courses even if I don't like what I am doing. When coursework is difficult, I give up or only study the easy parts (reverse coded). Even when course materials are dull and uninteresting, I manage to keep working until I finish.
Time management	I keep a record of what my assignments are and when they are due. I plan my work in advance so that I could turn in my assignments on time.
Environment management	I usually work in a place where I can read and work on assignments without distractions. I can ignore distractions around me when I study.
Self-efficacy	I'm certain I can master the skills taught in this course. I'm certain I can figure out how to learn even the most difficult course material. I can do almost all the work in class if I don't give up.

Notes: Each feature was calculated as the average of its associated items.

the variances in course design. Specifically, for each student in each course, we calculated the total number of clicks and total time spent over the first half of the course period. Time spent was calculated as the time lapse between adjacent click events. For the last click event of a student (with no subsequent event) or exceptionally lengthy lapses, we set a heuristic value of 90 seconds. The click counts and time spent were also broken down by categories, which were defined based on the URLs that click events pointed to, including "portal", "tasks", "content", "communication", "performance" and "miscellaneous." Restricting to the first half of course period speaks to the scenario of early identification of at-risk students for instructors. *Survey features* included four constructs of self-regulated learning skills and self-efficacy (Pintrich and De Groot, 1990) from pre-course surveys launched during the first week of these courses. The completion rates of these surveys ranged from 65% to 93% across the ten courses. All survey items were adapted from Motivated Strategies for Learning Questionnaire (MSLQ), a popular questionnaire to measure self-regulation skills in online learning (Pintrich et al., 1991). Each of the four constructs was measured by the average of corresponding survey items (Table 4.2).

As for outcomes, we defined two success measures. *Short-term success* was defined as a binary indicator of whether a student's final course grade was above the class median. Predicting this within-course outcome aligns with the needs of instructors to recognize struggling students in a timely manner (Forteza et al., 2017). Similarly, *long-term success* was defined as whether a student's average GPA in the year that followed the course was above the median of their classmates in that course. Predicting this longer-term outcome is of interest to academic advisors and institutional policymakers because it can help them make appropriate policy changes early in students' academic careers to increase student success and graduation rates (Luo and Pardos, 2018). We used class medians to construct these outcomes instead of certain grade thresholds in order to better compare short-term and long-term results.

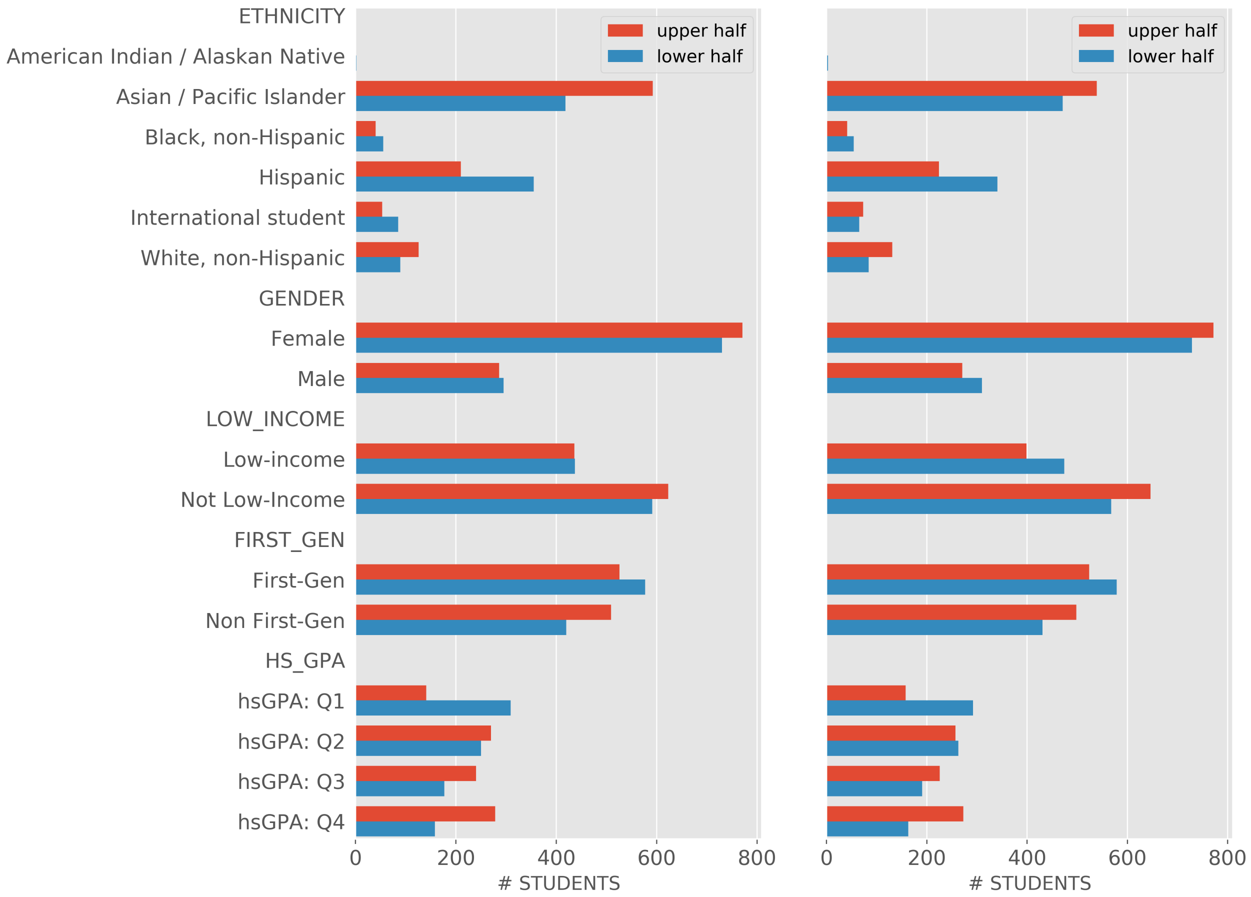
We examined all possible combinations of the three feature sets ($2^3 - 1 = 7$) regarding their ability to predict the two success measures. Therefore, a total of 14 binary classification problems were formulated. To fairly compare the prediction performance of these feature sets, students with missing values on more than 25% of all the individual features in Table 5.1 were dropped, which accounted for the decrease in sample size from 2,244 to 2,093. All continuous numerical features were standardized by centering to the median and scaling according to the interquartile range (IQR) to better handle outliers. For the remaining missing values, we performed multivariate imputation, i.e., modeling each feature with missing values as a function of other features.

4.3.3 Predictive Models

For each classification problem, we employed three common classification algorithms: logistic regression, support vector machines (SVM), and random forests. Course-level leave-one-group-out cross validation was used. In other words, the algorithm looped through the ten courses, and in each iteration used one course as the test set for the model trained on the remaining nine courses. Predicted values for each course were then put together from the ten iterations to evaluate the overall prediction performance. As our focus was the predictive power of different feature sets instead of models, we chose the classifier that produced the highest F-score for each combination of feature set and outcome. Because we used median splits to construct outcomes, class imbalance was not a concern and therefore no resampling was performed. The entire predictive modeling process was implemented using the scikit-learn Python library (Pedregosa et al., 2011).

4.3.4 Evaluation

We evaluated the prediction results via three metrics. Accuracy measures the overall predictive power of the features used. False positive rate (FPR) reflects the probability of



(a) Short-term success

(b) Long-term success

Short-term success: whether a student's final course grade was above the class median. Long-term success: whether a student's average GPA in the following academic year was above the class median.

Figure 4.1: Outcome distribution within different student subpopulations

missing out "at-risk" students or "overplacing" students. False negative rate (FNR), on the other hand, captures the chances of "underplacing" students (Scott-Clayton, 2012). These metrics can shed light on potential consequences of using certain data source(s) in different applications. From there, we can compare the utility of different data sources in a holistic manner.

We further evaluated each data source's contribution to the fairness of prediction results. Fairness was conceptualized as the performance parity across student subpopulations when the prediction was performed on the entire student sample. Specifically, we focused on an array of historically disadvantaged subpopulations and compared each of them with a corresponding reference group on the three metrics. For example, we compared the accuracy, FPR and FNR within Latinx students with those within white students. Figure 4.1a and 4.1b plot the group size and outcome distribution of these selected groups, where the last group under each category was the reference group.

Statistically, we computed the following disparity metrics for each disadvantaged group g:

$$acc\_disparity = acc_{ref}/acc_g \tag{4.1}$$

$$fpr\_disparity = fpr_g/fpr_{ref} \tag{4.2}$$

$$fnr\_disparity = fnr_g/fnr_{ref} \tag{4.3}$$

and separately tested whether each of this disparities was significantly larger than 1 using one-sided two proportion z-test. The larger these ratios were, the more this student group was "discriminated against" by the prediction model. We used the less flexible one-sided test because of the consistent evidence that traditionally underrepresented groups experience more inequities than their counterparts in academic settings (Allen-Ramdial and Campbell, 2014). All these ratios combined would characterize the comparative utility of different data sources for fair predictions of college success.

Table 4.3: Prediction performance on the entire student sample

Feature	Accuracy		FPR		FNR	
	Short	Long	Short	Long	Short	Long
Institutional	0.618	0.599	0.467	0.412	**0.299**	0.389
Click	0.602	0.613	0.485	0.385	0.313	0.389
Survey	0.534	0.557	0.599	0.385	0.336	0.502
Institutional+Click	0.670	**0.650**	0.351	**0.330**	0.310	0.370
Institutional+Survey	0.633	0.608	0.398	0.397	0.337	0.386
Click+Survey	0.609	0.604	0.431	0.457	0.353	0.335
Institutional+Click+Survey	**0.675**	0.638	**0.348**	0.402	0.303	**0.323**

Notes: The best result in each column was in bold. Short: predicting whether a student's final course grade was above the class median; long: predicting whether a student's average GPA in the following academic year was above the class median.

4.4 Predictive Utility of Different Data Sources

4.4.1 Overall Prediction Performance

Table 4.3 presents the prediction results on our full student sample across different feature and outcome combinations. In each column, the best-performing model is in bold to indicate which feature set(s) best predicted the corresponding outcome in the column header in terms of the given metric. Among the final sample of 2,093 students, 1,062 (50.7%) had short-term and 1,048 (50.1%) had long-term outcomes above their class median[3]. These numbers serve as the naïve baselines of prediction accuracy where all the students were simply predicted to be in the upper half (majority class).

When the three data sources were used separately, institutional features and click features both achieved an overall accuracy of around 0.6 for either short-term or long-term outcomes,

[3] The slight deviation from 50% was due to the drop of students with too much missing information on predictors, as described in Section 4.3.2.

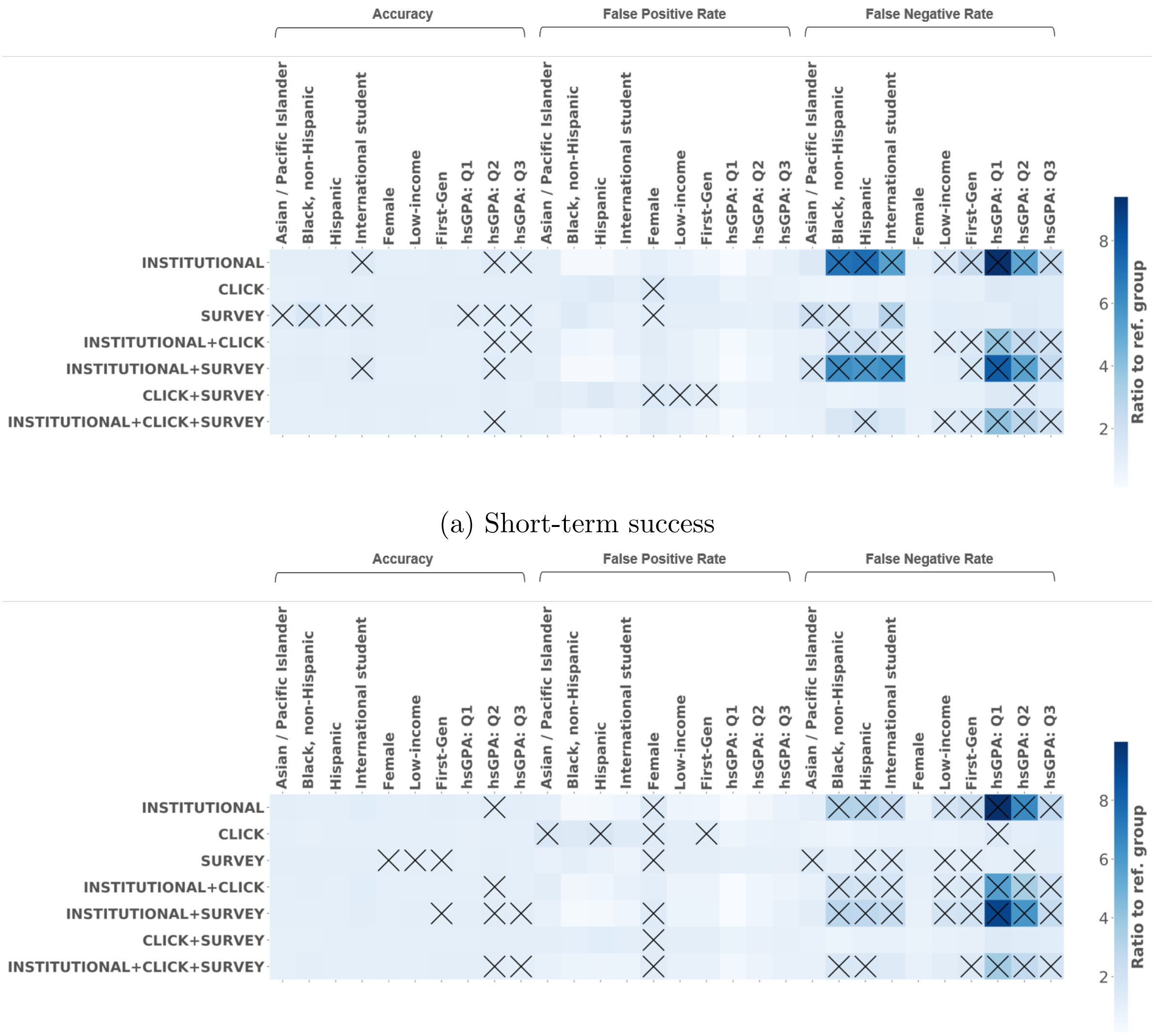
which was significantly higher than the baseline ($p < 0.001$ for all four cases). Specifically, institutional features appeared to be slightly more predictive of short-term success and click features predicted long-term success a little better, but neither of these comparisons was statistically significant. On the contrary, survey features had much weaker predictive utility because they predicted both outcomes with significantly lower accuracy than the worse of the other two features ($p < 0.001$ for short term and $p = 0.005$ for long term). When these feature sets were combined in different ways, we mostly saw improvement in the overall accuracy. The combination of institutional and LMS data led to the most noticeable accuracy increase in predicting both outcomes ($\Delta = 0.052, p < 0.001$ for short term and $\Delta = 0.037, p = 0.014$ for long term), evidencing complementary signals of student success in these two data sources. Survey data provided limited marginal utility as adding survey features to other feature sets never led to a statistically significant increase in accuracy and sometimes even had negative effects. However, the highest accuracy in predicting the short-term outcome was achieved when all three feature sets were used together.

Given the tradeoff between false positives and false negatives, overall best-performing feature sets did not necessarily have the lowest error rates. Among the three cases using a single data source, institutional features had both the lowest FPR and the lowest FNR for the short-term outcome ($p = 0.402$ for FPR and $p < 0.001$ for FNR compared to the second lowest). The same features also tied with click features for the lowest FNR in predicting the long-term outcome, while the latter led to the lowest FPR in the long term (tied with survey features). Combining these two data sources significantly lowered FPR ($\Delta = -0.116, p < 0.001$ for the short term and $\Delta = -0.055, p = 0.009$ for the long term) but not FNR. As for survey data, the patterns of error rates were more complicated than of overall accuracy. When used alone, survey features mostly led to higher error rates than the other two feature sets, except for FPR in the long term. On the other hand, adding survey features to other feature sets largely decreased FNR for long-term and FPR for short-term success predictions despite the fact that these metrics were exceptionally high in the case of using survey data alone.

4.4.2 Fairness of Predictions

Following Section 4.3.4, we computed and tested the extent to which each disadvantaged student subpopulation suffered discriminatory predictions (i.e., algorithmic bias) compared to their reference group under each combination of feature set and outcome. Figure 4.2a and 4.2b illustrate these results for short-term and long-term success prediction, respectively. Each cell colors a bias against a certain student subpopulation in a specific model. Darker cells suggest larger biases and crossed out cells represent those that were statistically significant ($p < 0.05$) after correcting for multiple testing within each background attribute. Subpopulations with fewer than 10 students were omitted as the error rates were less reliable.

Overall, there was no feature set that was entirely free from biased predictions. Across both outcomes, institutional features consistently led to higher FNR within various disadvantaged student subpopulations than within their peers. In other words, these students were more likely to be *underestimated* by the prediction model. This finding resonates with previous research that being aware of protected attributes (e.g., ethnicity) might induce identity-based biases in predictive analytics (Barocas et al., 2019). Adding other features to institutional ones alleviated some of these biases only in a marginal sense. That is, inclusion of institutional features seemed to largely determine the discriminatory behaviors of the model. Identity-blind LMS data was a fairer data source as the number of discriminated subpopulations was smaller. Compared to their reference groups, click features on their own significantly *overestimated* female students for both outcomes and Asian, Hispanic and first-generation college students for the long-term outcome. Survey data turned out to be neither accurate nor fair. When used alone, survey features led to significant biases against certain subpopulations across all metrics and outcomes. When combined with other feature sets, they did little to offset existing biases in most cases, except when they were used together with click features to predict long-term success. However, this latter case may suggest that survey data had equally low predictive utility for long-term success across different student



(a) Short-term success

(b) Long-term success

Notes: Each cell represents the algorithmic bias against a historically disadvantaged student subpopulation (compared to the corresponding reference group) in the specific scenario. Crosses represent statistically significant biases ($p < 0.05$) after correcting for multiple testing. Short-term success: whether a student's final course grade was above the class median. Long-term success: whether a student's average GPA in the following academic year was above the class median.

Figure 4.2: Illustration of prediction fairness

subgroups.

The plots also allowed for insights into the extent to which different student subpopulations were exposed to algorithmic biases across different scenarios. Ethnic minorities, students from low-income families and first-generation college students were more prone to underestimation. Female students were more likely to be overestimated than male students especially in the long term. Moreover, international students and students with lower high school GPAs suffered both more underestimation and less accurate predictions compared to their peers. Note that unlike other variables in the plots, high school GPA is an acquired attribute. Hence, our evidence of algorithmic bias implied that a student can be stigmatized due not only to their demographic attributes but to their past (academic) experience as well.

A Closer Look into Institutional Data

Reflecting on the consistent biases against disadvantaged student subpopulations when using institutional data, we also tested if removing a specific institutional feature (e.g., gender) would eliminate the bias against the corresponding disadvantaged group (e.g., female). Surprisingly, all the results looked qualitatively similar regardless of which feature we removed. This suggested the intersectionality of minority identities, i.e., a student from one disadvantaged group tended to have another disadvantaged characteristic as well. As such, simply removing individual background variables would not necessarily make the predictions fairer.

4.5 Discussions

4.5.1 Reflections on the Results

Our results shed light on the predictive validity of different sources of student data on college success. Our overall results agree well with those of Whitmer et al. (2019), where features from an assessment of socio-emotional skills were least predictive of course success, which is similar to the ineffectiveness of our survey data. On the other hand, they found that models using institutional variables and clickstream features performed better and comparably to one another, as we did. They also discovered that combining clickstream behaviors with socio-emotional skills outperformed institutional data alone, which we also saw with the FNR for the long-term outcome. Interestingly, they did not find additional predictive utility of higher-level behaviors (sequential features) from clickstream data, which we did not further investigate.

The limited ability of pre-course survey data to accurately predict either short-term or long-term success may suggest that self-reported measures of self-regulated learning are not key factors of online learning processes or performance. However, as suggested by previous research (DiBenedetto and Bembenutty, 2013), it may also suggest that students tend to overestimate their use of learning strategies in online courses. This is likely because students make estimations of their future behaviors based on memories of similar past events that are usually unreliable (Li et al., 2020). Thus, more research is needed to understand how to help students provide valid data of their learning skills as well as other psychological attributes in surveys (Osterhage et al., 2019).

When it comes to fairness, several interesting trends emerge. First, predictions using institutional data, which had the lowest FNR overall, were actually discriminatory when it comes to FNR for both outcomes. In particular, institutional data discriminated against students from underrepresented minority groups, low-income students, first-generation college students, and students with low high school GPA. This suggests that these models tend to disproportionately label students from these subpopulations as having below-median performance. In order to achieve higher overall accuracy, these models appear to be using a heuristic of classifying students as above or below median based on the majority class within the subpopulations that they belong to (see Figure 4.1). Therefore, one of the main sources of unfairness may just be the original class imbalance in different student subpopulations. When this imbalance results from historical inequities, the model will simply replicate those inequities and produce unfair predictions.

On the other hand, we found that using click features tended to be fair with respect to FNR, but instead somewhat discriminatory with respect to FPR, for several student subpopulations. Contrary to the discrimination brought by institutional features, this form of discrimination could occur just because the model is blind to individual background. More specifically, students coming from different backgrounds may on average exhibit similar learning behaviors, but their likelihood to succeed might differ due to factors that correlate with their socio-economic status. Since the click features do not have access to students' background information, they may predict that students from disadvantaged backgrounds are likely to succeed at a disproportionately high rate.

One specific and possibly counterintuitive trend is seen when it comes to gender biases. While none of the feature sets discriminated against female students in terms of FNR, almost all of the feature sets discriminated against them in terms of FPR for at least one of the two outcomes. In fact, female students tend to have higher GPA than their male peers in the dataset (see Figure 4.1). This reinforces the inference that for institutional features, the models classify students into the majority class of their subpopulations in order to maximize accuracy. On the other hand, the fact that using only LMS and/or survey data is also biased against female students in terms of FPR might be due to something else. This suggests that female students might (a) exhibit different click behaviors and survey responses from men, which tend to be predictive of better performance; or (b) have different baseline levels of engagement (e.g., likelihood of clicking on LMS pages) independent of their likelihood of success. If the former is true, click behaviors and/or survey responses could act as a weak proxy for gender, even though gender is not encoded in these features.

4.5.2 Practical Implications

In general, prediction errors are inevitable, but it is important to be aware of and minimize potential misplacement that may result in severe negative consequences. Below, we discuss three major scenarios where prediction models are used for educational decision making and the implications of our findings in these cases.

First, higher education has a long history of screening applicants for desirable educational opportunities such as merit-based scholarships, where the award is based on the prediction of student future performance. In this case, underestimating student performance may limit their educational development. While institutional data is one of the most widely used data sources for these purposes, our results suggest that institutional data alone might be more likely to underestimate achievement of students from disadvantaged background as compared to their peers. Moreover, these systematic biases do not go away easily even when other common data sources are added. Therefore, it is important for policymakers to cautiously employ predictive analytics for selecting students since it may result in unfair exclusion of already disadvantaged students from critical educational opportunities and access to social mobility through education (Haveman and Smeeding, 2006).

In community college settings, institutional data has also been used to evaluate students' readiness for college-level courses and assign students into remediation (Scott-Clayton, 2012), as well as to understand the impact of remedial and preparatory courses on subsequent college success (Nguyen et al., 2020). Put in this scenario, our results would suggest that students from historically disadvantaged subpopulations are more likely to be misplaced into remediation than their counterparts when they are actually capable of taking advanced courses. While remedial courses are designed to help academically underprepared students, they also increase students' cost and may delay student progression towards their degree goal (Bailey et al., 2010). For both this and the previous application scenarios, a potential algorithmic solution might be setting separate thresholds for different subpopulations to ensure fairness, as Kleinberg et al. (2018) suggested.

Finally, in the recent research and practice of online learning, LMS data have been commonly used to predict student performance and identify at-risk students (Wolff et al., 2013). Students who are identified as being at risk of low performance or dropout will often be placed into light-touch or optional academic support, such as receiving email reminders and tutoring services (Choi et al., 2018). In this context, it might be more concerning to overestimate student performance and ignore students in need than to underestimate student performance and place them to educational resources that they could opt out of. Our findings indicate that compared to males, female students would be especially likely to experience overestimation and therefore would not receive academic resources that they need. In this case, incorporating institutional data into the prediction might not be as problematic in order to leave no student behind.

4.6 Conclusion

In this paper, we responded to the call for research to evaluate and compare the utility of common student data sources (i.e., institutional data, LMS data and survey data) for building predictive analytics applications in the context of higher education (Fischer et al., 2020). We aimed to find out what data sources and their combinations predicted short-term and long-term college success both accurately and fairly across different student subpopulations. Our results suggest that overall, institutional data and LMS data on their own have decent predictive utility for either instructors' or policymakers' needs to identify students in need. Using them together further strengthens that predictive power. Survey data alone poorly predicts student success and only marginally helps alleviate some of the prediction errors in the presence of other data sources. With regard to fairness, institutional data consistently leads to higher false negative rate (underestimation) within historically disadvantaged students subpopulations than within their peers. LMS data, on the other hand, tends to overestimate some of these disadvantaged groups (e.g., female students) more often than their counterparts and these biases would be overridden by institutional data when the latter is added. Survey data makes very limited contribution to fair predictions. Interestingly, all sources of student data tend to overestimate female students who perform better than male students on average in our case. Also, students with lower prior achievement are no less affected by underestimation than underrepresented demographic groups.

These results combined suggest that using multiple data sources in college success prediction is beneficial for institutional stakeholders from both technical and ethical perspectives. Specifically, given the infancy and decent predictive utility of LMS data, institutions should feel encouraged to invest in the infrastructure to store, manage and analyze such data and integrate LMS-based behavioral measures into the routines of institutional research. On the other hand, utilizing multiple data sources still cannot guarantee fair predictions of college success especially for students who have less competitive academic records and who are historically disadvantaged in higher education. Therefore, it is advisable to combine the intelligence of experienced practitioners and data-driven applications for decision-making in the wild, in hopes of minimizing the risk that students are unfairly excluded from their optimal pathways due to biased algorithms or human judgement.

Our work has a few limitations which point to meaningful future work. First, the scope of our feature sets was limited and not representative of the full potential of different data sources. For example, for survey features we only used measures of self-regulation, but there are other psychological constructs that play equally important roles in learning processes. Therefore, our findings should be taken as a proof of concept in terms of systematically evaluating different data sources. Future work will extend the current piece to more comprehensive data sources that institutions have good control over (Hutt et al., 2019; Aulck et al., 2019) and to broader feature sets informed by existing research. Second, while we briefly reflected on the prediction results and practical implications, we did not formally examine how the biases illustrated in Figure 4.2 permeate through the predictive analytics pipeline. Future work will examine this aspect more thoroughly, as well as how to convey these sources of bias to stakeholders for more prudent decision-making on student data usage.

Chapter 5

Should College Dropout Prediction Models Include Protected Attributes?

5.1 Introduction

With the rapid development of learning analytics in higher education, data-driven instructional and learning support systems are increasingly adopted in classroom settings, and institution-level analytics systems are used to optimize resource allocation and support student success on a large scale. A common objective of these systems is the early identification of at-risk students, especially those likely to drop out of college. This type of prediction has significant policy implications because reducing college attrition has been a central task for institutional stakeholders ever since higher education was made accessible to the general public (Pantages and Creedon, 1978). As of 2018, fewer than two-thirds of college students in the United States graduated within six years, and this share is even smaller at the least selective institutions which serve disproportionately more students from disadvantaged backgrounds (Hussar et al., 2020). At the same time, the supply of academic, student affairs, and administrative personnel is insufficient to provide just-in-time support to students in need (Hussar et al., 2020). It is within these resource-strained contexts that predicting dropouts based on increasingly digitized institutional data has the potential to augment the capacity of professionals who work to support student retention and success. Starting with the Course Signals project at Purdue University, an increasing number of early warning systems have explored this possibility at the institutional level (Arnold and Pistilli, 2012; Jayaprakash et al., 2014; Ekowo and Palmer, 2016; Dawson et al., 2017).

Accurately forecasting which students are likely to drop out is essentially profiling students based on a multitude of student attributes. These attributes often include socio-demographic information that is routinely studied in higher education research. Although the analysis of historical socio-demographic gaps in retention and graduation rates is well established in higher education research (de Brey et al., 2019), it becomes controversial to use these same characteristics when making predictions about the future. For example, is it fair to label a black first-year student as *at risk* based on the higher dropout rate among black students in previous cohorts? The answer may be equivocal (Shum, 2020). On the one hand, the observed historical gaps capture systematic inequalities in the educational environment of different student groups, which may well apply to future students from the same groups and therefore contribute to similar gaps. In this sense, explicitly using socio-demographic data can result in more accurate predictions and improve the efficiency of downstream interventions and actions based on those algorithmic decisions (Paquette et al., 2020). On the other hand, from an ethics and equity perspective, the inclusion of socio-demographic variables may lead to discriminatory results if predictive models systematically assign differential predicted values across student groups based on the records of their historical counterparts. When these results are used for decision-making, stigmas and stereotypes could carry over to future students and reproduce existing inequalities (Kizilcec and Lee, 2020; Barocas et al., 2019).

In this paper, we investigate the issue of using protected attributes in college dropout prediction in real-world contexts. Protected attributes are traits or characteristics based on which discrimination is prescribed as illegal, such as gender, race, age, religion, and genetic information. We examine students in a residential college setting as well as students in fully online degree programs, which have been increasingly represented in formal higher education. In Fall 2018, 16.6% of postsecondary students in the United States were enrolled in exclusively online programs, up from 12.8% in Fall 2012 (Seaman et al., 2018; Snyder et al., 2019). The absence of a residential experience exposes students to additional challenges to accountability and engagement, and also makes it harder for faculty and staff members to identify problems with students' well-being and provide timely support. The COVID-19 pandemic has forced most colleges to move instruction online, which will likely increase the importance of online learning in the future of higher education (The Chronicle of Higher Education, 2020). Predictive analytics are therefore just as useful for online higher education as they are for residential settings for supporting student achievement and on-time graduation. Our findings in both residential and online settings offer practical implications to a broad range of stakeholders in higher education.

By systematically comparing predictive models with and without protected attributes in two higher education contexts, we aim to answer the following two research questions:

1. How does the inclusion of protected attributes affect the overall performance of college dropout prediction?

2. How does the inclusion of protected attributes affect the fairness of college dropout prediction?

This research contributes to the literature on predictive modeling and algorithmic fairness in (higher) education on several dimensions. First, we present one of the largest and most comprehensive evaluation studies of college dropout prediction based on student data over multiple years from a large public research university. This offers robust insights to researchers and institutional stakeholders into how these models work and where they might go wrong. Second, we apply the prediction models with the same features to both residential and online degree settings, which advances our understanding of generalizability across contexts, such as in which environment it is easier to predict dropout and to what degree key predictors differ. Third, we contribute some of the first empirical evidence on how the inclusion of protected attributes affects the fairness of dropout prediction, which can inform equitable higher education policy around the use of predictive modeling.

5.2 Related Work

5.2.1 College Dropout Prediction

Decades of research have charted the ecosystem of higher education as a complex journey with "a wide path with twists, turns, detours, roundabouts, and occasional dead ends that many students encounter" and jointly shape their academic and career outcomes (Kuh et al., 2007). Among the variety of factors that influence students' journey, background characteristics such as demographics, family background, and prior academic history are strong signals of academic, social, and economic resources available to a student before adulthood, which are substantially correlated with college success (Coleman, 1988). For example, ethnic minorities, students from low-income families, and first-generation college students have consistently suffered higher dropout rates than their counterparts (de Brey et al., 2019; Cataldi et al., 2018), and students who belong to more than one of these groups are even more likely to drop out of college. In addition to these largely immutable attributes at college entry, students' experiences in college such as engagement and performance in academic activities are major factors for success. In particular, early course grades are among the best predictors of persistence and graduation, even after controlling for background characteristics (Kuh et al., 2007).

With the advent of the "datafication" of higher education (Selwyn and Gašević, 2020), there has been an increasing thrust of research to translate the empirical understanding of dropout risk factors into predictive models of student dropout (or success) using large-scale administrative data (Aulck et al., 2019; Dekker et al., 2009; Jayaprakash et al., 2014; Del Bonifro et al., 2020; Beaulac and Rosenthal, 2019; Berens et al., 2019; Hutt et al., 2019). These applications are usually intended to facilitate targeted student support and intervention programs, and the extensive research literature on college success has facilitated feature engineering grounded in theory. For example, Aulck et al. (2019) used seven groups of freshman features extracted from registrar data to predict outcomes for the entire student population at a large public university in the US. The model achieved an accuracy of 83.2% for graduation prediction and 95.3% for retention. In a more application-oriented study as part of the Open Academic Analytics Initiative (OAAI),Jayaprakash et al. (2014) developed an early alert system that incorporated administrative and learning management system data to predict at-risk students (those who are not in good standing) at a small private college, and then tested the system at four other less-selective colleges.

While the recent decade has seen a steady growth in prediction-focused studies on college dropout, a large proportion of them are focused on individual courses or a small sample of degree programs (Hellas et al., 2018). Most of them investigate dropouts at brick-and-mortar institutions. Our study pushes these research boundaries by examining dropout prediction for multiple cohorts of students across residential and exclusively online degree programs offered by a large public university. The breath of our sample is rare in the dropout prediction literature and promises to offer more generalizable insights about the utility and feasibility of predictive models.

5.2.2 Algorithmic Fairness in Education

A central goal of educational research and practice has been to close opportunity and achievement gaps between different groups of students. More recently, algorithmic fairness has become a topic of interest as an increasing number of students are exposed to intelligent educational technologies (Kizilcec and Lee, 2020). Inaccuracies in models might translate into severe consequences for individual students, such as failing to allocate remedial resources to struggling learners. It is more concerning if such inaccuracies disproportionately fall upon students from disadvantaged backgrounds and worsen existing inequalities. In this context, the fairness of algorithmic systems is generally evaluated with respect to protected attributes following legal terms. The specific criteria of fairness, however, vary and largely depend on the specific application(s) (Verma and Rubin, 2018).

In the past few years, a handful of papers have brought the fairness framework to real-world learning analytics research. Most of these studies audit whether supervised learning models trained on the entire student population generate systematically biased predictions of individual outcomes such as correct answers, test scores, course grades, and graduation (Yu et al., 2020; Kung and Yu, 2020; Gardner et al., 2019; Hutt et al., 2019; Doroudi and Brunskill, 2019; Loukina et al., 2019). For example, Yu et al. (2020) found that models using college-entry characteristics to predict course grades and GPA tend to predict lower values for underrepresented student groups than their counterparts. Other studies have examined biases encoded in unsupervised representations of student writing (Arthurs and Alvero, 2020), or go further to refine algorithms for at-risk student identification under fairness constraints (Hu and Rangwala, 2020). Overall, this area of research is nascent and in need of systematic frameworks specific to educational contexts to map an agenda for future research.

When it comes to strategies to improve algorithmic fairness, a contentious point is whether protected attributes should be included as predictors (features) in prediction models. Most training data from the real world are the result of historical prejudices against certain protected groups, so directly using group indicators to predict outcomes risks imposing unfair stereotypes and reproduce existing inequalities (Barocas et al., 2019). In educational settings, it may be considered unethical to label students from certain groups as ”at risk” from day one, when in fact, these students have demonstrated an exceptional ability to overcome historical obstacles and might therefore be more likely to succeed (Shum, 2020). This concern motivated the research effort to “blind” prediction models by simply removing protected attributes (i.e. fairness through unawareness) or more complicated statistical techniques to disentangle signals of protected attributes from other features due to their inherent correlation (Calmon et al., 2017). In contrast, recent work has advocated for explicitly using protected attributes in predictive models (i.e. fairness through awareness) (Dwork et al., 2012). In particular, Kleinberg et al. (2018) showed in a synthetic example of college admission that the inclusion of race as a predictor of college success improves the fairness of admission decisions without sacrificing efficiency. Given the well-documented relationship between student background and their educational outcomes, a recent review also suggests that predictive models in education should include demographic variables to ensure that algorithms are value-aligned, i.e., all students have their needs met (Paquette et al., 2020).

To our knowledge, however, there is only limited empirical evidence to support either side of this debate. Our study therefore presents an in-depth examination of the consequences of including or excluding protected attributes on algorithmic fairness of a realistic, large-scale dropout prediction model.

Table 5.1: Features used for dropout prediction

Category	Features
Protected attributes	Gender (binary), first-generation college student (binary), underrepresented minority (URM; binary; defined as not Asian or White), high financial need (binary; FASFA-based expected family contribution under $5,500)
Incoming attributes	Age, high school GPA, math and verbal SAT/ACT scores, transfer student (binary), transferred credits, transfer GPA
Program information	Part-time student (binary), major, minor, STEM major (binary)
Course performance	Total courses enrolled, total units enrolled, percentage of courses that are required, credits received from different types of courses (lecture, seminar, etc.), levels of courses (100, 200, etc.), term GPA, mean and variance in course grades within each session during the term, percentage distribution of letter grades

5.3 Methodology

5.3.1 Dataset

We analyze de-identified institutional records from one of the largest public universities in the United States. This broad-access research university serves nearly 150,000 students with an 86% acceptance rate and 67% graduation rate. Its student population is representative of the state in which it is located, which makes it a Hispanic-serving institution (HSI). The university has offered many of the same undergraduate degree programs fully online to over 40,000 students. The dataset we use in this study focuses on undergraduate students and contains student-level characteristics and student-course-level records for their first term of enrollment at the university, including transfer students (except for those who transfer into their senior year). For our prediction task, we only keep students whose first term was in the Fall along with their course-taking records in their first term, including terms between 2012-18 (residential) and 2014-18 (online).

This sample comprises a total of 564,104 residential course-taking records for 93,457 unique students and 2,877 unique courses, and 81,858 online course-taking records for 24,198 unique students and 874 unique courses. The course-taking records include both a student's letter grade and course-level metadata (subject, course number, units, required for major, etc.). Student-level information includes socio-demographic information (age, gender, race/ethnicity, first-generation status, etc.), prior academic achievement (high school GPA, standardized test scores), enrollment information (transfer student status, part-time status, academic major and minor, etc.). These data are representative of what most higher education institutions routinely manage in their student information systems (SIS) (Aulck et al., 2019).

5.3.2 Prediction Target and Feature Engineering

The primary goal of a dropout prediction model is to alert relevant stakeholders to currently enrolled students who are at risk of dropping out of a degree program so that they can reach out and offer support at an early stage. While the general framework of dropout prediction is well established, the exact definition of dropout, or attrition, varies based on the specific context (Pantages and Creedon, 1978). In our context, we define dropout as not returning to school a year from the first time of enrollment. We only analyze students who first enrolled in Fall, so dropout means not returning in the following Fall. This final operationalization aligns well with retention, one of the two standard metrics of post-secondary student success in national reports of the United States (Snyder et al., 2019; Hussar et al., 2020).[1]

We use students' background characteristics and academic records in the first enrolled term (Fall) to predict dropout, because it would be beneficial to identify risks as early as possible and institutional records are usually updated and available at the end of each term. Informed by existing research in higher education and learning analytics (see Related Work),

[1] The other standard metric is graduation within 100% or 150% of the normative time (i.e. 4 or 6 years for four-year institutions). We do not examine this metric because the span of our dataset is only six years and we do not observe graduation outcomes for all student cohorts.

Table 5.2: Comparison of online and residential student populations

	Online	Residential
N	24,198	93,457
Dropout	40.7%	16.9%
Female	60.9%	47.9%
First-gen	42.4%	33.6%
URM	33.1%	34.6%
High need	61.9%	51.3%
Transfer	85.2%	31.8%
Part-time	77.2%	12.9%
Average age	27.1	19.7

we construct 58 features from the dataset for both residential and online students. Table 5.1 summarizes these feature by four categories. We include four protected attributes, which are the most commonly used dimensions along which to examine educational inequalities and set equity goals in policy contexts (Cataldi et al., 2018; Chen and Nunnery, 2019; Hussar et al., 2020).

Table 5.2 depicts the student profile in our analysis. The statistics reaffirm that, regardless of format, the institution serves a large proportion of students from historically disadvantaged groups. There are also major differences across formats. In line with the national statistics of exclusively online programs (Snyder et al., 2019), the online sample has a higher concentration of transfer and non-traditional (older, part-time) students, and also higher dropout rates compared to residential students. These characteristics validate that the current analysis is performed on student populations who are most in need of institutional support and allow us to scrutinize the generalizability of our findings across two distinct contexts of higher education.

5.3.3 Dropout Prediction

To investigate the consequences of using protected attributes in dropout prediction models, we generate two feature sets: the AWARE set includes all features shown in Table 5.1, while the BLIND set excludes the four protected attributes from the AWARE set. For convenience, we will refer to a specific model by the feature set it uses in the remainder of this paper. Given our binary target variable, the dropout prediction task is formalized as a binary classification problem. As we focus on identifying the effect of including protected attributes, we experiment with two commonly used algorithms – logistic regression (LR) and gradient boosted trees (GBT). We choose LR because it is a linear additive and highly interpretable classifier that can achieve reasonable prediction performance with well-chosen features. The choice of GBT, on the other hand, is for its ability to accommodate a large number of features, efficiently handle missing values, and automatically capture non-linear interactions between features.

We predict dropping out separately for online and residential students. For each format, we split the data into a training set and a test set based on student cohort: the last observed cohort (6,939 online and 14,275 residential students entering in Fall 2018) constitutes the test set and the remaining cohorts make up the training set (17,259 online and 79,182 residential students). There are two reasons for doing the train-test split by student cohorts. Practically, this split aligns with the real-world application where stakeholders rely on historical data to make predictions for current students (Jayaprakash et al., 2014). Technically, this approach alleviates the issue of data contamination between the training and test set (Farrow et al., 2019), as the features we use, especially the first-semester records, might be highly correlated within the same cohort but much less so across cohorts.

There are a few additional technical details about model training. First, we tune hyperparameters of the two algorithms by performing grid search over a specified search space and evaluating the hyperparameters using 5-fold cross-validation. Second, we add indicator variables for missing values in course grades, standardized test scores, and academic majors and minors. Third, we apply robust scaling to training features to regulate the influence of outliers. Fourth, because the class imbalance in both datasets can bias the model learning towards the majority class (i.e. non-dropout), we adjust the sample weights to be inversely proportional to class frequencies during the training stage.

The trained classifiers are then applied to the test set to evaluate the performance. The immediate output of each classifier is a predicted probability of dropping out for each student. To make a final binary prediction of dropout, we use dropout rates in the training data to determine the decision thresholds for the test set, such that the proportion of predicted dropouts in the test set matches the proportion of observed dropouts in the training set (Berens et al., 2019). Compared to the default of 0.5, this choice of threshold is more reasonable when we rely on the observed history to predict the unknown future in practice.

5.3.4 Performance Evaluation

We evaluate prediction performance based on three metrics: accuracy, recall, and true negative rate (TNR). In the context of dropout prediction, recall is the proportion of actual dropouts who are correctly identified, whereas TNR quantifies how likely a student who persists into the second year of college is predicted to persist. To examine the effects of including protected attributes on overall performance, we compute these metrics separately for each model and test whether each metric significantly changes from BLIND to AWARE models, using two proportion z-tests.

We operationalize fairness as the independence between prediction performance, measured by the three metrics above, and protected group membership. This definition of fairness with respect to the three metrics corresponds to the established notions of overall accuracy

Table 5.3: Overall prediction performance of AWARE and BLIND models trained with gradient boosted trees (GBT) and logistic regression (LR)

	GBT			LR		
Metric	AWARE	BLIND	Δ	AWARE	BLIND	Δ
Online (Non-dropout: 59.3%)						
Accuracy	75.8	75.6	0.2	75.2	75.4	-0.2
Recall	67.3	67.1	0.2	66.7	66.8	-0.1
TNR	82.4	82.3	0.1	81.9	82.0	-0.1
Residential (Non-dropout: 83.1%)						
Accuracy	83.9	83.9	0.0	83.6	83.6	0.0
Recall	54.1	54.1	0.0	53.2	53.3	-0.1
TNR	89.1	89.1	0.0	88.9	88.9	0.0

Note: None of the Δ values is statistically significant with $p < 0.1$.

equality, equal opportunity, and predictive equality, respectively (Kizilcec and Lee, 2020). Specifically, to quantify the fairness of a given model with regard to a binary protected attribute, such as URM, we compute the differences in each of the three metrics between the two associated protected groups, URM and non-URM students. We then compare how much these differences change between BLIND and AWARE models in order to quantify the effect of including protected attributes as predictors on fairness.

5.4 Results

5.4.1 Overall Prediction Performance

We first illustrate the effects of including protected attributes on overall prediction performance. Table 5.3 reports the overall performance of AWARE and BLIND models, trained with GBT and LR algorithms, on the test dataset. The last column under each algorithm reports the percentage point differences in performance between the two models (from BLIND to AWARE). The main finding is that including or excluding protected attributes does affect

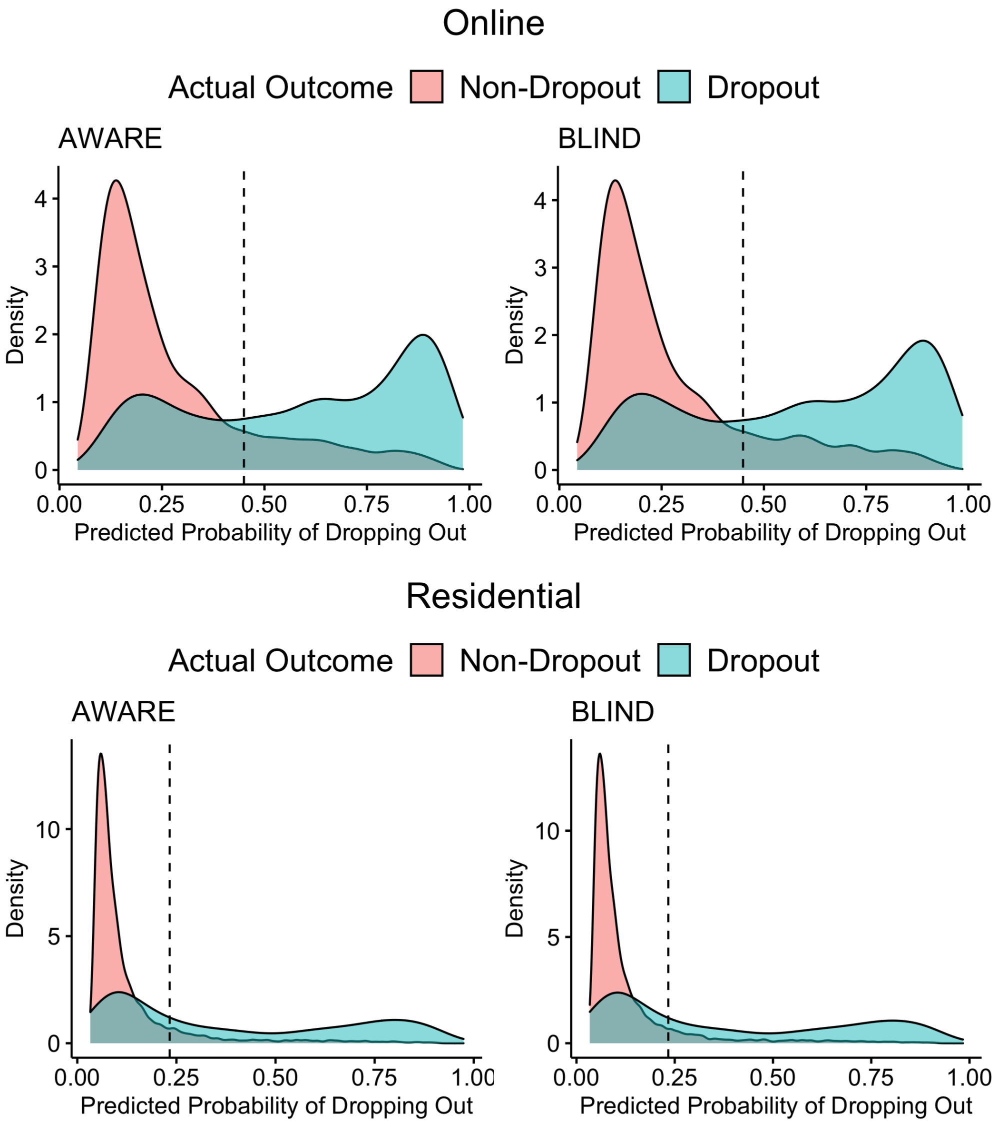


Figure 5.1: Distribution of predicted dropout probability

the performance of the dropout prediction in either context. None of the performance metrics (accuracy, recall, TNR) differs significantly between the BLIND and AWARE models. Additionally, while the more sophisticated GBT algorithm performs better than the simple LR on all metrics, the advantage is comparatively small (less than one percentage point on all metrics). Because of this, we restrict the following analysis to GBT-based models.

Compared to a naïve baseline which simply predicts every student to be the majority class (non-dropout) and achieves an accuracy equal to that majority's share, the predictive models can accurately predict online dropouts with a decent margin. However, the accuracy margin for predicting residential dropouts is fairly small. The other two metrics, which describe the accuracy among dropouts and non-dropouts respectively, achieve a higher value when the corresponding group has a larger share and vice versa. Specifically, the models are able to identify 67.3% of online dropouts and 54.1% of residential dropouts. This latter value is somewhat lower but still comparable to the recall performance in recent prior work on dropout prediction in residential programs (Berens et al., 2019; Del Bonifro et al., 2020).

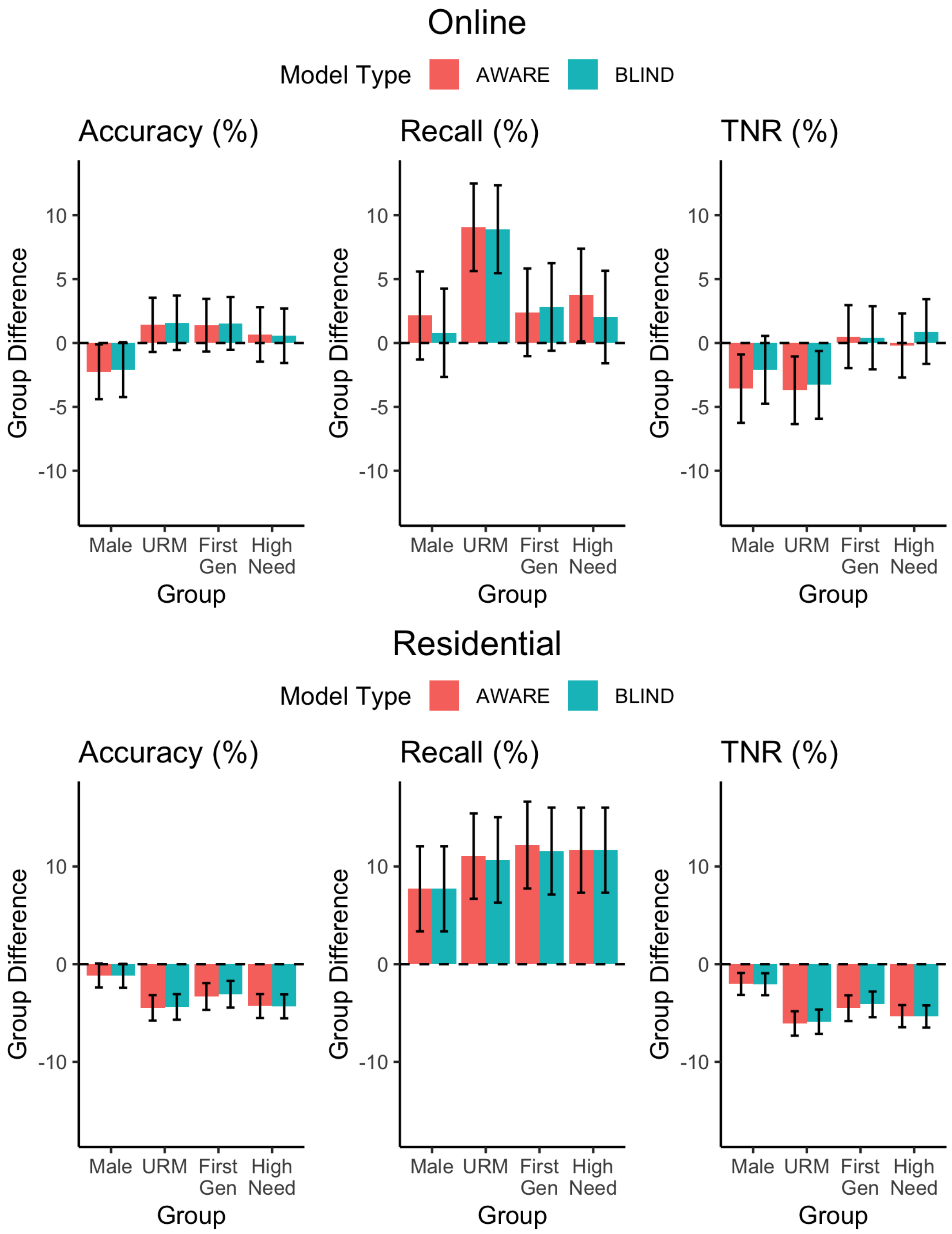
To take a closer look at the model predictions, beyond the three aggregate performance metrics, we examine whether including protected attributes alters the distribution of predicted dropout probabilities. As shown in Figure 5.1, the distributions are highly similar across the models which further validates the limited marginal impact of protected attributes. An additional insight from these plots is that dropouts might be much more heterogeneous than non-dropouts in terms of the features in Table 5.1, as their predicted probabilities are highly spread out, especially in residential settings where the majority of dropouts are assigned a small dropout probability. This pattern is consistent with the lower recall performance shown in Table 5.3.

This finding appears to conflict with prior research that demonstrates the critical role of demographic and background characteristics for student success in higher education (Kuh et al., 2007). In an effort to better understand our result, we explore two mutually compatible hypotheses inspired by the algorithmic fairness literature. One hypothesis is that dropping out, the prediction target, is not sufficiently correlated with protected attributes, and thus adding the latter to a dropout prediction model would not improve performance much. To test this, we fit separately for each enrollment format in the test data a logistic regression model that predicts dropout using all the possible interaction terms between the four protected attributes. We find that, even though a few coefficients are statistically significant, the adjusted McFadden's R^2 is as small as 0.006 for either format, lending support to our hypothesis.

The second hypothesis is that protected attributes are already implicitly encoded in the BLIND feature set, and adding them directly does not add much predictive power. We test this by fitting four logistic regressions for each format which use the BLIND feature set to predict each of the four protected attributes. Based on the adjusted McFadden's R^2, we find that only gender can plausibly be considered encoded in the other features (0.159 for online and 0.187 for residential). This lends partial support to our second hypothesis.

5.4.2 Fairness of Prediction

We further examine how the inclusion of protected attributes might affect the fairness of dropout predictions. As mentioned in the previous section, for each of the four protected attributes, we first measure fairness by the group difference in a chosen performance metric. For example, a prediction model that achieves the same accuracy on male and female students is considered fair in terms of accuracy (0% difference). Following this construction, Figure 5.2 visualizes these fairness results of the AWARE and BLIND models for each of the four protected attributes in terms of the three metrics. Each bar in a subplot depicts the difference in that metric between the labeled group and their counterpart (e.g., male - female). The closer the bar is to zero, the fairer that model prediction is. Overall, the figure

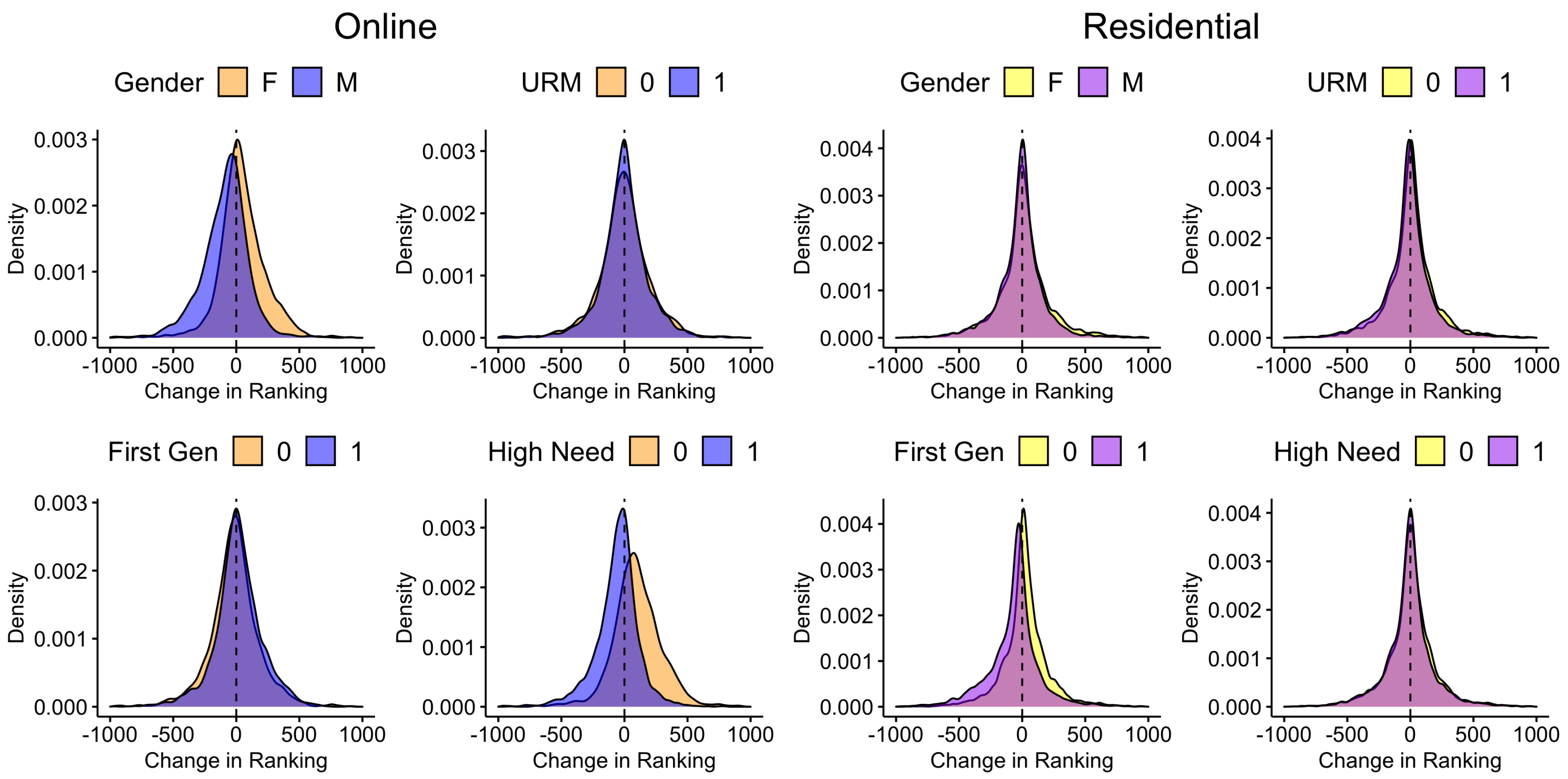


Notes: Positive group differences (y-axis) indicate higher values for the listed groups compared to their corresponding reference groups. Group differences closer to zero reflect higher levels of fairness. Error bars indicate 95% confidence intervals.

Figure 5.2: Fairness of AWARE and BLIND models in terms of accuracy (left), recall (middle), and TNR (right)

shows that both the AWARE and BLIND models are unfair for some protected attributes and some metrics, but fair for others. This lack of universal fairness is expected given the many dimensions of protected attributes, models, and metrics. However, for residential students, the model consistently exhibits unfairness across all protected attributes and metrics, especially in terms of recall. The inclusion or exclusion of protected attributes does not in general lead to different levels of fairness in terms of any metric in any enrollment format, as all adjacent error bars in the figure exhibit a high degree of overlap.

While the aggregated group fairness metrics do not differ with vs. without protected attributes, we take a step further to explore how individual-level changes in model predictions can shed light on the overall change in fairness. We examine changes in the individual ranking of predicted dropout probability among all predicted students (test set) from BLIND to AWARE model. Figure 5.3 plots the distribution of this ranking change for each protected group, where higher values represent moving up in the assigned risk leaderboard when protected attributes are included for prediction.



Notes: One unit increase means going up by one place in the AWARE model compared to the BLIND model.

Figure 5.3: Distribution of change in individual ranking of predicted dropout probability from BLIND to AWARE.

Table 5.4: Dropout rates among different protected groups in the test set

	Dropout rate	
	Online	Residential
Overall	40.7	16.9
Male	49.5	15.6
Female	40.7	14.0
URM	46.6	16.8
Non-URM	42.4	13.7
First-gen	43.7	17.9
Continuing-gen	44.1	13.5
High need	45.8	17.0
Low need	40.5	12.8

We find that overall the ranking change is centered around zero, but there are observable group differences in certain cases. In the online setting, the AWARE model tends to move up females and students without a high financial need on the dropout risk leaderboard simply based on their identity. Similarly, continuing-generation college students are moved up more in residential settings compared to their first-generation counterparts. We argue that these group differences suggest improved fairness if the group going up more in the ranking spectrum has lower dropout rates in reality, and vice versa. To formally evaluate this reasoning, we conduct a series of t-tests between pairs of protected groups on their ranking change. We also compute Cohen's d to gauge the standardized effect size. Comparing Table 5.5 which describes these results and Table 5.4 which presents the actual dropout rates of each group, we find that moving from BLIND to AWARE causes students from advantaged (lower dropout rates) groups to be assigned relatively higher risk rankings compared to their disadvantaged (higher dropout rates) reference groups, and that this effect size is larger when the two paired groups have larger gaps in dropout rates. Thus, adding protected attributes to the model is working against existing inequities to a marginal extent instead of reinforcing them.

Table 5.5: Welch two-sample t-test results and Cohen's d effect size of individual ranking change

Group (Avg. ranking change)		Rank Δ	Cohen's *d*
Online			
Female (50.5)	Male (-88.4)	138.9***	0.71
Non-URM (0.5)	URM (-0.9)	1.4	0.01
Continuing-gen (-13.9)	First-gen (18.6)	-32.5***	0.16
Low need (104.5)	High need (-60.0)	164.5***	0.87
Residential			
Female (15.9)	Male (-15.1)	31.0***	0.13
Non-URM (13.0)	URM (-22.8)	35.8***	0.15
Continuing-gen (27.2)	First-gen (-62.1)	89.3***	0.38
Low need (6.8)	High need (-7.0)	13.8***	0.06

Positive ranking change means increased predicted dropout risks from BLIND to AWARE model. *** p<0.001; ** p<0.005; * p<0.01

5.5 Discussion and Conclusion

We set out to answer a simple question: Should protected attributes be included in college dropout prediction models? This study offers a comprehensive empirical examination of how the inclusion of protected attributes affects the overall performance and fairness of a realistic predictive model. We demonstrate this finding across two large samples of residential and online undergraduate students enrolled at one of the largest public universities in the United States. Our findings show that including four important protected attributes (gender, URM, first-generation student, high financial need) does not have any significant effect on three common measures of overall prediction performance when commonly used features (incoming attributes, enrollment information, academic records) are already in the model. Even when used alone without those features, the group indicators defined by the protected attributes are not highly predictive of dropout, although the actual dropout rates are somewhat higher among minoritized groups. In terms of fairness, we find that including protected attributes only leads to a marginal improvement in fairness by assigning dropout risk scores with smaller gaps between minority and majority groups. However, this trend is not sufficiently large to systematically change the final dropout predictions based on the risk scores, and therefore the formal fairness measures are not significantly different between models with and without protected attributes.

In short, our results suggest limited effects of including protected attributes on the performance of college dropout prediction. This does not point to a clear answer to our normative question and prompts us to further reflect on the focal issue of using protected attributes. Recent work in the broader machine learning community has been in favor of "fairness through awareness" (Dwork et al., 2012), and has specifically suggested that race-aware models are fairer for student success prediction because they allow the influence of certain features to differ across racial groups (Kleinberg et al., 2018). Our findings resonate with these existing studies around fairness but only to a marginal extent. Notably, student groups with historically higher dropout rates are slightly compensated by being ranked lower in predicted dropout risks when protected attributes are used. This compensating effect, however, does not accumulate to statistically significant changes in predicted labels, possibly because the group differences in dropout rates were not sizeable in the past at the institution we study (see Table 5.4). In other words, protected attributes might have more to contribute to the fairness of prediction in the presence of substantial existing inequalities. Still, the existence of a weak compensating instead of segregating effect justifies the inclusion of these attributes. After all, a major argument for race-aware models, and more generally socio-demographic-aware models, is to capture structural inequalities in society that disproportionately expose members of minoritized groups to more adverse conditions. In addition, the deliberate exclusion of protected attributes from dropout prediction models can be construed as subscribing to a "colorblind" ideology, which has been criticized as a racist approach that serves to maintain the status quo (Burke, 2018).

Another contribution of this work lies in our approach to fairness evaluation. The analyses and visualizations we present are the result of many iterations to arrive at simple yet compelling ways to communicate fairness at different levels of aggregation and across many protected attributes. These methods can be used by those who seek to evaluate model fairness for research and practice. Prior research has mostly focused on evaluating one protected attribute at a time, but in most real-world applications we care about more than one protected attribute. We recommend comparing AWARE against BLIND models in terms of the individual ranking differences by group (Figure 5.3) as well as the group difference plots for multiple performance metrics and protected attributes (Figure 5.2). This approach offers a sensitive instrument for diagnosing fairness-related issues in various domains of application, which could easily be implemented in a fairness dashboard that evaluates multiple protected attributes, models, and performance metrics (Williamson and Kizilcec, 2021). This will remain a promising line of our future work.

This research has broader implications for using predictive analytics in higher education beyond its contributions to algorithmic fairness. With a common set of institutional features, we achieve 76% prediction accuracy and 67% recall on unseen students in online settings, that is, correctly identifying 67% of actual dropouts with their first-term records. For residential students, we achieve a higher accuracy of 84% but a lower recall of 54%. These performance metrics may seem somewhat lower than in prior studies of dropout prediction, but this might be because most existing studies examine a smaller sample of more homogeneous students, such as students in the same cohort or program (Del Bonifro et al., 2020; Dekker et al., 2009). This highlights the general challenge of predicting college dropout accurately. As suggested by the large variance in predicted probabilities for dropouts (Figure 5.1), widely used institutional features might not perform well in capturing common signals of dropout. This may point to important contextual factors that our institutional practices are presently overlooking. We view this as a limitation and important next step that will require both an interrogation of the theoretical basis for predictors and close collaboration with practitioners.

Further directions for future research in this area include exploring counterfactual notions of fairness in this context by testing how predictions would differ for counterfactual protected attributes, all else being equal. This would benefit the contemporary education system which relies increasingly on research that provides causal evidence. We would also like to move from auditing to problem-solving by evaluating correction methods for any pre-existing unfairness in predictions to see how the AWARE relative to the BLIND model responds (Lee and Kizilcec, 2020). We hope that this study inspires more researchers in the learning analytics and educational data mining communities to engage with issues of algorithmic bias and fairness in the models and systems they develop and evaluate.

Chapter 6

Concluding Remarks

This dissertation highlights the importance of equity themes in the emerging research field of educational data science (EDS). Four empirical studies in higher education contexts are presented to exemplify equity-oriented EDS research at the micro- or macro-level and with an explanatory or predictive paradigm. While these studies are topically different, they together convey a few important messages.

First, novel "big data" tracks educational processes at a much more granular level than common educational data in previous research and therefore can unveil the dynamic development of educational inequality typically measured in terms of static outcomes. The richness of such data also enables advanced statistical and computational methods to address challenges to causality in observational research contexts. For example, the availability of detailed discussion logs in Chapter 2 makes it possible to understand online peer effects at the individual post and response level and to construct low-level instrumental variables for causal inference, instead of treating interaction processes as a black box as in most previous research on peer effects.

Second, predictive analytics can facilitate different levels of educational decision making, and combining traditional data sources and novel “big data” in these algorithms might contribute to more reliable and equitable predictions. Importantly, unlike in some other application areas where highly complicated computational models can boost predictive performance by a great extent, educational predictive analytics rely much more on the choice of predictors which would lead to more desirable performance if based on solid explanatory education research. This is true for social science research in general (Salganik et al., 2020) and highlights the importance of theoretical advances in the age of “big data”.

In any case, equity-oriented EDS research is still an open space and more intellectual and practical efforts need to be made to define the agenda. Below are a few specific directions.

The first direction is the interplay between theory development and data analytics. As mentioned above, researchers need to filter massive data with existing knowledge to build responsible and cost-effective data science models to help act against existing educational inequities. On the other hand, computational techniques can dig out prominent patterns from massive data, which may complement existing theories and advance the understanding of educational processes. For example, data-mined behavioral sequences may strongly predict learning outcomes but are not well mapped to established theoretical constructs which tend to capture higher level processes than the action-level signals in the behavioral trace data. This then calls for the creation of finer-grained constructs to explain the mechanism of learning.

The second direction is unifying data infrastructure. Just as commonly used administrative or standardized testing data, some of the “big data” in education (e.g., behavioral logs) are universal and standard across schools and institutions, so the nuanced understanding of educational inequities or the predictive data analytics that help address them can potentially revolutionize the landscape of education research and practice. However, this potential can only be realized when the education system shares infrastructure and policies that facilitate the management and analytics of such non-traditional data, which are still yet to be constructed with more communal effort across stakeholders and institutions.

The third direction is holistic investigation of fairness, accountability, transparency and ethics of educational algorithms. This requires researchers to situate algorithms in the entire lifecycle of educational applications, including data collection, model development, deployment for decision making, etc. A mixture of methodologies should be leverage to connect these stages and build a thorough understanding of these algorithms in the wild. Example topics include understanding how measurement and representational biases in the input data translate into biased predictions, and how decisionmakers' cognitive bias and algorithmic bias interact. In addition, the effort to improve educational algorithms needs to move beyond borrowing from the generic computing community and integrate insights from education research to develop appropriate solutions.

Bibliography

Allen-Ramdial, S.-A. A. and Campbell, A. G. (2014). Reimagining the pipeline: Advancing stem diversity, persistence, and success. *BioScience*, 64(7):612–618.

Anderson, T. W. and Hsiao, C. (1981). Estimation of dynamic models with error components. *Journal of the American statistical Association*, 76(375):598–606.

Arnold, K. E. and Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pages 267–270.

Arthurs, N. and Alvero, A. (2020). Whose Truth is the "Ground Truth"? College Admissions Essays and Bias in Word Vector Evaluation Methods. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*, pages 342–349.

Aucejo, E. M., French, J., Araya, M. P. U., and Zafar, B. (2020). The impact of covid-19 on student experiences and expectations: Evidence from a survey. *Journal of Public Economics*, 191:104271.

Aulck, L., Nambi, D., Velagapudi, N., Blumenstock, J., and West, J. (2019). Mining University Registrar Records to Predict First-Year Undergraduate Attrition. In *The 12th International Conference on Educational Data Mining (EDM)*, pages 9–18, Montréal, Canada.

Bacher-Hicks, A., Goodman, J., and Mulhern, C. (2021). Inequality in household adaptation to schooling shocks: Covid-induced online learning engagement in real time. *Journal of Public Economics*, 193:104345.

Bailey, T., Jeong, D. W., and Cho, S.-W. (2010). Referral, enrollment, and completion in developmental education sequences in community colleges. *Economics of Education Review*, 29(2):255 – 270. Special Issue in Honor of Henry M. Levin.

Baker, R., Xu, D., Park, J., Yu, R., Li, Q., Cung, B., Fischer, C., Rodriguez, F., Warschauer, M., and 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*, 17:1–24.

Balaji, M. S. and Chakrabarti, D. (2010). Student Interactions in Online Discussion Forum: Empirical Research from 'Media Richness Theory' Perspective. *Journal of Interactive Online Learning*, 9(1).

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning*.

Beattie, G., Laliberté, J.-W. P., Michaud-Leclerc, C., and Oreopoulos, P. (2019). What sets college thrivers and divers apart? A contrast in study habits, attitudes, and mental health. *Economics Letters*, 178:50–53.

Beattie, G., Laliberté, J.-W. P., and Oreopoulos, P. (2018). Thrivers and divers: Using non-academic measures to predict college success and failure. *Economics of Education Review*, 62:170–182.

Beaulac, C. and Rosenthal, J. S. (2019). Predicting University Students' Academic Success and Major Using Random Forests. *Research in Higher Education*, 60(7):1048–1064.

Berens, J., Schneider, K., Görtz, S., Oster, S., and Burghoff, J. (2019). Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. *Journal of Educational Data Mining*, 11(3):1–41.

Bettinger, E., Liu, J., and Loeb, S. (2016). Connections Matter: How Interactive Peers Affect Students in Online College Courses. *Journal of Policy Analysis and Management*, 35(4):932–954.

Bettinger, E. P., Evans, B. J., and Pope, D. G. (2013). Improving college performance and retention the easy way: Unpacking the ACT exam. *American Economic Journal: Economic Policy*, 5(2):26–52.

Bettinger, E. P., Fox, L., Loeb, S., and Taylor, E. S. (2017). Virtual Classrooms: How Online College Courses Affect Student Success. *American Economic Review*, 107(9):2855–2875.

Blanden, J. (2020). Education and inequality. In Bradley, S. and Green, C., editors, *The Economics of Education*, pages 119–131. Academic Press, 2 edition.

Broadbent, J. and Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *Internet and Higher Education*, 27:1–13.

Burke, M. (2018). *Colorblind racism*. John Wiley & Sons.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems 30*, pages 3992–4001.

Cao, W., Fang, Z., Hou, G., Han, M., Xu, X., Dong, J., and Zheng, J. (2020). The psychological impact of the covid-19 epidemic on college students in china. *Psychiatry Research*, 287:112934.

Cataldi, E. F., Bennett, C. T., and Chen, X. (2018). First-generation students: College access, persistence, and postbachelor's outcomes (nces 2018421). Technical report, National Center for Education Statistics.

Chen, X. and Nunnery, A. (2019). Profile of very low and low-income undergraduates in 2015–16. Technical report, National Center for Education Statistics.

Choi, B. C. and Pak, A. W. (2005). Peer reviewed: a catalog of biases in questionnaires. *Preventing chronic disease*, 2(1).

Choi, S. P. M., Lam, S., Li, K. C., and Wong, B. T. M. (2018). Learning analytics at low cost: At-risk student prediction with clicker data and systematic proactive interventions. *Journal of Educational Technology & Society*, 21(2):273–290.

Cicchinelli, A., Veas, E., Pardo, A., Pammer-Schindler, V., Fessl, A., Barreiros, C., and Lindstädt, S. (2018). Finding traces of self-regulated learning in activity streams. In *Proceedings of the 8th International Conference on Learning Analytics Knowledge (LAK '18)*, pages 191–200. ACM.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, 94:S95–S120.

Conijn, R., Snijders, C., Kleingeld, A., and Matzat, U. (2017). Predicting student performance from lms data: A comparison of 17 blended courses using moodle lms. *IEEE Transactions on Learning Technologies*, 10:17–29.

Dawson, S., Jovanovic, J., Gašević, D., and Pardo, A. (2017). From prediction to impact: Evaluation of a learning analytics retention program. In *Proceedings of the 7th International Conference on Learning Analytics & Knowledge*, page 474–478.

de Brey, C., Musu, L., McFarland, J., Wilkinson-Flicker, S., Diliberti, M., Zhang, A., Branstetter, C., and Wang, X. (2019). Status and trends in the education of racial and ethnic groups 2018 (nces 2019-038). Technical report, National Center for Education Statistics.

Dekker, G. W., Pechenizkiy, M., and Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. In *Proceedings of the 2nd International Conference on Educational Data Mining (EDM 2009)*, pages 41–50.

Del Bonifro, F., Gabbrielli, M., Lisanti, G., and Zingaro, S. P. (2020). Student Dropout Prediction. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education (AIED 2020)*, pages 129–140. Springer.

DiBenedetto, M. K. and Bembenutty, H. (2013). Within the pipeline: Self-regulated learning, self-efficacy, and socialization among college students in science courses. *Learning and Individual Differences*, 23:218–224.

Doroudi, S. and Brunskill, E. (2019). Fairer but not fair enough on the equitability of knowledge tracing. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK '19)*, pages 335–339. ACM.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, page 214–226.

Ekowo, M. and Palmer, I. (2016). The Promise and Peril of Predictive Analytics in Higher Education: A Landscape Analysis. Technical report, New America.

Farrow, E., Moore, J., and Gašević, D. (2019). Analysing discussion forum data: a replication study avoiding data contamination. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 170–179.

Fischer, C., Pardos, Z., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., and Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education*, 44:130–160.

Forteza, D., Harfield, T., Whitmer, J., and Dietrichson, A. (2017). What does it take to predict student risk? Evaluating LMS data to determine readiness for predictive modeling. Technical report, Blackboard Inc.

Gardner, J., Brooks, C., and Baker, R. (2019). Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19*, pages 225–234, Tempe, AZ, USA. ACM Press.

Garrison, D. R. and Arbaugh, J. (2007). Researching the community of inquiry framework: Review, issues, and future directions. *The Internet and Higher Education*, 10(3):157–172.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.

Haveman, R. and Smeeding, T. (2006). The role of higher education in social mobility. *The Future of Children*, 16(2):125–150.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., and Liao, S. N. (2018). Predicting academic performance: A systematic literature review. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018 Companion)*, pages 175–199. Association for Computing Machinery.

Ho, C.-H. and Swan, K. (2007). Evaluating online conversation in an asynchronous learning environment: An application of Grice's cooperative principle. *The Internet and Higher Education*, 10(1):3–14.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., and Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, 595:181–188.

Hu, Q. and Rangwala, H. (2020). Towards Fair Educational Data Mining: A Case Study on Detecting At-risk Students. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*, pages 431–437.

Huckins, J. F., DaSilva, A. W., Wang, W., Hedlund, E., Rogers, C., Nepal, S. K., Wu, J., Obuchi, M., Murphy, E. I., Meyer, M. L., Wagner, D. D., Holtzheimer, P. E., and Campbell, A. T. (2020). Mental health and behavior of college students during the early phases of the covid-19 pandemic: Longitudinal smartphone and ecological momentary assessment study. *Journal of Medical Internet Research*, 22:e20185.

Hussar, B., Zhang, J., Hein, S., Wang, K., Roberts, A., Cui, J., Smith, M., Mann, F. B., Barmer, A., Dilig, R., and Nachazel (2020). The condition of education 2020 (nces 2020-144). Technical report, National Center for Education Statistics.

Hutt, S., Gardner, M., Duckworth, A. L., and D'Mello, S. K. (2019). Evaluating Fairness and Generalizability in Models Predicting On-Time Graduation from College Applications. In *The 12th International Conference on Educational Data Mining (EDM)*, pages 79–88, Montréal, Canada.

Jaggars, S. S. and Xu, D. (2016). How do online course design features influence student performance? *Computers & Education*, 95:270–284.

Jarke, J. and Breiter, A. (2019). Editorial: the datafication of education. *Learning, Media and Technology*, 44(1):1–6.

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

Jayaprakash, S. M., Scott, J. M., and Kerschen, P. (2017). Connectivist Learning Using SuiteC - Create, Connect, Collaborate, Compete! In *Practitioner Track Proceedings of the 7th International Learning Analytics & Knowledge Conference (LAK17)*, pages 69–76, Vancouver, BC, Canada.

Kent, C., Laslo, E., and Rafaeli, S. (2016). Interactivity in online discussions and learning outcomes. *Computers & Education*, 97:116–128.

Kizilcec, R. F. and Lee, H. (2020). Algorithmic fairness in education. *arXiv preprint arXiv:2007.05443*.

Kizilcec, R. F., Makridis, C. A., and Sadowski, K. C. (2021). Pandemic response policies' democratizing effects on online learning. *Proceedings of the National Academy of Sciences*, 118.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. (2018). Algorithmic Fairness. *AEA Papers and Proceedings*, 108:22–27.

Kreijns, K., Kirschner, P. A., and Vermeulen, M. (2013). Social Aspects of CSCL Environments: A Research Framework. *Educational Psychologist*, 48(4):229–242.

Kuh, G. D., Kinzie, J., Buckley, J. A., Bridges, B. K., and Hayek, J. C. (2007). Piecing Together the Student success puzzle: Research, Propositions, and Recommendations. *ASHE Higher Education Report*, 32(5):1–182.

Kung, C. and Yu, R. (2020). Interpretable Models Do Not Compromise Accuracy or Fairness in Predicting College Success. In *Proceedings of the 7th ACM Conference on Learning @ Scale (L@S '20)*, pages 413–416, New York, NY, USA. Association for Computing Machinery (ACM).

Lee, H. and Kizilcec, R. F. (2020). Evaluation of fairness trade-offs in predicting student success. *arXiv preprint arXiv:2007.00088*.

Li, Q., Baker, R., and Warschauer, M. (2020). Using clickstream data to measure, understand, and support self-regulated learning in online courses. *The Internet and Higher Education*, page 100727.

Loukina, A., Madnani, N., and Zechner, K. (2019). The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Luo, Y. and Pardos, Z. A. (2018). Diagnosing University Student Subject Proficiency and Predicting Degree Completion in Vector Space. In *Proceedings of the Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, New Orleans, LA, USA.

Matcha, W., Gašević, D., Uzir, N. A., Jovanović, J., and Pardo, A. (2019). Analytics of Learning Strategies: Associations with Academic Performance and Feedback. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19*, pages 461–470, Tempe, AZ, USA. ACM Press.

McCormick, A. C., Kinzie, J., and Gonyea, R. M. (2013). Student engagement: Bridging research and practice to improve the quality of undergraduate education. In Paulsen, M. B., editor, *Higher Education: Handbook of Theory and Research*, pages 47–92. Springer, Dordrecht.

McFarland, D. A., Khanna, S., Domingue, B. W., and Pardos, Z. A. (2021). Education data science: Past, present, future. *AERA Open*, 7:233285842110520.

Means, B. and Neisler, J. (2020). Suddenly online: A national survey of undergraduates during the covid-19 pandemic. Technical report, Digital Promise.

Miller, A. L. (2012). Investigating social desirability bias in student self-report surveys. *Educational Research Quarterly*, 36:30–47.

Motz, B., Quick, J., Schroeder, N., Zook, J., and Gunkel, M. (2019). The validity and utility of activity logs as a measure of student engagement. In *Proceedings of the 9th International Conference on Learning Analytics Knowledge (LAK '19)*, pages 300–309. Association for Computing Machinery.

Nguyen, H., Wu, L., Fischer, C., Washington, G., and Warschauer, M. (2020). Increasing success in college: Examining the impact of a project-based introductory engineering course. *Journal of Engineering Education*.

Osterhage, J. L., Usher, E. L., Douin, T. A., and Bailey, W. M. (2019). Opportunities for self-evaluation increase student calibration in an introductory biology course. *CBE—Life Sciences Education*, 18(2):ar16.

Pacansky-Brock, M., Smedshammer, M., and Vincent-Layton, K. (2020). Humanizing online teaching to equitize higher education. *Current Issues in Education*, 21(2).

Pantages, T. J. and Creedon, C. F. (1978). Studies of college attrition: 1950—1975. *Review of Educational Research*, 48(1):49–101.

Paquette, L., Li, Z., Baker, R., Ocumpaugh, J., and Andres, A. (2020). Who's learning? Using demographics in EDM research. *Journal of Educational Data Mining*, 12(3):1–30.

Pardos, Z. A., Fan, Z., and Jiang, W. (2019). Connectionist recommendation in the wild: on the utility and scrutability of neural networks for personalized course guidance. *User Modeling and User-Adapted Interaction*, pages 1–39.

Park, J., Yu, R., Rodriguez, F., Baker, R., Smyth, P., and Warschauer, M. (2018). Understanding Student Procrastination via Mixture Models. In *Proceedings of the 11th International Conference on Educational Data Mining (EDM)*, Buffalo, NY, United States.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Picciano, A. G. (2002). Beyond Student Perceptions: Issues of Interaction, Presence, and Performenace in an Online Course. *Journal of Asynchronous Learning Networks*, 6(1):21–40.

Pintrich, P. R. and De Groot, E. V. (1990). Motivational and Self-Regulated Learning Components of Classroom Academic Performance. *Journal of Educational Psychology*, 82(1):33–40.

Pintrich, P. R., Smith, D. A. F., Garcia, T., and McKeachie, W. J. (1991). A manual for the use of the motivated strategies for learning questionnaire (mslq). Technical report, Ann Arbor, MI.

Ragusa, A. T. and Crampton, A. (2018). Sense of connection, identity and academic success in distance education: sociologically exploring online learning environments. *Rural Society*, 27(2):125–142.

Richardson, J. C., Maeda, Y., Lv, J., and Caskurlu, S. (2017). Social presence in relation to students' satisfaction and learning in the online environment: A meta-analysis. *Computers in Human Behavior*, 71:402–417.

Rodriguez-Planas, N. (2022). Covid-19, college academic performance, and the flexible grading policy: A longitudinal analysis. *Journal of Public Economics*, 207:104606.

Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., Morgan, A. C., Pentland, A., Polimis, K., Raes, L., Rigobon, D. E., Roberts, C. V., Stanescu, D. M., Suhara, Y., Usmani, A., Wang, E. H., Adem, M., Alhajri, A., AlShebli, B., Amin, R., Amos, R. B., Argyle, L. P., Baer-Bositis, L., Büchi, M., Chung, B.-R., Eggert, W., Faletto, G., Fan, Z., Freese, J., Gadgil, T., Gagné, J., Gao, Y., Halpern-Manners, A., Hashim, S. P., Hausen, S., He, G., Higuera, K., Hogan, B., Horwitz, I. M., Hummel, L. M., Jain, N., Jin, K., Jurgens, D., Kaminski, P., Karapetyan, A., Kim, E. H., Leizman, B., Liu, N., Möser, M., Mack, A. E., Mahajan, M., Mandell, N., Marahrens, H., Mercado-Garcia, D., Mocz, V., Mueller-Gastell, K., Musse, A., Niu, Q., Nowak, W., Omidvar, H., Or, A., Ouyang, K., Pinto, K. M., Porter, E., Porter, K. E., Qian, C., Rauf, T., Sargsyan, A., Schaffner, T., Schnabel, L., Schonfeld, B., Sender, B., Tang, J. D., Tsurkov, E., van Loon, A., Varol, O., Wang, X., Wang, Z., Wang, J., Wang, F., Weissman, S., Whitaker, K., Wolters, M. K., Woon, W. L., Wu, J., Wu, C., Yang, K., Yin, J., Zhao, B., Zhu, C., Brooks-Gunn, J., Engelhardt, B. E., Hardt, M., Knox, D., Levy, K., Narayanan, A., Stewart, B. M., Watts, D. J., and McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403.

Schippers, M. C., Scheepers, A. W., and Peterson, J. B. (2015). A scalable goal-setting intervention closes both the gender and ethnic minority achievement gap. *Palgrave Communications*, 1(1):1–12.

Scott-Clayton, J. (2012). Do High-Stakes Placement Exams Predict College Success?

Seaman, J. E., Allen, I. E., and Seaman, J. (2018). Grade Increase: Tracking Distance Education in the United States. Technical report.

Selwyn, N. and Gašević, D. (2020). The datafication of higher education: discussing the promises and problems. *Teaching in Higher Education*, 25(4):527–540.

Shapiro, D., Dundar, A., Huie, F., Wakhungu, P., Bhimdiwala, A., and Wilson, S. (2019). Completing College: A State-Level View of Student Completion Rates (Signature Report No. 16a). Technical report, National Student Clearinghouse Research Center, Herndon, VA.

Shum, S. B. (2020). Should predictive models of student outcome be "colour-blind"? `http://simon.buckinghamshum.net/2020/07/should-predictive-models-of-student-outcome-be-colour-blind`.

Siemens, G. (2005). Connectivism : A Learning Theory for the Digital Age. *International Journal of Instructional Technology and Distance Learning*, 2(1):1–7.

Snyder, T. D., de Brey, C., and Dillow, S. A. (2019). Digest of education statistics 2018 (nces 2020-009). Technical report, National Center for Education Statistics.

Stahl, G., Koschmann, T., and Suthers, D. (2014). Computer-Supported Collaborative Learning. In *The Cambridge Handbook of the Learning Sciences*, pages 479–500.

The Chronicle of Higher Education (2020). The post-pandemic college. Technical report.

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement*, 8(2):63–70.

Tu, C.-H. and McIsaac, M. (2002). The relationship of social presence and interaction in online classes. *The American journal of distance education*, 16(3):131–150.

U.S. Department of Education Office for Civil Rights (2021). Education in a pandemic: The disparate impacts of covid-19 on america's students. Technical report.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE.

von Keyserlingk, L., Yamaguchi-Pedroza, K., Arum, R., and Eccles, J. S. (2021). Stress of university students before and after campus closure in response to covid-19. *Journal of Community Psychology*.

Vygotsky, L. S. (1978). *Interaction between Learning and Development*. Harvard University Press, Cambridge, MA, USA.

Wang, S. and Noe, R. A. (2010). Knowledge sharing: A review and directions for future research. *Human Resource Management Review*, 20(2):115–131.

Whitmer, J., Pedro, S. S., Liu, R., Walton, K. E., Moore, J. L., and Lotero, A. A. (2019). The Constructs Behind the Clicks. Technical report, ACT, Inc.

Williamson, K. and Kizilcec, R. F. (2021). Learning analytics dashboard research has neglected diversity, equity and inclusion. In *Proceedings of the 8th ACM Conference on Learning @ Scale*.

Wise, A. F. and Cui, Y. (2018). Unpacking the relationship between discussion forum participation and learning in MOOCs. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge - LAK '18*, pages 330–339, New York, New York, USA. ACM Press.

Wolff, A., Zdrahal, Z., Nikolov, A., and Pantucek, M. (2013). Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 145–149.

Wolters, C. A. (1998). Self-regulated learning and college students' regulation of motivation. *Journal of educational psychology*, 90(2):224.

Xie, J., Essa, A., Mojarad, S., Baker, R. S., Shubeck, K., and Hu, X. (2017). Student Learning Strategies and Behaviors to Predict Success in an Online Adaptive Mathematics Tutoring System. In *Proceedings of the 10th International Conference on Educational Data Mining*, pages 460–465, Wuhan, China.

Xu, D. and Jaggars, S. S. (2013). The impact of online learning on students' course outcomes: Evidence from a large community and technical college system. *Economics of Education Review*, 37:46–57.

Xu, D. and Jaggars, S. S. (2014). Performance Gaps Between Online and Face-to-Face Courses: Differences Across Types of Students and Academic Subject Areas. *The Journal of Higher Education*, 85(5):633–659.

Xu, D. and Xu, Y. (2020). The ambivalence about distance learning in higher education. In *Higher Education: Handbook of Theory and Research*, pages 1–52. Springer, Cham.

Yu, R., Lee, H., and Kizilcec, R. F. (2021). Should college dropout prediction models include protected attributes? In *Proceedings of the 8th ACM Conference on Learning @ Scale (L@S '21)*, pages 91–100, New York, NY, USA. ACM.

Yu, R., Li, Q., Fischer, C., Doroudi, S., and 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 (EDM 2020)*, pages 292–301.