

# UC Irvine

## UC Irvine Electronic Theses and Dissertations

### Title

Predicting Learning Behavior Using A Unified Framework: Looking Beyond the Clicks

### Permalink

<https://escholarship.org/uc/item/6j3333vd>

### Author

Mohammed, Shafee

### Publication Date

2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,  
IRVINE

Predicting Learning Behavior Using A Unified Framework: Looking Beyond the Clicks

DISSERTATION

submitted in partial satisfaction of the requirements  
for the degree of

DOCTOR OF PHILOSOPHY

in Education

by

Shafee Mohammed

Dissertation Committee:  
Associate Professor Susanne M Jaeggi, Chair  
Professor Mark Steyvers  
Assistant Professor Shayan Doroudi

2020



## DEDICATION

To

my family, friends, and colleagues

in recognition of their insurmountable support through thick and thin  
days and nights  
for never leaving my side

for constant words of encouragement  
and  
push for tenacity

A special feeling of gratitude to my supportive parents,  
Razia Begum and Abdul Rahaman Mohammed  
For all they have done for me and my siblings,  
Rafee and Reshma Mohammed,  
And teaching us to prioritize compassion and introspection  
Above everything else.

A special thanks to Snigdha Kamarsu, for being the best(est) friend  
And for making this journey enjoyable and feel less strenuous.

A dedication to you, the reader of this dissertation,  
hoping to complete your own missions. You got this!

“If you can't fly then run, if you can't run then walk, if you can't walk then crawl, but  
whatever you do you have to keep moving forward.”

Martin Luther King, Jr.

## TABLE OF CONTENTS

	Page
LIST OF FIGURES	iv
LIST OF TABLES	v
ACKNOWLEDGMENTS	vi
CURRICULUM VITAE	vii
ABSTRACT OF THE DISSERTATION	xii
CHAPTER 1: INTRODUCTION	1
CHAPTER 2: Predicting learning in the context of working memory training	54
CHAPTER 3: From Intentions to Actions: Understanding Students' Self-Reported Study Plans, Adherence to Study Plans, Click Behaviors, and Their Relationship to Learning	95
CHAPTER 4: Dynamic of Motivations, Priorities, and Student Learning: Looking Beyond Click Behavior to Predict Student Learning in An Online Course	124
CHAPTER 5: General Discussion	198
BIBLIOGRAPHY	211
APPENDIX A: Table of Survey Measures	227

## LIST OF FIGURES

	Page	
Figure 1.1	Features of learning	15
Figure 2.1	Data Splitting Protocol	75
Figure 2.2	Prediction accuracy with Logistic Regression	84
Figure 2.3	Adjusted R-squared	85
Figure 3.1	Grade point distributions of the students	106
Figure 3.2	Final Letter Grade Distribution of Students	109
Figure 3.3	Results from Regression Models for Study-2 RQ1	112
Figure 3.4	Results from Classification Models for Study-2 RQ1	113
Figure 3.5	Results from Regression Models for RQ2	115
Figure 3.6	Results from Classification Models for RQ2	116
Figure 4.1	Timeline of Data Collection	148
Figure 4.2	Word Clouds	163
Figure 4.3	LDA generated Word Clouds	164
Figure 4.4	Final Study Grade Distributions for Study-3	165
Figure 4.5	Weekly Review Quiz Score Distributions	166
Figure 4.6	Results from Regression Models for RQ1	176
Figure 4.7	Results from Classification Models for RQ1	178
Figure 4.8	Results from Regression Models for RQ2	180
Figure 4.9	Results from Classification Models for RQ 2	181
Figure 4.10	Results from Regression Models for RQ3	182
Figure 4.11	Results from Classification Models for RQ3	183
Figure 5.1	Screenshot of “New Analytics”	207

## LIST OF TABLES

		Page
Table 1.1	An overview of the three studies	50
Table 2.1	List of studies, sample and training details for study-1	65
Table 2.2	A description of sample population, demographics, and training	68
Table 2.3	Confusion Matrix	77
Table 2.4	Models Tested and List of Features for Study-1	83
Table 3.1	Models Tested and List of Features for Study-2	104
Table 3.2	Descriptive Statistics of students	107
Table 3.3	Descriptive Statistics of exams	109
Table 3.4	Feature Importance	117
Table 4.1	An Adaptation of MACM	145
Table 4.2	A List of Sample Questions	150
Table 4.3	Descriptive Statistics of quizzes	166
Table 4.4	Models Tested and List of Features for Study-3	173
Table 4.5	Feature Importance	185

## ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my committee chair, Associate Professor Susanne M. Jaeggi, who has the mindset and the substance of a genius: she continually shows tenacity, spirit, and confidence through her research. I am immensely grateful for her scientific rigor and expertise, her unlimited and unconditional support, guidance, and encouragement to pursue my intellectual endeavors. Susanne has been the cornerstone of this work throughout its journey, from inception to completion. She provided feedback, advice, and challenged my thinking when I am off-track and helped me learn, improve, and advance. Thank you, Susanne, for everything.

I am grateful to all of my collaborators, including Martin Buschkuehl, Mark Steyvers, Shayan Doroudi, Ksenia Korobkova, and Fernando Rodriguez for their advice, help, support, and letting me pick on their brain for intellectual stimulation. I would like to thank my fellow graduate students for collaborating, sharing data, and helping me design and develop my data analysis plans. Special note of thanks to Peter McPartlan, Osman Umarji, and Masha Jones for supporting me every step of the way directly or indirectly. I am very thankful for my research assistants that worked with me over the years and helped me shape my work. Henry, Caleb, Dylan, and Fanbo worked closely with me these past few years while continually helping me, challenging me, and collaborating with me. I cannot thank you all enough for everything for the endless hours of effort you poured into our work. Finally, I could not thank the graduate support groups at UC Irvine, specifically, graduate school administrative staff at SoE, the librarians, the graduate writing division for always delivering on their services with a smile. Your work ethics and dedication inspire me. I promise to emulate you and pay your kindness forward.

I am thankful to every graduate, undergraduate, manager, and support staff that has worked with me and the rest of my lab group these past few years. Every one of you made my time away from my homeland and my family less miserable and more tolerable. Very special thanks to Snigdha Kamarsu for being my endless supply of positive energy, hope, and tenacity. Thank you, again, Snigdha. This would have been difficult without your support.

In the beginning and in the end, it was the continued and steadfast support of my family that helped me endure difficult times, gave me a reason to keep pushing myself, assured me that the 11000 mile distance is but a number, and that we are always in it together. It would have been extremely difficult for me to take even a single step forward without you. I am forever thankful and eternally grateful. Oh and...finally, I would like to thank every single person who was involved in the process of helping me with the endless consumption of caffeine needed to keep my brain wired. I could not have done this without you.



## CURRICULUM VITA

### Shafee Mohammed

- 2011 B.S. Optometry, Birla Institute of Technology and Sciences, Pilani, India.
- 2013-14 Teaching Assistant, School of Optometry, University of Houston
- 2015-20 Teaching Assistant, Graduate School of Education, University of California, Irvin
- 2016 fMRI Training Program, University of Michigan
- 2016-17 Graduate Student Researcher, WMP Lab, University of California, Irvine
- 2018 IBM AI Data Scientist
- 2019 M.A. in Education, University of California, Irvine
- 2020 Ph.D. in Education, University of California, Irvine

## FIELDS OF STUDY AND RESEARCH FOCUS

Human learning and behavior and Learning Analytics

Research interests include individual differences in learning, cognition, executive functions, the mediators and moderators of learning and predicting learning quality using predictive analytics.

## PUBLICATIONS

Mohammed, S., Flores, L., Deveau, J., Hoffing, R. C., Phung, C., Parlett, C. M., Sheehan, E., Lee, D., Au, J., Buschkuehl, M., Zordan, V., Jaeggi, S. M., & Seitz, A. R. (2017). The Benefits and Challenges of Implementing Motivational Features to Boost Cognitive Training Outcome. *Journal of Cognitive Enhancement*, 1(4), 491–507. <https://doi.org/10.1007/s41465-017-0047-y>

**In preparation:**

Mohammed, S., Katz, B., Parlett, C. M., Pahor, A., Buschkuehl, M., Seitz, A. R., Shah, P., Steyvers, M., Jonides, J., & Jaeggi, S. M. Predictive modeling of changes in a working memory training performance.

Mohammed, S., Mcpartlan, P., Umarji, O., Rutherford, T., Hanwen, Y., Pitts, C., Rodriguez, F., Warschauer, M., From Intentions to Actions: Understanding student's adherence to study plans and their relation to learning outcomes.

Mohammed, S., Mcpartlan, P., Umarji, O., Rutherford, T., Hanwen, Y., Pitts, C., Rodriguez, F., Warschauer, M., Dynamics of Motivations, Priorities, and Student Learning: Looking Beyond Clicks

### **Conference Presentations:**

Mohammed, S., Katz, B., Jaeggi, S. M., & Buschkuehl, M. (2016, May 6th). *Individual Differences in Working Memory* – Presented at the Fostering Literacy and Learning with Text and Data Mining, UC Irvine, CA.

Mohammed, S., Katz, B., Jaeggi, S. M., & Buschkuehl, M. (2016, May 20th). *Individual Differences in Working Memory* – Presented at the The 29th Annual Conference on the Neurobiology of Learning and Memory, UC Irvine, CA.

Mohammed, S., Katz, B., Jaeggi, S. M., & Buschkuehl, M. (2016, June 28th). *Getting at Individual Differences in Working Memory: A Data Mining Approach* – Presented at the Games for Engaging Learning Annual Scientific meeting, Boston University, MA.

Mohammed, S., Katz, B., Jaeggi, S. M., & Buschkuehl, M. (2016, Nov 19th). *Individual Differences in Working Memory Training* – Poster presented at the 57th Annual Meeting of the Psychonomic Society, Boston, MA.

Parlett, M. C., Mohammed, S., Katz, B., Shah, P., Zabel, C., Au, Jones, M., Buschkuehl, M., & Jaeggi, S. M., (2017, Jan 29th). *Individual Differences in Working Memory* – Presented at the Dallas Aging and Cognition Conference, Dallas, TX.

Mohammed, S., & Jaeggi, S. M. (2017, Apr 21st). *Can Machine Learning Help Predict Working Memory Training Gains?* – presented at the AGS Symposium, UC Irvine, Irvine, CA.

Mohammed, S., Katz, B., Buschkuehl, M., Shah, P., Steyvers, M., Jonides, J., & Jaeggi, S. M. (2017, July 27th) Predicting Individual Differences in Working Memory

Training Gain. Virtually presented (due to Visa issues) at the CogSci 2017, London, UK.

Mohammed, S., Flores, L., Deveau, J., Hoffing, R. C., Phung, C., Parlett, C. M., ... & Zordan, V. (2017, July 28). The benefits and challenges of implementing motivational features to boost cognitive training outcome. Presented at 2017 Games for Change Festival, New York, NY.

Mohammed, S., Flores, L., Deveau, J., Hoffing, R. C., Phung, C., Parlett, C. M., ... & Zordan, V. (2017, Nov, 10). The benefits and challenges of implementing motivational features to boost cognitive training outcome. Virtually presented at 58th Annual Meeting of the Psychonomic Society, Vancouver, Canada.

Mohammed, S., (2017, Oct 5) *Can Machine Learning Help Predict Human Learning?* - Presented as part of a panel titled "Media and Technology for Cognitive Enhancement Throughout the Lifespan."

Mcpartlan, P., Rutherford, T., Mohammed, S., Rodriguez, F., (2019, Apr 5th). *Modality Motivation: Selection Effects and Motivational Differences in Students Who Choose to Take Courses Online.* – Presented at the AERA Annual Conference, Toronto, Canada.

#### FELLOWSHIPS

2018 - 19	Pytorch Scholarship, Udacity
2018	Data Science Certificate, UCI
2016-2016	fMRI Training Program, University of Michigan
2013-2014	Presidential Fellowship, University of Houston
2013-2014	Graduate Student Fellowship
2011-2012	Fellowship in Orthoptics and Vision Therapy

#### HONORS AND AWARDS

2014	Best International Graduate Student Award
2007-2011	World Health Organization Academic Excellence Award
2016	Honorable Mention - Fostering Literacy and Learning with Text and Data Mining, UC Irvine, CA.

#### RESEARCH AND WORK EXPERIENCE

**Graduate Student Researcher** – University of California-Irvine, School of Education, Irvine CA; Working Memory and Plasticity Laboratory: advised by Dr. Susanne Jaeggi (sept 15 – present)

### **Teaching Assistant for courses -**

University of California-Irvine, School of Education, Irvine CA; (Sept 15 - present)

**Product Design Consultant** – Freelance (Oct-14 – Jul-15)

**Research Assistant** - University of Houston, School of Optometry, Houston Tx; Advanced Anatomy and Histology Lab advised by Dr. Micheal Twa (Jun-14 – Aug-14)

**Teaching Assistant** – University of Houston, School of Optometry, Houston Tx; Optics: Dr. Han Cheng (Sep-13 – May-14)

**Senior Optometrist** – Reliance Vision Express, Hyderabad India; (Dec-12 – Jul-13)

**Junior Optometrist** – Sankara Eye Care Institutions, Bangalore India; (Jul-11 – Aug-12)

**Optometrist (Intern)** – L V Prasad Eye Institute, Hyderabad India; (May-10 – Jul-11)

### **VOLUNTEER WORK AND POSITIONS OF RESPONSIBILITY**

#### **Committee work:**

Associated Graduate Students UCI, Campus Communications Director

Associated Graduate Students UCI, Council Member – Internal and International Committee

Associated Doctoral Students in Education – AGS Representative, School of Education, UCI

Vice-President – Surabhi, Graduate Indian Students Association, UCI

AGS Symposium Society and Urban Planning Session Moderator

AGS Symposium Physics and Chemistry Session Moderator

Mentor – DECADE, UCI

### **TEACHING EXPERIENCE**

“Case Study Analysis” Professional Development Session in course EDUC 151/PSY BEH 192V, Language and Literacy – Spring, 2016

“Effective Poster Design” Professional Development Session in course EDUC 151/PSY BEH 192V,

Language and Literacy – Spring, 2016

Final Exam Review in course EDUC 40, Theories of Development and Learning Applied to Education – winter, 2016

“Educational data mining” in EDUC 212, Literacy and Technology – Fall, 2016.

“Expertise and Transfer” in EDUC 285, Theories of Learning and Cognition – Winter, 2017.

“How to create a good documentary? A Tutorial” in EDUC 124 - Fall, 2017.

“A Workshop Tutorial on Video Editing Softwares” in EDUC 30 - Fall, 2017.

“Bias in Technology: How does it occur and what to do about it” in EDUC 30 – Fall, 2017.

“Does Technology Promote or Bias Against Multiculturalism?” in EDUC 124 – Fall, 2017.

“Is School Enough?” in EDUC 30 – Winter, 2017.

“Understanding Bias in Technology: A critical view” in EDUC 130 – Fall, 2018.

“An Introduction to Machine Learning” in ED 15 – Winter, 2019.

“Multiculturalism, Bias, Technology, and Inclusion” in ED 124 – Spring, 2019.

“Multiculturalism, Bias, Technology, and Inclusion” in ED 124 – Fall, 2019.

“Navigating grad school applications: Dos and Don'ts” in Ed 124 – Fall, 2019.

“Critical reflections of media and understanding counter-narratives” in Ed 125 – Winter 2020.

“Dimensionality Reduction – 7 critical steps to navigate multidimensional data” in Ed 288B – Spring 2020.

## **ABSTRACT OF THE DISSERTATION**

Predicting Learning Behavior Using A Unified Framework: Looking Beyond the Clicks

By

Shafee Mohammed

Doctor of Philosophy in Education (LTCD)

University of California, Irvine, 2020

Associate Professor Susanne M. Jaeggi, Chair

Predicting learning and human behavior in general is a challenging endeavor. Machine learning driven predictive modeling have been an increasingly popular means to understand disparities in student performance. With more than a handful of approaches to predictive modeling, the current literature of predicting learning is plagued with issues such as lack of standards for predictions, lack of work to understand the lower and upper bounds to context specific predictions, and explanatory models perceived to be at odds with predictive models. To overcome these issues, I use a single predictive modeling framework across three different learning contexts that involve four key steps: using approaches that are prediction task specific while reporting all metrics; using a baseline model for comparison; using context-specific early learning to predict later learning; systematically introducing extrinsic feature sets to derive actionable insights. In my first study, I use this framework to predict learning in a working memory training context. Results suggest that later learning can be predicted from early learning behavior better when using extrinsic features. In the second study, this framework is applied in the context of a blended learning environment. Results

suggest that students' study spacing intentions, demographics, and past achievements predict later learning, while students' click behaviors in the learning environment do not improve predictions. In the third study, this framework is applied in the context of a fully online learning environment. Results show that students' self-reported motivational, affective, and social-emotional data are more predictive of early learning than context-agnostic click-behaviors. Overall, the current work proposes and evaluates a framework that can be used to compare results across learning contexts, between models, and approaches to make predictions that may inform future prescriptions. Specifically, the framework acts as a means to understand the relationship shared by the three key aspects to making successful predictions – 'how well', 'how soon', and 'how much information' – and relevance of looking beyond simple accuracies.

## CHAPTER 1: INTRODUCTION

*“Education is the most powerful weapon which you can use to change the world”*

---- **Nelson Mandela**

For millennia, human learning has been an important factor for societal sustenance and advancement. The two most remarkable advancements in the Pleistocene era that molded human evolution, and arguably, the contemporary human society were evolution of languages and construction of socio-cultural practices. Together, these promoted the acquisition of skills and knowledge from older generations to younger generations (Farina, 2013; Kivinen & Piironen, 2018). Human society emphasizes learning and skill acquisition as means for becoming competent adults. From exploratory and self-directed learning in hunter-gatherer cultures to apprentice learning to the structured curricular training in typical school settings that exist today, the concept of learning as a method for behavioral alteration remains. These have been the differentiating factor of humans from other species throughout the ages since learning and skill acquisition promoted division of labor and coordinated social sustenance through actions.

Before I proceed further, I wish to clarify my positionality on the definition of “learning.” There is no simple or singular answer to the question, “What is learning?”. As Visser (2012) discussed in their work, learning is a complex phenomenon that is tightly intertwined with the notion of living and developing across lifespan. While learning might be a phenomenon which is typically defined with an emphasis on some form of manifest behavioral change that is quantifiable, it goes beyond measurable terms. In the current



work, I will take the narrow delineation of the definition which is limited to changes in ability to perform a task or achieve grades within a course. However, the approach to predicting and understanding learning discussed here may be applicable to any setting or context, as long as there is a means to express that learning occurred (e.g., change in the ways a learner engages with their biological and physical environment, the ways in which a learner asks questions and explores answers to those questions, etc.)

Regardless of the ways in which one defines or measures learning, the cultural practice of transferring knowledge and skills from generation to generation formed the basis of converting an individual's potential into proficiencies (Gray, 2009). Such socio-cultural practices and norms, in turn, act as a cornerstone to 'human behavioral modernity' (Farina, 2013; Kivinen & Piironen, 2018). The society we live in today with near ubiquitous presence of some form of information and communication technologies, determine our shared and cultural affordances such as online learning (which I will discuss shortly). Ideally, these affordances that facilitate cultural coexistence are equitable, and the resulting proficiencies are identical for everyone. However, the extent to which the proficiencies are achieved within any learning community of practice, as gauged by standardized or non-standardized measures, varies across individuals. Such differences arise as a function of how much learning occurred at the individual level as well as due to the pedagogical practices that did not tailor learning for an individual's needs, and subsequently due to the way that learning was measured. Research to understand individual differences in the quality and quantity of individuals' learning stemmed from the need to identify and reduce these gaps in achievement and performance. Lately, an increasingly higher emphasis is being placed on using digital

data to transform online learning with educational technologies and the providers of such technologies becoming hopes for educational access and equality. The hopes of 'generating data to close achievement gap', 'protecting data to ensure privacy', and 'using data to expose inequality' have been some examples for upward trajectory of online learning. Macgilchrist (2019) described these endeavors to be enacting 'cruel optimism' where the desires of edu-technical transformation is tied to a fantasy of good education. The hope that educational equity is somehow linked to educational technologies and increased data is optimistic. However, the author relates this idea as being cruel given that this idea is tied to a 'fraying fantasy' of a good life. Eventually, this hope, or the 'cruel optimism' is that understanding data of such individual differences is going to lead to reduced achievement gap and improved equity or access (of learning and beyond) to cater to everyone's need through personalization. Unfortunately, the hope that data can at least empower learners is far from achieved as I will discuss later.

With the advent of machine learning and artificial intelligence, researchers have sought to use predictive models that understand the relationship of learners with teachers, learning materials, digital learning environments, and user experiences within these learning environments. Older generations of personalization efforts included intelligent tutoring systems and content-sequencing via semi-automated approaches (i.e., rule-based systems) which were later replaced by data-driven personalization approaches that may sometimes be determined by neural networks that are difficult to interpret (Baker, 2016; Graesser, Hu, & Sottolare, 2018; Kulik & Fletcher, 2016). Specifically, in the past decade, revolutionary improvements occurred in the field of information technology that led to a significant change (quantitatively and qualitatively)

in online learning environments, smartphones, and social networking. These transformations led to widely available data that were not possible in the past. While “Data Mining” as a field emerged several decades ago, it is only recently that it has gained popularity and traction in the field of education (C. Romero & Ventura, 2007; Cristobal Romero & Ventura, 2013). Educational data mining (EDM) evolved as a separate field of research since it was proposed to be different from conventional data mining approaches due to the co-dependence (and hierarchical nature) of features that promote learning. The emphasis in EDM is typically placed on the multi-level hierarchical modeling that are based on psychometric, socio-cultural, and cognitive theories of learning. Therefore, EDM as a field evolved to be an amalgamation of theories and techniques drawn from fields such as computer science, learning science, psychology, and mathematics (Dutt, Ismail, & Herawan, 2017). Learning analytics (LA), a similar field that is gaining popularity recently, also has similar aims as EDM. Both EDM and LA aim to improve education by analyzing large datasets to extract useful information that stakeholders can utilize to promote learning. However, while EDM places a focus on automated discovery of information by reducing the information to its components, LA places a focus on human judgement by empowering instructors and students (Cristobal Romero & Ventura, 2020). While both fields are interrelated and often work together, each have their own issues that are yet to be addressed. For instance, stakeholders (learners and instructors) are often not knowledgeable in the ways of EDM and LA and how to use the insights derived from these methods. Another issue is the diversity of terminology in use to represent the overall approach and a lack of standards to compare the literature. Furthermore, not every published work related to

LA accounts for the inherent hierarchical nature of the data generated in learning contexts. Romera and Ventura (2020) discuss how the fields of EDM and LA have evolved enormously over the past decade and show that the extant literature used terms such as ‘Academic Analytics’, ‘Institutional Analytics’, ‘Data-Driven Education’, ‘Big Data in Education’...etc., which lead to increased confusion among the stakeholders.

In my thesis, I assess the state of the current theories of practice in EDM and LA, specifically looking at the various predictive modeling approaches used, and discuss the shortcomings of some of the approaches taken. I take the view that leveraging data pertaining to the affective, motivational, and socio-emotional process of the students are critical to improving predictive modeling of online learning. I demonstrate a highly reproducible stepwise approach to predictive modeling across learning scenarios to evaluate the predictive value of learning performance trends, demographics, automatically generated data from learning environments (click behaviors), and learner-centric data. In addition to presenting a general predictive modeling framework, this thesis is comprised of three major components: (1) establish and validate *a multivariate predictive modeling framework* that utilizes on-task performance and demographics in a learning environment that is *not constrained* by factors that affect learning in classroom settings (e.g., differences in learners’ subject knowledge). Using a dataset in working memory training context, I show that earlier predictions are stronger when predictive models learn from demographic data and information related to the training environment, facilitated by data mining. I show that using information of on-task performance for the first few sessions, predictive models can reliably differentiate

learners who perform above the median level. (2) Next, I use this approach to evaluate predictive performance within a blended learning environment while incorporating learners' study intentions and corresponding click behaviors within the learning environment into the models. I show that understanding learners' intended study spacing are useful for improving predictions whereas click behaviors (including both the quantity and frequency) do not improve predictions. (3) With a conclusion that click behaviors do not improve predictions in the blended learning environment that I evaluated, I further investigate the validity of click behaviors and their predictive value in a fully online learning environment. Furthermore, I evaluate the extent to which dynamic changes in individuals' motivation, affective, and socio-emotional processes measured throughout the learning phase affect the performance of predictive models. Relying on survey measures of the dynamics of features central to learners ('learner-centric' measures), I show that learners' self-reported responses on these measures are more valuable than artificially derived indicators of learner traits such as engagement, diligence, and procrastination through click behaviors. Ultimately, the predictive models tested here can be utilized to identify potentially low performing or high performing individuals in various learning scenarios at the earliest possible time during the learning period or compare predictions at different time points during the learning period given the data constraints. Subsequently, I discuss the potential of predictive modeling approaches have to understand and identify learner-centric measures that are critical to making accurate and robust predictions of future learning without relying on click behaviors that do not provide context of the learners' behaviors. These three

components and the overall discussions and implications of the current work will be discussed in the next four chapters.

In the rest of this chapter, I introduce and define learning (as applicable for this dissertation) followed by a review of prevalent predictive modeling approaches, their heterogeneity and the diversity of results and interpretations. Next, I introduce the need to compare the predictive models across learning settings (not just in classroom settings such as online or blended learning classrooms). Then, I discuss the overall methodology of predictive modeling using standard machine learning models, the metrics used to evaluate these models, and their limitations. In chapter 2, I introduce predictive modeling in working memory training context. I discuss the benefits of this comparative model to establish the stepwise approach I use for the next two chapters within classroom learning contexts. In chapter 3, I evaluate, replicate, and build upon the model from Chapter 2 within a blended learning context. I discuss the differences between the learning contexts in Chapters 2 and 3, compare the results, and discuss its shortcomings. In Chapter 4, I evaluate, replicate, and build upon models from the two previous chapters in a fully online learning context to validate the need for learner-centric measures for making optimal future learning predictions. Finally, in Chapter 5, I discuss the implications of the current work, its advantages, its limitations, and connect them back to the extant literature. Furthermore, I discuss some approaches that can be taken in the future to empower predictive models that aim to personalize and promote learning. Finally, I provide a summary of the key contributions of my work and I present my thoughts on how to improve predictions of learning occurring within any setting. Now, more than ever, the importance and limitations of online learning are evident. We

are living in the times of one of the greatest threats to education across the globe. According to a recent report published on “Education for Global Development” (Saavedra, 2020), more than 1.6 billion children and youth are unable to attend schools across 161 countries due to COVID-19. Amid these challenging times, much of the learning has moved to online venues in hopes of continuing to educate the future generations, leading to soaring rates in the usage of online learning. Thus, online learning has changed from being *an* option to being the *only* option for many students. As we are moving our imperfect brick-and-mortar offline education system to an online education realm, the need to ensure that we surpass the shortcomings of learning environments is greater (Fernandez & Shaw, 2020). Thus, I hope that this dissertation work adds to the discussions of how we should approach promoting educational equity through thoughtful and careful implementation of our understanding of learners’ needs, rather than by understanding data generated through context-agnostic clicks. After all, there is no true “personalization” without knowing the “person” and their needs.

### **1.1 Defining Learning.**

Learning and other forms of experience-dependent changes are multi-faceted and complex phenomena. For instance, learning begins through the attainment of declarative memories, the development of appropriate cognitive as well as sensory skills either through instruction or practice, followed by the organization of knowledge acquired into general representations. In the end, these changes require continual work towards improvisation that occurs through experimentation and exploration (Olesen, Westerberg, & Klingberg, 2004; Wickelgren, 1981). In general, the processes and products of various dimensions (e.g., cognitive, socio-cultural, and motivational factors)

work together to enable behavioral changes that are classified as learning. For instance, the cognitive dimension of knowledge is assumed to be a result of interactions between objects (stimuli) and people (organisms). An individual that engages with a prolonged stimulation and undergoes consistent interaction with the stimulus such as the working memory training (described in detail further below), often demonstrates experience-related changes in the underlying cognitive architecture which result in learning on the working memory task (Olesen et al., 2004). The extent to which the learning occurs, however, is dependent on many different factors beginning with the ways in which an individual chooses to learn.

An individual who wishes to learn something new or gain a new skillset, has a few different options: observation, imitation, guided instruction, discovery learning etc. Let us briefly compare two options: *imitation* or *guided instruction*. Imagine that an individual wants to learn to play Chess. Observation and subsequently imitation are the primary means to learning basic moves of the game (Bandura & Jones, 1962; Moore, 2004). Several aspects of the learning behavior such as patterns, states, actions, or desirable outcomes maybe imitated via observation. However, the extent to which an individual learns solely by imitation is limited. In certain critical situations, it might even be considered hazardous (e.g., aircraft piloting). Guided instructions by an expert are considered far more effective in many scenarios especially if a learner wants to advance from being a novice to being proficient at the learned craft. Instructions that are based in behaviorism or cognitive theories do not expect learners to derive strategies and rules independently. Instead such an approach depends on the learner's expected skill-level at a given point, their cognitive, and non-cognitive limitations. The content being taught



at any given point of time depends on what the learner needs to know until the learner becomes proficient (Reid & Stone, 1991). An expert or a master of the craft, however, transcends the skills acquired through guided learning and relies on situational and intuitive cues that are specific to any given situation. For instance, going back to the example of chess, a chess master makes appropriate moves every time he/she sees a meaningful chess array drawing on past experiences (Dreyfus & Dreyfus, 1980).

Regardless of the means, the resulting learning that might have happened is broadly considered to take two different forms. First, learning manifests as *knowledge acquisition*. For example, when a person is said to have learned a subject, e.g., biology, we assume that this person understands a significant amount of the concepts, theories, and the underlying mechanisms that connect these concepts to each other and to living organisms. The broader (or deeper) the knowledge, the greater the person's ability to explain more scenarios pertaining to the field. This type of learning is commonly demonstrated via curricular training and formal education. According to Michalski and colleagues (1983), knowledge acquisition is defined as "learning new symbolic information and [the] ability to apply this information effectively." Secondly, learning connects to *skill refinement*. For example, when a person is said to have learned to juggle, we assume that this person has improved the motor and cognitive skills required to master juggling through sustained practice and/or training. Much of this form of learning process, i.e., skill refinement, involves very little learned symbolic knowledge. Rather, it focuses on the necessary 'refinement of skills', either motor, perceptual, or cognitive, via repeated practice and improvement at the subconscious level (Michalski et al., 1983). While the quantity or quality of learning depends on how the learning is

facilitated (e.g., imitation or guided instruction) and what the end goal of the intended learning is (e.g., learning a new skill or to increase the depth and breadth of a subject knowledge), there are other context and setting specific factors that are known to effect learning to varying degrees. An in-depth understanding of these factors is critical to understand the process of learning and its key determinants. Before I discuss these factors, it is critical to differentiate online learning and offline learning since the key constraints and determinants of learning occurred can sometimes differ (T. Anderson, 2004; Reinig, 2010; Van Bruggen, 2005).

## **1.2 What is Online Learning?**

Online courses, once defined by van Bruggen (2005) as “courses that use the World Wide Web to deliver some form of instruction to learners separated by time, distance, or both”, are a growing venue of learning with millions of users. Online courses, as I have mentioned earlier, have become the most ubiquitous form of learning in the past three months with learning taking place across internet-based learning avenues that go beyond the world wide web (as of May 2020). Online learning has been gaining increased enrollments, even without consideration for the increased numbers due to the pandemic we are experiencing. Online course enrollment had grown at a very rapid pace in the past decade (> 9% year on year growth) with at least one in 3 students in the United States taking an online course (Seaman, Allen, & Seaman, 2018). While there is no peer-reviewed literature to report the numbers for early 2020, we can suspect that virtually all students with access to the Internet and adequate hardware right now might be taking some form of an online course. Several works by Bowen and his colleagues have shown over the past few years that these numbers

have been increasing due to the growing expenses of higher education and since online learning has been pitched as a frugal alternative (H. R. Bowen, Fincher, Bowen, & Fincher, 2019; W. Bowen, 2013; W. G. Bowen, Delbanco, Gardner, Hennessy, & Koller, 2013; Škrinjarić, 2014). The numbers were also boosted due to the effects of recession across United States after 2008 where stakeholders at the state policy level urged the public university systems to embrace online courses to increase access to high-quality education (W. G. Bowen, Chingos, Lack, & Nygren, 2013). For instance, universities across California have increased their catalog of fully online courses ever since, that are available for cross-campus enrollment as well as made available on some private online learning platforms such as Coursera and Udacity. This push for online learning as a cost-effective solution has added benefits of enabling learning without being limited by time and space constraints.

In the existing literature, before evaluating the effectiveness and value of online learning, it is critical to examine the differences between online and blended learning environments first. Terms such as “e-learning”, “cyber-learning”, “web-based learning”, and “internet-based learning” are used interchangeably with “online learning”. While, “blended learning” and “hybrid learning” are used synonymously. The key difference between the two is the extent to which a course involves the use of internet (Bienkowski, Feng, & Means, 2014; Hubackova & Semradova, 2016). Typically, a course is considered “online” if more than 80% of the content is made available via internet. On the other hand, a course is considered ‘blended’ if at least 30% of the content is available online and 21% of the content is taught face-to-face. Any course that does involve online elements but limited to less than 30% are referred to as “web-

enhanced” learning (Alpert, Couch, & Harmon, 2016; McPartlan, 2020). Let us now consider the factors that predict learning, whether it is in an online, offline, or blended learning environment and the need to account for individual differences while optimizing learning.

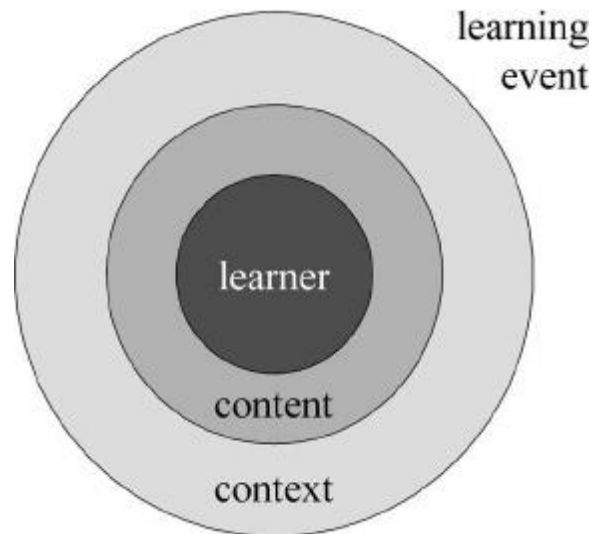
### **1.3. What predicts learning?**

Individual differences in learning determine the extent of success of an individual as measured by standardized tests. For example, the differences in learning attained is a consequence of how an individual has learned/practiced to respond to symbolic systems and social cues over the past several years of his/her life, and due to the differences in his/her aptitudes (Corno et al., 2001). Therefore, much of the research relevant to learning in the past century had a primary focus on identification of student characteristics and learning environments that are optimal for learning outcomes. Several theories and models of learning have been postulated in the past few decades to describe overall learning and to identify individual differences that explain differences in learning. However, we do not yet have a parsimonious and coherent theory that explains the performance of learners and their individual differences (Cronbach, 2003) due to the complexities and the sheer number of domains that contribute towards learning.

It is well established that students’ behavioral, cognitive, and non-cognitive characteristics before, during, and after the periods of instruction (or training) interact with various environmental factors to produce differential learning outcomes. The differences in what an individual knows at any given point of time and the subsequent learning outcomes may be driven by individual differences in several dimensions that

may broadly fall into three categories (see Figure 1.1) – *the learner, the content being learned, and the context in which the learning is happening* (Beckmann & Birney, 2012).

According to Beckmann & Birney's review of adult learning literature, *learners'* cognitive and neural processes that drive learning undergo changes with age. These changes occur in two mutually conflicting directions. Ageing is often synonymous with age related *declines* in domains such as working memory capacity, information processing speeds, motor skills, inhibitory control, and attention controls. Whereas, the increase in knowledge over the course of age, experience related learning, development of schemata that trigger optimal and automated responses to situations characterize the *growth* of learning. While these two opposing general rules are common for most individuals, it is an individual's opportunities to indulge in learning and the levels of engagement that determine the amount of learning. Differences in various trajectories in these two mechanisms differ for each individual, thus, resulting in a wide range of diverse learner profiles (Khribi, Jemni, & Nasraoui, 2009; Premlatha, Dharani, & Geetha, 2016). While the differences in learners' abilities may be due to cognitive capacities, many non-cognitive characteristics of individuals such as attitude, self-efficacy, implicit beliefs have major contributions to the outcomes of learning. Studies show that these non-cognitive characteristics are vital later in life of individuals due to increased self-directed learning, autonomy, goal-orientation, and the declines in cognitive and perceptual skills (Bo, Borza, & Seidler, 2009; Newport, 1990).



*Figure 1.1.* Features of learning -- A simple representation of the relationship between the learner, content, and context reproduced from Figueiredo & Afonso (2006). In any given learning event, the factors that are central to the learner interact with the content being taught, and the context of learning (environment).

The *content* being learned by an individual might be determined by the level of selectivity of the individual's learning goals. Learning might occur as "learning for the sake of learning" that focuses on conceptual knowledge acquisition that a learner is interested, to a more selective form of learning that focuses on skill refinement. This transition in learning that is mediated by selectivity of learners' needs is targeted towards skill acquisition and to succeed in the skill-labor market (Beaudry, Green, & Sand, 2016). The selectivity that drives the transition from knowledge acquisition to skill refinement is suggested to serve two functions – selectivity in order *to overcome age related decline* in cognitive and executive functioning resources and selectivity in order *to learn for an immediate skill application* such as a career or more generally, to solve specific problems (Beckmann & Birney, 2012). As a result of the interactions between the selectivity of learning and its underlying motivation to engage in learning activities, differences in learning quality might arise differently for each individual. The levels of interaction of each individual with the content being learned might also differ by the

experiences an individual might have incurred with similar learning material in the past. As a result, the individual differences in the experiences accumulated over time also contribute towards differential learning outcomes. While interest in a particular learning material might be the driving force for learning, the quality of learning might not simply be a function of how interested an individual is in the content being learned. The amount of learning is also driven by whether a learner is going out of their comfort zone to learn a material rather than being confined to their comfort zone. As a result, many lines of research have been focusing on making the content optimal for best desired learning outcomes. Inducing desirable difficulties -- where the conditions of learning are varied over time and space, learning is interleaved, learning is spaced as opposed to massed/crammed, where learners are required to take frequent tests -- are introduced into the learning content to enhance learning (Bjork & Bjork, 2009).

The *context of learning* adds further complexity in the ways in which learning occurs. As the population of learners move from a highly structured classroom setting to an independent learning or a skill-performance based setting (e.g., software coding certification programs) the diversity of learning contexts increases. This move from structured to semi-structured or unstructured learning is, typically, closely tied to the biological age of an individual. Relatedly, a major focus of research in the past few decades have been on the differential learning experiences of online (semi-structured/unstructured) and traditional classroom (highly structured) settings which are detailed further below. Overall, a mixture of factors that are internal to the individual (interest in the learning material, level of cognition, and cognitive load), to communities of learning practices, as a function of societal, historical, and socio-cultural affordances,

factors that are external to the individual (learning at home vs learning at school, public vs private education, tutoring vs lack thereof, etc.) drive the learning of each individual differently. The individual differences in these three dimensions (learner, content, and context) interact with each other to produce learning to various degrees during any learning event. However, the exact mechanisms by which these interactions drive learning are yet to be fully elucidated.

Several models have attempted to understand the ways in which the three factors affect learning and the ways in which they are interlinked. Walberg's educational productivity theory and Hattie's synthesis of meta-analyses for higher education pedagogy (Hattie, 2008) are a few theories that empirically tested a wide range of factors that might determine learners' outcomes (H.J. Walberg, 1981; Herbert J. Walberg, 1984; later reviewed by M. C. Wang, Haertel, & Walberg, 1993). Wang and his colleagues' work (1993) used an accumulated evidence from three methods - a series of meta-analyses on 3000 studies, reviews of handbooks and narratives, and opinions of experts – to estimate the differential influence of various factors on learning. Specifically, the work resulted in a "*knowledge base*" which included 228 variables across 30 categories that influenced learning. Furthermore, Wang and his colleagues identified 11 factors that had the most impact on learning: student's developmental level (or age), student's ability (prior achievement, motor and non-motor skill levels included), motivation, amount of instruction, quality of instruction, psychological environment of the classroom, home influences (including parental influences), peer group influences, exposure to mass media, classroom environment, and adaptive instruction (instructional delivery system, program design, and implementation included). Their work also



suggested that *distal/background influences*, the factors that are not central to the learner such as state, district, and school policies, are less important to learners' outcomes. In contrast, *proximal influences* such as psychological and motivational factors, quality and quantity of instruction, and variables relevant to home environment (e.g., parental involvement in child's learning) are the most influential to learning.

Furthermore, this work suggested that interventions which focus on student-level variables such as cognitive, meta-cognitive, socio-behavioral, motivational, and affective characteristics will yield improved learners' outcomes as these factors are most malleable to targeted interventions. Ultimately, the purpose of these interventions should be to provide significant positive learning experiences that are related to learning outcomes. Haertel, Walberg, and Weinstein (1983) performed a comprehensive review of 8 models of school learning that continue to influence interventions. Their review included: (a) Glaser's psychological learning theory (1976) which investigated resources, actions, and practices that foster competence in children such as learning-to-learn, and reinforcement of learned materials with an emphasis on the cognitive aspects of the learner, (b) Carroll's model of school learning (J. Carroll, 1963; J. B. Carroll, 1989), which investigated the relationship of efforts spent on a learning task and the efforts required to accomplish the task. This model emphasized the importance of learners' aptitudes to the extent of learning, (c) Cooley and Leinhardt's model that investigated classroom processes (1975), which described a model for classroom performance of the learners with an emphasis on students' initial general abilities and prior achievement such as SATs and cumulative GPAs to understand the extent of learning. This review by Haertel and her colleagues also included other influential

models of learning put forth by Bloom (1976), Bennett (1978), Gagne (1977), Harnischfger & Wiley (1978), and Bruner (1966). A majority of these eight models included very similar constructs to understand learning behavior despite the differences in the nomenclature of several constructs across models. Learners' abilities, motivations, affects, state of mind, and the quality and quantity of learning are identified as the most recurrent constructs that were represented in these models. These models, however, did not include several important factors that influence learning such as school and classroom environments, peer influences, parental influences, and the influence of the mass media that effect learning. Newer models were built on the foundations of Haertel's et al.'s review to integrate the key factors that were missing from these eight models to better determine learning. For instance, Zins and colleagues (2004) explored the importance of learners self-regulation, social-emotional skills, and motivational alignments in determining learning and found that social-emotional aspects of a learner add a significant value in understanding learning behaviors.

Wang (1993) and Haertel's (1983) work has been a cornerstone in learning literature. Together, these works reviewed and integrated over 3000 research studies and eight prominent models of learning to identify the relative importance of key dimensions of individual differences in learners' characteristics that predict learning. Of those learners' characteristics, factors such as cognitive, motivational, affective, psychomotor, social, and behavioral domains are identified as the means to various trajectories of learning (and the user profiles that are related to these trajectories). These characteristics were suggested as the domains that may be most malleable to policies and interventions that seek to improve education and learning outcomes.

However, practical applications of these models to generate novel policies or interventions are significantly challenging due to several reasons. For instance, since each of these domains are built on a wide range of multivariate variables that encompass the broader domain, educators find it difficult to effectively target specific aspects of the learners to produce positive outcomes. The level of complexity of the underlying variables that determine the learning quality and quantity increases in the absence of the overarching latent variables. For example, Cattell-Horn-Carroll (Gf-Gc) theoretical model (CHC model) evaluated over 70 specialized cognitive abilities that load into eight cognitive domains (Schneider & McGrew, 2012) that have varying effects on learning. Creating interventions in real-life scenarios, that target cognitive abilities alone, may not yield the best results, given that one needs to account for a complex network of multivariate predictors to sufficiently understand the impacts of this intervention, while assuming other influences on learning are “*controlled*” for. In addition to difficulties in creating meaningful and effective interventions, the underlying complexity of the networks of domains further increases the levels of difficulty in building theoretical and practical frameworks than can be empirically tested for the relative contributions of each of those variables to learning.

In summary, one outstanding limitation of the comprehensive models that seek to explain learning is that each of these models make use of latent construct variables that hope to capture the meaning of large sets of underlying variables which may not fully represent the meaning of the variables’ interactions. These latent constructs, often, do not account for the richness of the data underneath. Some of the key research conducted on individual subdomains, or those that are studied in isolation -- such as

motivational orientations, aptitude, cognitive prerequisites, social-behavioral aspects, and self-regulation of the learners -- identified and proposed several theories of learning that impact students' outcomes in isolation. Such models come with their own limitation, in that they speak for isolated influences of each variable (or sets of variables) separately. This limitation is further exaggerated due to the nature of using averages to explain the contributions of these isolated variables. While averages of the latent variables may speak for the general relations of these variables to the learning outcomes, they fall short in identifying the differential learning trajectories of each individual. Part of the solution to this problem is to create both explanatory and predictive models in parallel using a single multivariate framework to understand learning, predict learning, and apply solutions that drive personalization of learning.

As an overview, learning is influenced by several dimensions of variables that are in-turn influenced by multivariate sets of predictors. These dimensions include cognitive, meta-cognitive, motivational, affective, state, socio-cultural factors, etc. Much of the existing empirical research focuses on a very narrow set of these dimensions to understand learning. A brief look at the research on motivational factors, both extrinsic and intrinsic to the organism, show the complexities of factors that tend to drive learning. Besides cognitive factors, motivation, and the levels of engagement of an individual on a task also significantly effect learning. The amounts of engagement, in turn, are often driven by the socio-cultural needs of individuals. The most optimal learning occurs when a few conditions are sufficiently met. One such ideal learning scenario might involve an individual engaging in a task that is of interest to him/her, consists of leaning tasks that fall at the difficulty level that is desirable, provides

satisfaction, enhances the sense of belonging in a community, ensues peer, teacher and/or parental favor, and subsequently a prolonged engagement with the task in a spaced learning setting (Shanks, Holyoak, & Medin, 1996). Models and theories that fall short in explaining the connections between the factors, especially at individual levels may not lead to effective interventions. The complexity of learning as a phenomenon poses several challenges, especially where practitioners seek to improve curriculum designs, developing clear goal and sequences of the teaching materials, and to reinforce later learning. While it is important to establish a one-on-one relationship between each of the variables to the learners' performance, practical applications of this understanding to real-life situations may not fully reflect the anticipated levels of improvement in learning.

Learning as a phenomenon involves many means and many ends which differ as a function of all the factors mentioned earlier. When the various models of learning are incorporated into a single framework using empirical research or using meta-analytical approaches, the average values or the relative contributions of each factor (e.g., measured as T-scores by Haertel et al., 1983) do not explain the variances of learning at an individual level. There is no known one-size-fits-all approach to learning. Thus, in the recent past, a demand for personalized learning platforms such as intelligent tutoring systems increased in popularity (Hong, Chen, Chang, & Chen, 2007).

#### **1.4. A case for personalization.**

In typical grouped learning environments (e.g., classroom) that facilitate skill refinement or knowledge acquisition, educators strive to make the learning optimal for all students in that setting. However, the optimal settings laid out by an instructor for one

student (or the average student) may not fit the needs of other students due to many differences among individuals. Take the following scenario as an example: In the 1950s, the United States Airforce investigated the difficulties that pilots were having in controlling the flights. As part of the investigation, Hertzberg & Daniels (1952) found that the cockpits were designed to fit the average pilots of 1920s. They then proposed that cockpits that are designed based on updated measurements to fit the needs of pilots in 1950s (with an understanding that the average 'size' of pilots increased by 1950s) will fix the problem. Thus, they measured the size of more than 4000 pilot on ten dimensions (height, weight, ...etc.) to update the cockpit design based on the new average. However, it was later found that the human errors in aircraft piloting continued since the updated cockpit did not fit the needs of any of the pilots due to the individual differences in the measured dimensions. Follow up work later showed that only 3% of the pilots fell into the average range even after restricting the dimensions to only three of the ten. Not even a single pilot fit the averages on all ten dimensions. This finding, eventually, led to growth of adjustable and ergonomic design approaches (Barbé, Mollard, & Wolff, 2014) that might better fulfill the individual pilot's needs. This is similar to the concurrent situation of fixed learning environments that tailor the contents for average individuals.

At a group level, learning is known to be improved by optimizing the routine activities conducted in a classroom. For instance, consistently organizing information that a student needs to process (e.g., format of the learning materials), regular classroom activities (e.g., frequently test-taking), or reflection of content and goals at the end of each learning session (Gu & Johnson, 1996; McDaniel, Roediger, & McDermott,

2007; Roediger & Karpicke, 2006) have a positive impact on learning. Detailed classroom learning objectives implementation, clear formative assessments and feedback on tasks and performance of students, and reinforcements such as praise from the teachers that satisfy the learners' motivational needs are also known to facilitate learning (Nicol & MacFarlane-Dick, 2006; Norcini, 2010). However, accomplishing these broader goals do not fulfill individuals' learning needs. For instance, if a learner has other (non-academic or non-course related) goals that are conflicting with the learning demands such as time spent on a learning task, the general pedagogical principles may not satisfy the requirements for an individual to improve on the learning task (Deci, Ryan, Vallerand, & Pelletier, 1991; Reiss, 2012).

Three broad strategies were suggested to improve learning at individual levels by overcoming the limitations of the pedagogical principles applied to group settings – *individualization, differentiation, and personalization* (Amanda Stedke, 2017). *Individualization* of learning is one strategy where instructions of learning material are adapted to different speed levels (pacing of the learning) based on the learners' needs. This approach typically does not alter the overall learning goals of the course or the ways in which course material is delivered (in terms of different methods of teaching). All students in the learning setting are expected to learn the same material, via the same modality, and with the same learning goals. Each student, however, is allowed to self-pace their learning. Students are allowed to learn material faster or slower as they need, skip content, re-read content that they need additional help with (Greeno, Collins, & Resnick, 1996; Heift & Schulze, 2015).

*Differentiation* of learning is the one of the strategies where unlike in the Individualization, the instructions are different by person. The learning goals, speed of learning, and the material being learned are same for all the students and the difference arises in the medium in which the learning material is provided to the students (Gibson & Gibson, 1955; Thomas & McKay, 2010). Based on a notion that individuals have preferred learning styles (verbal and visual learners) referred to as meshing hypothesis, researchers used differentiation as a means to improve learning, university-level teaching strategies, and tutoring services. Following a critical review of the literature relevant to learning styles, Pashler and colleagues (2009) concluded that the empirical evidence for learners having varying styles does not exist. Subsequent studies by Rogowsky and her colleagues (2015) supported this evidence against learning styles. However, the ability to differentiate the learning material into different learning modalities such as auditory and visual material is known to have benefits as long as the material in different modes support and supplement each other as opposed to adding conflicting or noisy information (Moreno & Mayer, 2007; Picciano, 2009).

*Personalization* of learning refers to the overall tailoring of learning experiences. In this strategy, learning material is altered to fit the needs, goals, speeds, and to the learners' preferences. This refers to environments that are fully personalized and thus encompasses the ideas of differentiation as well as individualization. Whereas the other two strategies focus on the teacher-environment level characteristics, personalization focuses on the individuals' characteristics. Thus, personalization gives the choice and voice to the learners and makes the learners active participants in the design of their



own learning. In the 21<sup>st</sup> century, personalization is made possible due to the advent and increasingly available online learning platforms (Amanda Stedke, 2017).

### **1.5. Online Learning and personalization.**

Personalization of learning, especially of the learning that occurs via online platforms, has been a focus of research in recent years. Since online courses are a growing venue of learning with millions of users, and since they are conducted in information technology-driven platforms, personalization by means of tailoring online content has increased in popularity in the recent past. Personalized learning via online-platforms emphasize independent, student-led, and out-of-class learning experiences. The ability of online-learning platforms, in theory, is said to promote learning by allowing the students to monitor the progress of learning, pace the learning, and by setting personal goals that they are keen to achieve. Personalization via technology is also hypothesized to improve “non-grade band curricular frameworks,” where individuals of different age groups and grade levels can learn the same content at the same difficulty level, based on their current skill level and personal goals. While such technology enabled personalized learning is a key to improving individual level learning, the existing research does not support the idea that increased diversity in learning personalization translates to improved learning overall. One of the predominant issues that plague online personalized learning platforms, such as Coursera, is the immense rates of dropout (Patterson & McFadden, 2009; Rivard, 2013). Online degree completion rates remain around 30% in most developing countries. While many institutions offer academic support programs that aim to increase retention and course completion rates, the problem with dropouts persist. However, the amount of dropout rates could be due

to factors such as the students learning the content that they are interested in (as opposed to taking a course for the sake of completion or a certification/degree), which is one of the key purposes of personalized learning platforms. This shows that the ways in which the adaptive algorithms define “completion” are in some cases, non-adaptive. The notion of large dropout rates being attested to online learning do not hold true when the online learning occurs at a higher education institution since the goals of learning are set by instructors and the course is required to be completed within the set academic time frame (i.e., semester or quarter). Thus, some researchers posit that despite the high attrition rates, the issue of retention rates may not be as severe as the numbers denote. Rather, it indicates a need to rethink the way we define completion and students’ success and to improve the adaptive algorithms that determine the goals and mark completion of the said goals (Glance & Barrett, 2014; Rizzardini, Chan, & Guetl, 2016). Aside from the high attrition rates, another issue that remains is the lack of unified models for predicting and personalization efforts.

Many higher education institutions in the U.S. use a software based learning platform, usually referred to as Learning Management system (LMS) that integrates teaching materials, learning activities, course administration tools, and exam administration tools all in one elaborate information management ecosystem (Dahlstrom, Brooks, & Bichsel, 2014). With the help of the tools in LMS, teachers and university management, often collect students’ “digital trails” – when, where, and how the information from the LMS is accessed by the students in order to understand the students’ engagement with the course resources such as lecture materials, exams, and other resources (Mah, 2016). While the goal of many LMS tools is to improve student

engagement, understanding, and eventually learning outcomes via personalization, there is a need to empirically test frameworks that optimize LMS platforms to accomplish these goals.

In her review of “Learning Analytics” and “Digital Badges”, Mah (2016) explains the importance of understanding these two key words to build predictive models that may understand the learning trajectories of each individual that aims to improve personalization of learning. As a quick recap, *Learning Analytics*, a method to analyze and mine data from online course learning management systems (LMS) has been a crucial research endeavor in the recent years due to its importance in understanding students’ learning and behavior. For instance, learning analytics are employed to examine the frequency of students’ engagement with the learning material in relation to the quality and quantity of learning of individuals. *Digital Badges* are described as a “new way to capture and communicate what an individual knows and can demonstrate”, often in an online learning environment or on social media platforms such as LinkedIn. Digital badges act as digital stickers of achievements, coding skills, language mastery, digital or non-digital competencies, and affinity or affiliation towards skill sets. Historically, the use of digital badges lies in the world of video games where achievements and mastery of aspects of game leads to badges or levels as rewards that the players use to showcase their skills (Ahn, Pellicone, & Butler, 2014; Hickey, Jovanovic, Lonn, & Willis, 2015; Hickey, Willis, Jovanovic, & Lonn, 2015). Digital badges act as a proxy to measure the current skill level of an individual, their levels of motivation and engagement with the learning material, and as a means to alter his/her self-regulation (Jovanovic & Devedzic, 2015). Overall learning analytics act as powerful

ways to assess, analyze, and predict the quality and quantity of learning as well as the learning environment and may improve retention of learners in MOOCs (Ifenthaler, 2015). While digital badges are said to be useful markers for advancement or progress in learning a skill or subject, the implementation of badges in training, learning, and online courses conducted at universities (that are not part of the open-source Massive Open Online Courses - MOOCs) are sparse.

### **1.6. A case for predictive modeling.**

Humans have attempted using predictions of future to guide the present since ancient times. From shamans in the ancient world seeking signs of success and hope in inanimate objects to scientists using data-driven models to making predictions, the field of predictions as a science has come very far. Machine learning driven approaches are at the heart of current predictive modeling efforts with practical applications in nearly every field. Social scientists are historically known for seeking explanations of social phenomenon and human behavior via interpretable causal mechanisms (Breiman, 2001; Veltri, 2017). Predictive accuracy of these causal models was often ignored in favor of reproducible causal modeling with unbiased estimates from individual parameters. In contrast, fields such as physical sciences, have embraced predictive modeling to drive the field forward. Partly, the reason for this trend is due to the unambiguous predictions and widely available data (Hofman, Sharma, & Watts, 2017). Lately, due to the massive proliferation of data streams, in terms of available data and the quality of the data itself, social scientists are increasingly turning towards machine learning driven predictive modeling approaches. Another reason for the rise of predictive modeling is also due to the increasing concern over the paucity of replicability

of results, and counter-narratives regarding validity of existing causal models (Gilbert, King, Pettigrew, & Wilson, 2016; Open Science Collobaration, 2015). In a recent review, Hofman and colleagues (Hofman et al., 2017) have proposed that three major issues require prompt resolution before social sciences can benefit from machine learning driven predictive models. These include standardizing practice for evaluating predictions from models, establishing theoretical limits of predictive accuracy in complex social systems, and complementing predictive accuracy and interpretability (via explanatory models) and not treating these disparate approaches as substitutes for each other. In order to extrapolate these three key aspects within the context of predicting online learning, first, we need to understand the current state of predictive modeling within online learning contexts.

Online learning has been criticized for not being an effective mode of education since its inception and continues to be questioned and doubted, specifically about its value and tradeoffs as it aims to cut costs (Deming, Goldin, Katz, & Yuchtman, 2015). Thus far, very little concrete evidence has been gathered to address the issues of quality and value of online learning over offline and blended learning. This is expected, and perhaps, very hard to gain evidence for, given that it is near impossible to randomly assign students to one or the other group in a full randomized controlled trial setting (Bienkowski et al., 2014; H. R. Bowen et al., 2019). A fairly recent meta-analysis of the existing research by Lack (2013) and a slightly older review by Means and colleagues (Means, Toyama, Murphy, Bakia, & Jones, 2009) show that combining face-to-face elements is relatively more effective approach than fully online environments for student performance. In the 45 studies considered for meta-analysis by Lack, not a single study

involved random assignment. A few recent papers that investigated the value of online courses using random assignment suggested that the results are grimmer for online learners compared to face-to-face courses (for an example, see Alpert et al., 2016). Furthermore, on a very large-scale study with a sample size over 200,000 comparing face-to-face courses versus their equivalent counterparts taught by the same instructors following the same course materials, showed that online courses are statistically significantly worse for performance and dropout rates. In fact, the very reasons that learners have for taking online courses by themselves were associated with nearly half a standard deviation drop in students' performance. These results were also worse for below-average performers, youth of color and minorities (Bettinger & Loeb, 2017; Figlio, Rush, & Yin, 2013). Given the current levels of online learning enforced on learners, it is even more critical to validate online learning, entertain the idea of personalizing learning, and promote ways to discover insights that can help predict learners' performance at the earliest possible time.

Learning analytics have the capacity to use dynamic information of the learners in real-time to model, predict, and optimize the processes of learning. Scheffel and his colleagues (2014) introduced a framework with five key markers to determine the quality of learning analytics that can accomplish these goals. The five qualities include:

- i. learning measures and output (i.e., comparability of the analytics, effectiveness and efficiency of the analytics, and helpfulness of the analytics)

- ii. coherent objectives (i.e., awareness of the current learning goals, motivations, and behavioral changes of the learners, and the ability to derive insights and reflections that are meaningful to learners)
- iii. learning support (i.e., classification and detection of students at risk and requiring additional support, perceived usefulness of the learning material available, recommendations of things to do to improve)
- iv. data management (i.e., maintaining standards of data, ensuring privacy and transparency), and
- v. organization level benefits (i.e., ease of implementation, ability to engage non-learner stakeholders, supporting organization level changes, and ability and access to training modules).

While a great learning analytics platform is expected to meet all of these qualities, the most critical of these qualities is the ability of the learning analytics to support learners' self-reflection and predicting their future learning. In a fully personalized learning platform, the personalization begins with the learner. The learner is given an opportunity to reflect, or critically evaluate their current state of knowing, define their own goals, determine the difficulty of content they wish to tackle, and decide on their own pace of learning. The next step in personalization of learning occurs when predictive models and algorithms can appropriately determine the state of learner based on their explicit inputs from the reflection, levels of performance in the early phases, and their motivations to detect learning trajectories and to empower students who are at risk of failure or dropout (Dillenbourg, Schneider, & Synteta, 2001). One thing to note is that while learners are allowed to set pace and difficulty, teachers play a critical role in

determining constraints of these parameters that each student can then manipulate. For instance, students should be able to tweak the difficulty of content within an upper and lower bound which is set by a teacher, just enough, to neither lose interest nor be discouraged completely to tackle the content.

As discussed in the earlier sections, there are many ways to measure learners' outcomes and experiences. In the online learning contexts, often, the measures used to predict students' performance involve variables that account for students' readiness (or current state of knowing measured using ACT, SATs, GPAs - Conley, 2008; Komarraju, Ramsey, & Rinella, 2013; Porter & Polikoff, 2012; Venezia & Jaeger, 2013), expectations to succeed (Nadelson et al., 2013), levels of participation and engagement in online course activities, engagement with the learning material on LMS measured via click behavior (quality, quantity, and digital trail of clicks), time spent on resources such as learning material or quizzes that are made available on LMS – time elapsed between two consecutive clicks, and participant demographics (A. Anderson, Huttenlocher, Kleinberg, & Leskovec, 2014; Bayer, Bydzovská, & Géryk, 2012; Chaturvedi, Goldwasser, & Daumé lli, 2014; Guo & Reinecke, 2014; Hershkovitz, Baker, Gowda, & Corbett, 2013; Huang, Dasgupta, Ghosh, Manning, & Sanders, 2014; Ramesh, Goldwasser, Huang, Daume, & Getoor, 2014; Seaton, Bergner, Chuang, Mitros, & Pritchard, 2014; Wilkowski, Deutsch, & Russell, 2014).

However, these studies tend to isolate the effects of these predictors of learning leading to widely differing predictions of later learning (accuracy range of 50%-98%) painting an incomplete picture of the learners' characteristics and how these predictors are related to each other (Mohamed, Husain, & Rashid, 2015). A handful of studies that



used Neural Networks were able to predict learning outcomes with accuracies exceeding 95% but the results from these studies did not provide any means for the end users to determine what worked, what did not work, and what are the most influential factors to their learning (see Kumar & Vijayalakshmi, 2012 for an example). Specifically, work by Kumar and Vijayalakshmi achieved a prediction accuracy of 98% to predict the 5<sup>th</sup> semester performance using the grades received in the past 4 semesters using Neural Networks. While numerically achieving 98% accuracy is astounding, the approach simply accounts for past performance in the 4 previous semesters to make predictions in the 5<sup>th</sup> semester with a conclusion that if you had good grades according to your academic record, you will be above-median performers. Unfortunately, that result is not usable for teachers to improve their own teaching or for the students to improve their own learning.

Additionally, results from a neural network with many hidden layers are very hard to interpret. Factors that are central to the learner, factors that can be manipulated, and factors that the teachers and students have direct control over (and subsequently create action plans for) are rarely included in these predictive modeling efforts. In the review by Mohammed and colleagues (Mohamed et al., 2015), only 4 out of the 30 studies reviewed included student measures such as motivation with accuracies ranging from 65% to 83% in online and offline settings. For example, work done by Sembiring and colleagues (2011) produced 83% prediction accuracies using machine learning based approaches (Kernel K-means and Smooth Support Vector Machines) using features such as study behavior, engagement times, beliefs, family support (4 learner-centric features) and classified the students into 5 categories (excellent, very good, good,

average, and poor). This study included data from 1000 student participants spanning across three different courses and utilized questionnaires for the learner-centric features across 4 categories for students enrolled at Universiti Malaysia Pahang (offline courses). The four learner-centric measures were categorized into high, medium, or low instead of using the full scale of responses. One interesting finding is that when classifying the learners into 5 groups, the prediction accuracies for the top 20% vs. bottom 20% yielded a significantly better prediction rates (over 94%) than for the rest of the 60% students (61%-75%). While these results are promising, they do not provide a means to evaluate the relative importance of learner-centric features compared to models that only use demographics and past performance. Overall, those studies that yielded very high prediction accuracies were either using neural networks that are difficult to evaluate or understand, did not provide sufficient guidance for stakeholders to understand the results, or provide a reproducible framework to understand the relative predictive value of each feature.

Let us now evaluate these results in terms of the three key challenges that Hofman and colleagues identified (2017). First, there is a lack of standards for predictions. Typically, when making predictions using machine learning driven models (within the context of online learning), researchers have two choices: classification or regression. For instance, *classification* models would involve categorical identification and grouping of learners using some form of measure of success within the learning context (i.e., predicting dropout (yes/no), predicting letter grade, predicting success above or below average...etc.) On the other hand, *regression* models would involve predicting outcomes of interest on a continuous scale (e.g., predicting the scores of

learners on final exam, predicting quiz performance...etc.) Each of these two approaches include many different model variants that can accomplish these goals (to classify or to regress). Not every model is ideal for all tasks. Furthermore, the ways in which the performance of these models is evaluated, is also different. Typically, the researchers make choices regarding which goal they seek to accomplish, what model(s) to use, what metrics to report, and how to report these results. This makes it very difficult to cohesively compare results from different studies. Although, researchers in the fields of machine learning and artificial intelligence typically use 'simple-to-understand' quantitative metrics, results within any given field will not be comparable unless there is a consensus on which metrics to use and report. Thus, it is important to report all commonly used metrics where no standards exist.

Next, it is critical to establish the limits of predictive modeling in any given context. Predictability of human behavior is highly context specific. For instance, in a given year, if the probability of someone sleeping in their own bed every night is 80%, then it is easier to achieve a prediction accuracy of 80% by simply training a model that can understand this heuristic. On the other hand, predicting rare events or "black swan" events are near impossible even with more complex models due to the inherent nature of uncertainty associated with these events. For instance, a catastrophic day in stock markets or finding a million dollars under a tree outside a park are rare events and inherently harder to predict using typical modeling efforts. Online learning typically falls somewhere in the middle with intermediate predictability, where learners with good grades are known to do predictably well. However, as discussed earlier, learning behavior is more nuanced and the outcomes resemble either a flip of a coin or finding

the proverbial needle in a haystack, depending on the context. Thus, establishing the baseline and best possible predictions from the modeling efforts for a given context, given a specific dataset, is important. First, it is important to establish a robust baseline model for comparing the results against. Next, it is important to establish the best possible prediction performance given a particular dataset. Note that the best possible predictions need not be a model that gets the predictions correct 100% of the times. Rather, it is the best iteration (or model) that yielded the highest accuracy compared to the baseline. The baseline model is important since it is important to achieve predictions beyond known or expected heuristics. Going back to the probability of sleeping in one's own bed every night, simply guessing the highly probable event for every prediction will yield 80% accuracy (albeit with a very high false positive rate). In the context of predicting above or below average performance within a classroom setting, simply predicting above average for every prediction will yield an accuracy of 50% (since every event is likely to be correct one-half times for each prediction). Establishing the best possible predictions for a given context is important since, often subpar predictions are ascribed to insufficient data and/or poor model quality. However, it is not always true that higher amounts of data yields better accuracies since the inherent increase in noise due to higher amounts of data is known to reduce the quality of predictions (Kwon, Lee, & Shin, 2014). Similarly, more sophisticated modeling by itself does not solve issues that go beyond limitations of modeling efforts. For instance, rare events are harder to predict, inherently due to the nature of rarity and unpredictability associated with the event. Furthermore, it is reasonable to assume that, within learning contexts, the best possible prediction accuracies are a function of context specific factors as well as the

constraints of the task itself. For instance, learning an inherently repetitive task, such as WM training, is likely to be a function of cumulative advantage of practice, structured repetition, and ability at baseline. Whereas, learning in the context of online learning is more nuanced and is influenced by higher number of extrinsic factors (e.g., dynamic motivations that are discussed later), making it inherently more challenging to predict. I discuss this possibility in more detail, separately, in the next chapter.

The next critical issue that needs to be resolved is the coexistence of predictions and interpretations. The primary concern with predictive modeling applied to complex social structures and human behavioral problems is that predictive models involve complex models that are hard to interpret. While this is an important aspect to address, Hofman and colleagues (Hofman et al., 2017) argue that predictive modeling efforts are not at odds with explanatory modeling for three reasons:

a) “simple models do not necessarily generalize better than complex models” (i.e., role of Occam’s razor - (Domingos, 1999)) since generalization depends on the entire research process (including the choices the researcher makes regarding the modeling). Generalization errors can subsequently be reduced using ensemble methods such as boosting and bagging. This approach partially overcomes the tradeoff of generalization despite increasing model complexity.

b) there is a growing amount of evidence that shows that the trade-off between predictive and explanatory models is minimizing. It is possible to achieve an interpretable model that also provides insight into mechanisms that drive phenomenon by reducing generalization errors and by using simplest models that achieve the same prediction accuracies.

c) the notion of being able to understand a phenomenon should be in terms of ability to interpret the gathered data (via explanatory models conducted ex post) as well as being able to successfully account for the patterns within data to make future predictions ex ante.

Within the context of predicting online learning, there is a need to implement solutions to these three issues to successfully apply predictive modeling that are easy to compare across settings. Furthermore, addressing these issues would also help accomplish the goals of being able to replicate results, understand the most important predictors, and eventually to personalization of learning. So, what does a framework that accomplishes these goals look like?

### **1.7. A test framework**

An ideal solution that seeks to solve the three goals discussed earlier, should utilize predictive modeling that follows rigorous optimization and systematic testing of models and features that are specific to a given context. This can be accomplished by following a single framework that is built on common core elements that can be used for comparison across all contexts and settings. In the current work, I use a framework with 4 key elements that seek to accomplish this goal.

1. Reporting all prediction task specific metrics: The first step of this solution starts with identification of the question that needs to be answered. Predictive modeling using machine learning approaches can be used to solve two main types of problems: supervised and unsupervised. While solving supervised problems, an algorithm is trained on a set of features with a set end goal with known labels such as above average or below average performers or end result such as quiz

scores (classification or regression). On the other hand, unsupervised problems require algorithms to organize data without any known set goals (such as known categories of the outcomes). For the context of this thesis, I will confine the discussion to supervised predictive modeling approaches since all of the datasets in the current work are analyzed ex post with known outcomes. Specifically, I will discuss classification models and regression models. Problems that require classification models will require identification of the categorical outcome measures of interest such as classification of learners into above or below median performance or letter grades. Problems that require solutions on a continuous scale such as predicting performance on final exams use a regression model. While the available selection of models differs based the question variant, the next steps are applicable to either variant. Note that I am not including any discussions regarding all of the clustering methods applied in the context of EDM and LA. For a review, please see the work of Dutt and his colleagues (2017). Identification of clustering and validation of the identified clusters are specific to the data, the nature of the clusters, and what a researcher is trying to seek within these clusters that are worth exploration to empower learners. Thus, the current framework neither take clustering approaches into account nor does it provide a lens to evaluate the results of clustering approaches without significant modifications.

2. Baseline Model: The next step involves identifying a baseline model. A baseline model acts a reference point for comparing the results of the rest of the models. The performance on the baseline model provides a means to measure the

absolute increase in performance as well as lift ratios for each subsequent model. There are many ways to establish baseline performance of predictive modeling. Some of the most common approaches for classification models involve predicting true classes using dummy data, predicting most frequent class label,, predicting the class following a prior probability or heuristic, predicting the class uniformly at random, and predicting a constant class to check for expected false positive rates. For regression models, baseline models typically involve predicting median or average or a constant value. For the sake of consistency across all models (classification and regression with class balanced or imbalanced cases), given that the nature of classes, number of classes, or ranges of continuous data typically vary, it is more reliable to generate dummy data for any given context using the average and standard deviations of the features of interest<sup>1</sup>. Example baseline models for classification and regression models are shown in the next chapter.

3. Predicting using early learning: Once a baseline model is established, it is important to understand if the data from learning itself is valuable for predictions. This is an important step given the emphasis placed on existing work. For instance, over 80% of the studies (18 out of 22) included in a recent review included historic performance to make predictions (Mohamed et al., 2015). Perhaps, this model is also important to establish how well and how soon one

---

<sup>1</sup> Across all three studies, standard baseline modeling approaches were also implemented with similar results. However, we decided to use dummy data to make our baseline predictions. We used the dummy data to make predictions on both classification and regression tasks to determine the predictions our models would yield without using any real data to demonstrate that using randomly generated data around the true means are sufficient to derive near chance predictions from our models.



can predict future learning without relying on any other extrinsic variables. In a given context, if early learning behavior alone is able to make reliable and robust predictions of future learning early enough, perhaps the predictive modeling will not require any other features. This would help minimize the use of resources spent on gathering, managing, processing, storing, and removing the noise associated with non-essential features. However, it is important to understand and establish the best possible potential predictability within the context-specific learning scenario. In addition, it is important to assess context specific extrinsic (to the measurable performance trends) features, that might improve performance of the models.

4. Using extrinsic feature sets: The next step involves evaluating the importance of predictors that go beyond early learning behavior. Given that there are many potential differences in individuals' learning behavior and since these differences are driven by context-specific factors and constraints, it is important to establish the validity of these features. However, it is a challenging task to systematically test and understand the value of each predictor within any given environment without leading to spurious results. This issue is likely to be worse if researchers that use new tools out of excitement potentially do not understand the mechanisms of the models. Furthermore, an emphasis on crunching numbers solely to maximize "accuracy" is prone to errors such as overfitting and non-generalizability. Some seemingly safe choices that are made to improve prediction accuracies could lead to wrong conclusions. For instance, reporting the prediction accuracies alone does not account for the costs associated with

false positives and false negatives. Another issue that is associated with predictive models, is the lack of decision-making tools offered to the stakeholders within the contexts of learning. For instance, a highly accurate neural network that uses all features available (and therefore attesting equal significance to all features) can explain why their predictions are correct. However, it does not provide any means to understand why their predictions might be wrong and does not provide a way to exploit the predictability to improve learning in any way. Typically, the first step to evaluating the value of the overall predictive model has been placed on whether or not these models offer any prescriptions for future actions -- decisions that teachers, students, and administrators can take to improve the overall quality of learning. This requires identification of features that are malleable to prescriptions such as training a skill or targeted improvement of motivational and affective traits. Some features that are commonly used for making predictions do not offer any fundamental actionable insights. For instance, in a typical learning environment demographic features such as age and gender, while important predictors of learning, do not offer any means for future manipulation. If age predicts learning in a given context, the most reasonable directives for the future would be to make the content age-generic, or to limit the learning to specific age groups, or to provide additional support to specific age groups. Each of these options, while optimal for the given conclusions drawn from the model, do not offer any real solution given that not everyone within an age group band is likely to perform equally. Furthermore, none of these solutions account for the interest of the learners in the course

(regardless of their age). Thus, for the sake of simplicity, in the current work, I group the extrinsic features into two sets:

- a. Features that are not malleable: While evaluating the value of features within a learning context, it is important to evaluate the features that are inherent to the learner, such as demographics and precursor variables (prior knowledge and experience/skill). These features should be selected based on theoretical importance and empirical evidence.
- b. Features that are malleable: Next, the models can evaluate concomitant and/or post-learning features such as course/training specifics, group characteristics, situational impacts, social desirability, shifts in motivations and affective measures, changes in work load, and other commitments.

The first two steps (deciding on a prediction task and including a baseline model) are used for determining the best fit models and to establish a lower bound for the prediction accuracy. The next two steps are used for determining if there is a need for extrinsic features for making robust predictions beyond the learning itself. Furthermore, the two components within the fourth step can be broken down into multiple steps or combined into a single step depending on the amount of data available, importance and relevance of the features that can improve predictions. In addition, within this framework, I propose to use *all* commonly used metrics that are relevant for the choice of prediction task (discussed in detail in the next chapter). For instance, instead of reporting prediction accuracies alone for classification tasks, it is important to report precision, recall, F1-score, as well as AUC scores. For regression tasks, it is important to report  $R^2$ , adjusted  $R^2$ , RMSE, and MAE as required. This provides a means for

comparing cross-context and cross-model results with relative ease. Note that while this framework enables a means to incorporate a step-wise modeling approach to evaluate predictive models, this does not describe the necessary steps needed to reduce generalization error such as bagging and boosting (see Chen & Guestrin, 2016; Dietterich, 2000; Schapire, 2003). However, in this thesis, I will demonstrate the ensemble techniques used for improving generalization of the models for each study separately given that the sample sizes and features are varying across the three studies.

### **1.8. Key research questions**

Overall, the low accuracies of the existing predictive models that seek to understand learners performance may be linked to (a) inadequate understanding of the dynamic nature of students' learning behavior, especially, the dynamic nature of the situational motivations and their influence on students' performance, (b) not accounting for the study habits of the students (both the intentions and the extent of implementation of those intentions.), and (c) not having a single predictive modeling framework to understand context-specific learning. These three issues are interlinked since solving one of these issues at a time without accounting for the other two issues, will lead to an incomplete solution. For instance, one can use the best possible data-driven approach to consistently predict learning and yield really good prediction accuracies (like the neural networks described earlier). However, if the factors that are considered for these models and the ways in which these factors can be exploited for improving student learning are not readily understood by the teachers or the learners, the issue of achieving personalization remains unresolved. Thus, in the current work I focus on

predicting learning on three datasets across three different learning settings (a Working Memory (WM) training task, a blended learning scenario, and a fully online learning scenario) in order to understand the dynamic nature of learning over short periods of time or *short-burst time spans* (training periods or length of course). This approach provides an opportunity to use net intraindividual variability across the learning periods on various dimensions to examine the changes in performance over time.

The current work focuses on predictive models that seek to understand the contribution of on-task behaviors and a few of the individuals' characteristics discussed earlier such as demographics, spacing intentions, and the dynamic nature of motivations. The key questions I seek to address in this thesis pertains to the dynamic nature of learning and its predictability using a single framework. Specifically, I apply models that examine intraindividual variability of on-task performance over short spans of time alongside characteristics that pertain to the learning settings and the learners themselves (see Figure 1.1). Specifically, I use performance on the *content* (or mastery -- of students' grades or WM training performance), *context* specific predictive features, and *learner*-centric features, in hopes of systematically evaluating a single predictive framework. Furthermore, I investigate the relative contribution of these factors in early detection of low performers. In doing so, I address the gap in the literature pertaining to predictive models of learning by accounting for the dynamic nature of predetermined spacing of learning, learners self-reported study spacing intentions, adherences to the study-intentions, and dynamics of learners' motivations in learning contexts. I hope that the results of this work will add to the conversation around creating a standard

predictive modeling approach that is reproducible across multiple settings and contexts, and eventually, improve the personalization tools for learning.

The aforementioned multivariate predictive model learns from performance during the early phase of learning alone, changes in learning, and later the participant characteristics in a series of stepwise predictive models to understand the quality of learning. These models will be applied in the context of a working memory (WM) training scenario first to establish its utility, and later, in two “real-world” scenarios, specifically, in two different sophomore biology courses held at the University of California, Irvine to make the results more concrete and ecologically valid. I focus on participants’ on-task learning behavior (i.e., training or classroom performance) in the first model to understand how early learning behavior might predict later learning behavior. Next, I incorporate a set of participant characteristics that may improve the performance of the predictive models.

In *Study-1*, I build models that predict learning on a WM training task which targets a crucial component of the brain that acts as a very short-term information storage and processing unit that helps accomplish tasks. I set-up a binomial classification problem to predict whether a learner ends up above or below median performance level after 15 sessions of WM training. I investigate the predictive values of training performance in the base model isolated from the other variables and then incorporate an array of predictors to the full model. Furthermore, by using one session increments leading up to the 15<sup>th</sup> training session, I investigate the changes in accuracy as I incorporate session by session training performance into the predictive models. The final predictive model not only contributes towards understanding the individual differences in learning on a

WM task, but also helps evaluate the importance of each of the features' predictive value, and to understand how much learning information might be required to attain predictions of the final performance in a learning scenario that is not constrained by subject knowledge or socio-emotional and affective processes that effect online learning discussed earlier.

In *Study-2*, I seek to replicate the utility of the predictive framework and test the performance of the predictive models that have a similar structure from Study-1 in a blended classroom setting with context specific features that captures the use of an online learning platform to deploy lectures and exams. Similar to Study-1, I set-up a binomial classification problem to predict whether a learner ends up above or below 50% of the peers in the classroom at the end of the 10-week long course. First, I investigate the predictive values of students' performance on the assignments and tests. Then, I use students' performance scores and their demographics into the model. Finally, I introduce self-reported study behaviors of the students along with click behaviors that are purported to capture the implementations of spacing and procrastination behaviors within the blended learning system. In doing so, I evaluate the relative predictive value of the features that are considered important in blended learning contexts. Similar to Study-1, performance on each of the various tests scores will be added to the models in their chronological order one test at-a-time leading up to the final exam to investigate how much information might be necessary to accurately predict the final grades' outcome. The final predictive model from Study-2 is used to understand the comparability of the results from Study-1 in terms of predictability of learning , while establishing the utility of the framework in an ecologically valid scenario

and how study intentions and click behaviors will enhance predictions. Furthermore, this study helps understand the utility of self-reported study habits and a data-driven measure of students' behavior that approximate those habits (albeit context-agnostic click behaviors), to students' learning and the predictability of the quality of learning.

In *Study-3*, in addition to replicating and testing the models from *Study-2*, I investigate effects of individuals' observed behavior (number and distribution of clicks and time spent on resources) and day-to-day changes in self-reported motivations during online learning to predict the quality of learning. I investigate the importance of intimately understanding the learners' motivations on a day-to-day basis in making rigorous predictions of their future learning. In this part of the proposed work, I investigate learning in a fully online course, that lacks a face-to-face contact with the teacher which increases the need for students' self-regulated learning during the entirety of the learning phase. This poses an interesting question as to how the dynamic nature of students' motivations on a daily basis might lead to differences in learning at individual levels. This final study helps understand the effects of dynamic nature of students' motivations on independent learning. A detailed description of the predictors and the rationale for their inclusion, the specific research questions and relevant discussions are provided separately per study. A short summary of the three studies is provided in Table 1.1. Note that in studies 2 and 3, instead of a median split, I split the students into 2 near equal groups with at least B- or below B- grade groups instead of arbitrary median splits. This was done to make the splits with near equal group sizes as well as to avoid a split where having above and below the median would yield the same



grade. The group splitting procedures are provided in more detail for each study separately.

Table 1.1

*An overview of the three studies of the proposed work.*

<b>Study</b>	<b>Dataset (s)</b>	<b>Sample n (age range)</b>	<b>Predictive Model Variants</b>	<b>Outcome Measure</b>	<b>List of Predictors</b>
<b>Study-1</b>	Working Memory Dataset-1 (lab studies)	775 (7-86)	(1a) Training only  (1b) Training + non-training data	Performance on 15 <sup>th</sup> session of training  Binomial Classification models – <i>above</i> or <i>below median performance</i> (for both WM datasets)	Training performance (sessions 1-14) Age Supervision Stimulus type Learning rates
<b>Study-2</b>	Online learning with a face-to-face component	459 (19-29)	(1a) Test performance (Quizzes, Assignments, Homeworks, and Midterms) only  (1b) Test performance + demographics + study intentions  (2) Test performance + demographics + study intentions +	Final Grade - Binomial Classification models <i>at least B-</i> or <i>below B- grade</i>	Performance on assignments, quizzes, and HWs Age Ethnicity Low income status GPA SAT Study Intentions (spacing) Adherence to spacing

---

			click behaviors		intentions (click behaviors)
<b>Study-3</b>	Fully Online learning	147 (19-25)	All models from Study-2 (1a, 1b, 2)  (3) Final model in Study-2 + dynamic changes in motivation	Final Grade - Binomial Classification models  <i>at least B- or below B- grade</i>	Performance on all tests Age Ethnicity Low income status GPA SAT Study intentions (spacing) Adherence to spacing intentions Dynamics of learner-centric metrics

---

A typical limitation of using a predictive modeling approach lies in its lack of explanatory power in understanding the fundamental connections of the predictors to each other and to the students' grades. However, the predictive modeling approach used in this work goes beyond the traditional statistical modeling in that it looks for hidden interactions between various motivational and human characteristics at each individual's level beyond looking for average scores or traditional statistical hypothesis testing. In the current work, I will test models that are based on a pre-determined conceptual framework I described earlier leading up to the motivational dimensions of the individuals' performance beginning with their measured performances. First, I focus on the applicability of the framework that is devoid of subject knowledge, motivation,

affective, and emotional constraints to understand the value of sequentially introducing features to the predictive models. Next, I evaluate whether or not self-reported measures of motivation and the students' self-attested importance to a specific motivational construct determines the quality of their learning. This may allow the models to explicitly look for interactions of variables of interest. The advantages and limitations of each study are discussed in their respective chapters.

As I have briefly discussed, learning can be a complex phenomenon that is driven by learner, content, and context level factors. While all of these factors are known to help understand learning using predominantly an explanatory approach, there are not many studies that evaluated how well these factors help understand learning using a predictive approach. The existing literature is divided on the value and utility of the predictive approaches in general. Unfortunately, the widely varying success of predictive models that can be attributed to varying degrees of features, differences in modeling approaches, and lack of standards to compare results only add more confusion. Ideally, a good synergy between explanatory, as well as predictive modeling approaches are necessary to understand learning *ex post* and *ex ante*. As more and more predictive modeling efforts are being put forth, I believe that we need to take a step back and understand the current state of literature better in order to understand how to move forward. Predictive models are an important addition to the scientific repertoire that may, one day, help enable personalized educational technology. I believe that we need a better framework to be able to compare the utility of each idea, approach, and predictions that we make. Ultimately, we can only hope that the predictions we are making today, will have a great impact on the learning of future

generations. I hope that this thesis will provide a means to better understand our efforts and subsequently the impact of our efforts.

## **CHAPTER 2: PREDICTING LEARNING IN THE CONTEXT OF WORKING MEMORY TRAINING**

As a summary of the introduction chapter, I have discussed the importance of predictive modeling for online learning, specifically, that predictive modeling can be utilized to make predictions of later learning from early learning in order to understand who needs more help during the early learning. Many studies that have used this approach resulted in varying results, using varying approaches, and using varying feature sets. In the current work, given that the goal is reuse a single framework to evaluate the relative importance of features, here, I explore the idea of using predictive modeling to evaluate a learning scenario in order to understand how predictive modeling can be used, specifically, to predict later learning, and what to expect from such an approach in short-burst time spans. The models that I will employ in this chapter will act as a foundation for the next two chapters. Here, I focus on employing predictive modeling in the context of a working memory (WM) training dataset, where click behaviors and subject knowledge are not relevant. While dynamics of motivation are important for any learning setting, typically, WM training literature does not evaluate this aspect of training performance. In addition, performance within WM training during the later learning phases (or training 'sessions') are highly correlated with early learning performance as learners typically become progressively better with training on average. Therefore, WM training dataset that I discuss here acts a meaningful testing ground to evaluate the validity of using early performance alone to determine later performance. In addition, since WM training performance are affected by demographic features (that I

discuss later), it provides an opportunity to evaluate how utilizing such demographic features along with training features might improve the predictions of the model. The key point I explore in this chapter is the idea that understanding the learners' early learning trends lead to better predictions of later learning, however, that using demographics and other extrinsic features will improve predictions early on but have diminishing value later.

### **2.1. A brief introduction to working memory and its training.**

Working memory (WM) is the cognitive system that allows for storage and manipulation of information, allows for handling of ongoing information while engaging with a task, and is responsible for higher order cognitive skills (Cowan, 2017). WM capacity is critical to successfully and efficiently perform a wide range of activities such as reading, mathematics, mental arithmetic in a classroom environment, or for planning and task execution (Conway, Kane, & Engle, 2003; Engle, 2002; Klingberg, 2010). Improving WM capacity via training programs and WM related skills have far-reaching practical as well as theoretical implications - from improving mechanisms of learning to reducing learning disorders, altering cognitive, and perceptual capacities (Engle, 2018). However, many WM training programs that aim to improve individuals' WM capacity or efficiency yield inconsistent results both within and across studies.

In general, most WM training protocols follow the following structure. One or more groups of individuals are assigned to a training task where the participants are required to perform multiple sessions of training within a given time span (often involving a pre-determined amount of sessions and distribution between sessions). Before and after the training period, each group performs one or more assessments to

determine whether the training led to increased performance on these untrained tasks with many variants in how the interventions are designed and implemented (Pergher et al., 2019). For example, researchers implement a broad range of training stimuli (e.g., letters, symbols, objects, auditory signals), training tasks (e.g., simple span, complex span, N-back), training settings (in lab with supervision, without supervision, at home), training dosage and spacing between subsequent training sessions – all of which might affect training outcome. To date, the full extent of such factors, and how they interact with each other are poorly understood. Thus, to determine and to further our understanding of the quality of on-task learning, it is important to understand whether these features predict learning outcomes beyond demographics. In the current work, we focus on N-back training, one of most commonly used approaches to WM training (S. M. Jaeggi, Buschkuhl, Jonides, & Shah, 2011; Soveri, Antfolk, Karlsson, Salo, & Laine, 2017). In a typical WM training session using N-back task, learners are presented with a stimulus such as letters (visual, auditory, or mixed) one at a time with spaced by a fixed amount of time (e.g., 3 second per stimulus used in Susanne M Jaeggi, Buschkuhl, Jonides, & Perrig, 2008) with the objective of detecting whether the current stimulus matched the stimulus from ‘N’ items back in the series.

Individual differences across multitude of factors, differences in study settings, and the differences in execution of each study are attributed to the inconsistent results (Au, Buschkuhl, Duncan, & Jaeggi, 2016; Au et al., 2015a; Bogg & Lasecki, 2015; Soveri et al., 2017). Much of the extant research operates with an emphasis on identifying causal mechanisms that result in training behaviors (i.e., on-task learning and transfer), very little research is done to understand the predictability of the

performance of individuals on training task. Since WM is a system with limited capacity and since working memory training is geared towards improving individuals' WM capacities, it is important to understand if an individual is improving within the training task. The predictive modeling approach discussed earlier might be applicable to working memory training scenarios in that it can be used to predict who might be a good learner and who the training is working for. In the current work illustrates (i) whether predictive modeling can be utilized to predict the later performance from early training performance alone, and (ii) evaluate whether demographic information and training specific information can improve the predictions. In doing so, we evaluate the differential contribution of training performance metrics (performance on the WM task on any given session) and non-training performance metrics (participants' background information including demographics, training parameters) on predicting the later performance of individuals. Additionally, we evaluate the ways in which the data from the training studies can be mined for an improved understanding of the performance over the course of training to improve predictions of later learning.

## **2.2. Why predict working memory training?**

So far, the focus of the dissertation has been around the discussions of learning, personalization, difficulties associated with the process (specifically of predictive modeling approaches), and the challenges of the existing predictive models. The work discussed here ties into the broader scientific goals of understanding and evaluating (perhaps, eventually improving through personalization) working memory training. However, in the current work, I treat working memory as an example of skill-acquisition as I have discussed earlier. Therefore, first I will provide the rationale for using



predictive modeling within working memory training context. Working memory training involves repeated practice on a specific skill using a specific task (which is often increasingly difficult to adapt to the learner akin to increasing complexity of learning material). Working memory training data that I analyze here shares several similarities with online (and blended) learning –

- i. Use of a digital device: The data for WM training that I analyze here are collected using digital devices much like online learning. Machine learning driven predictive modeling (and personalization) are, typically applicable and feasible for digital environments since data analysis is driven by algorithms. Of course, it is possible to digitize data from non-digital sources before making predictions, however, this requires significant amounts of work and the predictions will not be readily available for automated data-driven personalization.
- ii. Evaluation: Both WM and Online learning involve some form of measure to evaluate the learners' performance. In case of WM training data, an individual who starts training at a certain level of WM capacity, undergoes training. Subsequently, a measure of performance can be attained by inspecting the performance level at the end of training, or as a difference between the initial level of performance and the final level of performance. In case of online learners, evaluation is conducted using assignments, quizzes, or any other form of metric that the instructor determines appropriate for the course. Typically, the measure of final performance, would be in terms of grades obtained by learners.
- iii. Individual differences: WM and online learning are influenced by many different factors and as a result, the performance differ across individuals. These

differences are characterized by magnitudes of differences in performance (within subject – growth and between subjects – e.g., high performance/better grades).

While there might be a few other shared similarities (e.g., both have the potential to influence life outcomes, albeit in different ways), these three similarities are critical for the ensuring that the current predictive modeling is applicable. Since, the goal is to predict later learning from early learning as measured from data generated using a digital platform using individual differences in learning, it is necessary to have these similarities. Note, that these similarities are broadly defined and there might be critical differences (for instance, online learning involves the use of internet, working memory training does not have to). These differences are not critical for the sake of predictive modeling. There are also a few critical differences, however, that make the predictive modeling more challenging in case of online learning scenarios.

- i. Knowledge: While, for the context of the current work, it is not necessary to differentiate the core knowledge structures (also referred to as mental models, schemas, or conceptual frameworks (Day, Arthur, & Gettman, 2001)) that determine WM and classroom learning, one fundamental difference is that learning in a classroom setting rely on specific subject knowledge (and related subject domains), WM training performance does not depend on subject knowledge.
- ii. Repetition: In a typical WM training protocol, the learners are required to practice the same task (sometimes at a higher or lower difficulty level) over the course of the training or learning period. On the other hand, in a typical online learning

context, the progression of course structure and the content being taught could be diverse and complex.

- iii. Motivation and Affective traits: Motivation of learning is a very complex topic and difficult to evaluate across various learning contexts. For instance, self-regulated learning is known to play a very crucial role in learning within online contexts (Kuo, Walker, Schroder, & Belland, 2014; Schunk & Zimmerman, 2012), but there is no empirical evidence that suggests that self-regulated learning influences WM training. Furthermore, typical WM training studies do not include motivation related measures with a few exceptions (for e.g., Katz, Jaeggi, Buschkuhl, Stegman, & Shah, 2014; Mawjee et al., 2017). However, the scales used to measure motivation, expectancies, and socio-emotional aspects of learning in online learning contexts are not used for valid reasons (motivation does not play a similar role in two contexts, learning in academic contexts carry high costs associated with failure and have direct impact on life choices such as career.)

Thus, it is reasonable to argue that WM training can be considered a context of learning where learning occurred shares similarities with online learning yet different since it is a less complex prediction problem (since subject knowledge and motivation and affective processes are not considered in WM training evaluation). In addition, since, WM training is sequential repetition of a training task spaced in time with varying difficult levels, it retains the overall structure (single or similar tasks used throughout the training process). Furthermore, the WM training task that I analyze involves training regimen that were assigned by the researchers and sometimes conducted in the lab with supervision. This is not the case in fully online learning settings, where most of the

learning habits are driven by the students' motivations and effort. Thus, it is reasonable to argue that WM training is a relatively easier prediction problem that does not need to account for changing motivations, knowledge requirements, or differences in lecture delivery (and differences induced by instructor). WM training data provides exceptional opportunity to evaluate the step-wise predictive modeling approach (and prior to using it) within a learning context that is considerably more complex. Therefore, I hypothesize that the predictions within WM training context will lead to comparably better results than in online learning.

### **2.3. Specific Research Questions:**

**RQ 1** – To what extent can we predict learners' performance on later WM training sessions from early WM training sessions, demographics, and training related information?

There are two subcomponents to RQ 1 that need to be addressed before we can fully answer "to what extent" we can predict learners' performance. By answering each of these subcomponents, we can determine the extent of predictions, the constraints, and the advantages.

- (a) How accurately can we predict learners into those who learn more (above median) or those who learn less (below median) in the later training sessions? How accurately can we predict learners' actual performance in the later training sessions?

This subcomponent pertains to determining if the data gathered for addressing RQ 1 are useful to classify or differentiate learners into groups/categories or if the data are useful for predicting the performance in the later sessions. To address this part of the

research question, we employed two different predictive models – classification models and regression models. Classification models are aimed at categorizing learners into those who learned more and those who learned less compared to the median. Regression models are aimed at predicting the performance levels of the learners. Each approach uses a range of different machine learning models which are discussed in the analytical approach.

- (b) What features are the most predictive of overall later learning? What are the relative predictive values of learning during the early WM training sessions, demographics, and training related information?

This subcomponent is aimed at determining which of the features are most predictive of later learning. To address this component, following the proposed framework, we created four models. Model 1a was the baseline model. This model was used to understand if the other three models are performing as expected (better than the baseline). Model 1b included training performance data of the learners. Model 1c included training performance as well as demographics. This model was used to estimate the predictive value of the non-malleable features available for the current dataset. Model 1d included training performance as well as demographics and training related information. Here, we chose to incorporate demographics and training related information to evaluate the predictive value of potentially malleable training related information. Note that these training related information are context specific (to WM training) and are not available for the other two datasets I evaluate later. For each model, we provided the predictive model with increasingly more information about the training performance to determine how much training data is necessary for making ‘good’ predictions.

Combined, these two steps address questions related to “How much information is necessary?”, “How soon can be predict?”, and “How well can we predict?” Note that given the nature of the step-wise inclusion of information, answer to one question will change in response to the other two questions. For instance, how well we can predict is a function of how soon we are hoping to make predictions and how much relevant information is available for making the predictions. We hypothesized that prediction accuracies during early learning are highest with all features included compared to models that only rely on training information alone (Model 1b performs worst early on). Later, once the models can reliably understand and learn from the learners training trends, we hypothesized that both models perform similarly.

## **2.4 Methods**

The analytical approaches that are discussed here, specifically, the machine learning based prediction model selection and predictive modeling approach are applicable to all three studies that are included in this dissertation. First, we describe the data source and participants. Next, we discuss the machine learning model selection process following the proposed framework. Then, we provide the list of features used for each model. Finally, we detail the analysis for the models.

**2.4.1. Dataset.** We used a cumulative dataset from an array of WM training studies that were mostly conducted in a lab setting (with a few exceptions listed in Table 2.1). These studies were selected as data sources since all of them used a single training protocol (N-back) with some differences as detailed in Table 2.1. We revisit N-back training data from several studies conducted either at our lab or collaborating labs. We included data

from 15 studies which were conducted between 2008 and 2019. Details of the studies are presented in Table 2.1.

Data from 739 participants (mean age  $\pm$  SD,  $24.73 \pm 21.11$ ; range 7-86) were included. Participants were included if they completed an adaptive N-back training over the course of at least 15 sessions. During the N-back task, the participants are presented with a sequence of stimuli (spatial, verbal, object or in combinations of these stimuli) one at a time. The participants were asked to decide if the current stimulus was the same as the one presented 'N' trials ago ( $N = \{1, 2, 3, \dots\}$ ). The higher the N-level, the more difficult the task is expected to be. Typically, an adaptive algorithm (such as a staircase method) is used to increase or decrease difficulty adjusting to participants' performance. Previous work showed that longer training periods lead to higher gains with significant performance on the gain scores after 2 weeks of training (Jaeggi et al., 2008). Thus, we excluded the participants who completed less than 15 sessions of N-back training from our final dataset. As noted earlier, machine learning models typically perform better when provided with more data. Thus, we restricted the current analysis to the first 15 sessions in order to maximize the available data for our machine learning models while retaining a theoretically meaningful level of training information. Note that around 40% of the participants completed more than 15 sessions of training. After

Table 2.1.

List of studies, sample and training details for the studies included.

Study	N	Population	Location	Length of training sessions	Training Context	C (Y/N)	Training Condition
Jaeggi et al. (2008). <i>PNAS</i>	15	Young adults	Bern, Switzerland	20 rounds	lab	Y	Dual N-back
Jaeggi et al. (2010). <i>Intelligence</i>	46	Young adults	Taipei, Taiwan	15 rounds	Lab (in group)	N	Dual/Single N-back (Spatial)
Jonides et al. (2010). <i>Presented at the Office of Naval Research Contractor's meeting, Arlington, VA</i>	22	Young adults	Ann Arbor, Michigan	20 rounds	Lab	Y	Dual N-back
Seidler et al. (2010) <i>Technical Report No. M-CASTL 2010-01, University of Michigan, Ann Arbor.</i>	18	Older adults	Ann Arbor, Michigan	20 rounds	Lab (not closed cubicles)	Y	Dual N-back
Jaeggi et al. (2011). <i>PNAS.</i>	32	Typically developing children	Ann Arbor & Detroit, Michigan	10 rounds	Home	N	Single N-back (Spatial)
Angera et al. (2012). <i>Behavioral Brain Research.</i>	29	Young adults	Ann Arbor, Michigan	20 rounds	Lab (not closed cubicles)	Y	Dual N-back



Jaeggi et al. (2014). <i>Memory and Cognition</i>	50	Young adults	Ann Arbor, Michigan	15 rounds		N	Dual/Single N-back (Auditory/verbal)
					Home		
Stpankova et al. (2014). <i>Developmental Psychology</i>	20	Older adults	Prague, Czech Republic	20 rounds	Home	Y	Single N-back (Verbal)
Zhang et al. (2014). <i>26th Annual Convention of the Association for Psychological Science, San Francisco, CA</i>	26	Young adults	College Park, Maryland	15 rounds	Home	Y	Single N-back (Object)
Katz et al. (under review)	55	Typically developing children	Ann Arbor, Michigan	20 rounds	School (large groups)	Y	Single N-back (Spatial)
Tsai et al. (in prep)	30	Adolescent	Irvine, California	15 rounds	School (small groups)	Y	Single N-back (Object)
Jones et al. (2019). <i>Journal of Attention Disorders</i>	39	Children with ADHD	Irvine, California & Ann Arbor, Michigan	15 rounds	Home	Y	Single N-back (Spatial)

Jaeggi et al. (in press) <i>Journal of Gerontology: Psychological Sciences</i> .	79	Older adults	Irvine, California & Ann Arbor, Michigan	20 rounds	Home	Y	Single N-back (Object)
Katz et al. (2018) <i>Learning and Memory</i>	36	Young adults	Ann Arbor, Michigan	20 rounds	Home	Y	Dual N-back
Pahor et al., (under review)	242	Young adults	Irvine, California & Riverside, California	20 rounds	Lab (not closed cubicles)	Y	Single N-back (Object)

---

*Note. C stands for Compensation.*

excluding those who did not complete at least 15 sessions of training (N=54) and 2 other participants who had data missing on various variables of interest, the final sample included 683 participants (mean age  $\pm$  SD, 22.51  $\pm$  16.11). The sample included more female participants than males (58% female). Overall, the N-back performance at baseline (i.e. average performance in the first three sessions) was 2.61  $\pm$  0.92 N-back levels (range = 1 to 6.25 N-back levels). The N-back final performance (i.e. average performance in the last three sessions) was 3.64  $\pm$  1.58 N-back levels (range = 1.04 to 9.05 N-back levels). A description of the sample population, demographics details, and training details can be found in Table 2.2. Specifically, we have provided the proportions of participants that were supervised during the training process (trained in a lab setting), participants from USA (given that our sample included data from more than one study – this was not used for the training purposes), Single N-back and Dual N-back proportions.

Table 2.2

*A description of the sample population, demographics details, and training details*

feature	Mean	SD	Proportion
Age (range 7-86 years)	25.37	18.54	
Gender (proportion female)			0.58
Supervised participants			0.50
USA based participants			0.82
Single N-back type			
Spatial			0.42

---

Verbal		0.10
Object		0.12
Dual N-back type		0.36
N-back level at Session 1	2.30	0.79
N-back level at Session 15	3.21	1.56
Change in performance (Session 15 – Session 1)	1.03	1.01

---

**2.4.2. Raw Data Collection and Data Preprocessing.** The flow of the data beginning from the preprocessing to final predictions are presented below. De-identified raw data were collected in secure files, processed, and analyzed using custom Python scripts. To achieve this, we collected raw data into csv files (separately for each study) and preprocessed the data by cleaning (e.g., by dropping the data from participants who may have missing data on the key variables described below).

**Defining the Class variable (P).** Next, we extracted the preliminary feature of interest – the outcome measure by categorizing the learners into “above” or “below” median performers based on the change in performance on the 15<sup>th</sup> training session<sup>2</sup> compared to the 1<sup>st</sup> training session (see below).

The next step in ‘Feature-extraction’ involved transforming data to generate features of interest that are expected to carry high predictive values. This step was conducted to encode the outcome variable as well as the features that are used to predict the outcome variable. The primary goal of feature extraction was to create a set

---

<sup>2</sup> A preliminary attempt to calculate the final performance of each individual based on how many sessions they completed (instead of making an arbitrary cut-off at the 15<sup>th</sup> session) showed very similar findings.

of features that were expected to influence the performance of the individuals on working memory training task. We selected the variables following two criteria: 1. the features were theoretically grounded in the literature. For instance, demographic details and training details alongside performance on the training task and other extrinsic features such as supervision, and compensation associated with participation are known to drive learning. 2. The variables selected carried high variance and low correlation with other variables. In other words, the dimensionality of the datasets was reduced by dropping features that had no mutually exclusive information. Specifically, based on previously mentioned meta-analyses (Au et al., 2015; Karbach & Verhaeghen, 2014; Melby-Lervåg & Hulme, 2013; Shipstead, Redick, & Engle, 2012; Wager & Smith, 2003), a set of features were extracted to be used as predictors. We included the following key variables for each individual participant:

**Performance trends.** The first goal of this study was to understand the value of on-task performance to understand learning without the presence of any other information to predict learning. The on-task performance act as true measures of intraindividual variability regardless of the cause for the variability. Thus, we use the raw performance as well as changes in performance in the first few sessions as measures of training trends.

**(a) Performance on each session.** The average performance of each individual from training session 1 through 14 (N-back levels) were used as measures of variability within task across the micro timespan of training.

**(b) changes in performance.** Changes in performance from session 1 to session 2 ( $t_2-t_1$ ); session 2 to session 3 ( $t_3-t_2$ ); session 1 to session 4 ( $t_4-t_1$ ) were

used as features. The changes in subsequent training sessions represent the changes in learning rate on the N-back task per session and capture the intraindividual variabilities for the first four sessions. Fast learners show larger learning rates in each subsequent training sessions. Although the learning rate can be calculated for each subsequent training session, we restricted this to the first four sessions since our interest was in predicting the later gains using minimal early training information. We added the cumulative learning rate for the first four sessions ( $t_4-t_1$ ) to account for the total learning attained during this period with a hypothesis that understanding the learning rate would improve prediction accuracies.

**Demographics.** Next, we used non-malleable learner level measures to understand the overall learning and their predictive values.

**(a) Age.** The age of participant (in years) was used as a feature based on some of the previous work that showed a positive predictive value (see Au et al., 2015; Jaeggi et al., 2014; Jaeggi, Karbach, & Strobach, 2017). Several other features that were derived based on age, including groups of age (children, young adults, older adults, etc.), decade of age (less than 10 years old, 10-20 years old, etc.) were tested. We included age in years as a feature since the other variables did not yield any further predictive value.

**(b) Gender.** Based on Au and his colleagues' meta-analytical work (2015), the performance of individuals did not differ by gender. Furthermore, a preliminary analysis using validation dataset yielded a low predictive value of gender. Thus, we ultimately excluded participants' gender as a feature.

**Training details.** As discussed earlier, learning is influenced by the context of learning, the content features, and other environmental factors. Thus, we used type of the trained stimulus and context of training as measures of learned activity.

**(a) Stimulus type.** Two key features were generated from the type of stimulus material used during training, a binomial dummy variable to represent whether the type of N-back training was single or dual (i.e. only one stream or two streams of stimuli), and the type of N-back stimuli (Spatial, Verbal, Object, and Dual). The feature N-back stimuli type did not yield additional predictive value beyond a simple binomial variable for single or dual N-back. Thus, we dropped type of N-back stimuli as a feature from the final modelling.

**Extrinsic features.** Based on the literature, extrinsic features are known to influence the performance of learners. Thus, we included two measures of extrinsic features that are known to influence WM training performance.

**(a) Supervision during training.** A dummy variable was created to represent supervision during training and used as a feature. Specifically, individuals who trained at home were labeled as unsupervised and those who trained in the lab setting were labeled as supervised with a hypothesis that learning about supervision during training might improve predictions of performance (Au et al., 2015).

**(b) Compensation.** A dummy variable was generated to represent whether the individual received compensation or not. Compensation is an extrinsic motivational factor that is hypothesized to increase the prediction accuracies since earlier studies have shown that being compensated leads to negative impact on performance.

The final steps of analysis included training, hyperparameter optimization (correcting for false positives and false negatives), post-processing (final model selection, defining and calculating performance metrics for the selected model), and then testing and evaluating the model on the test dataset.

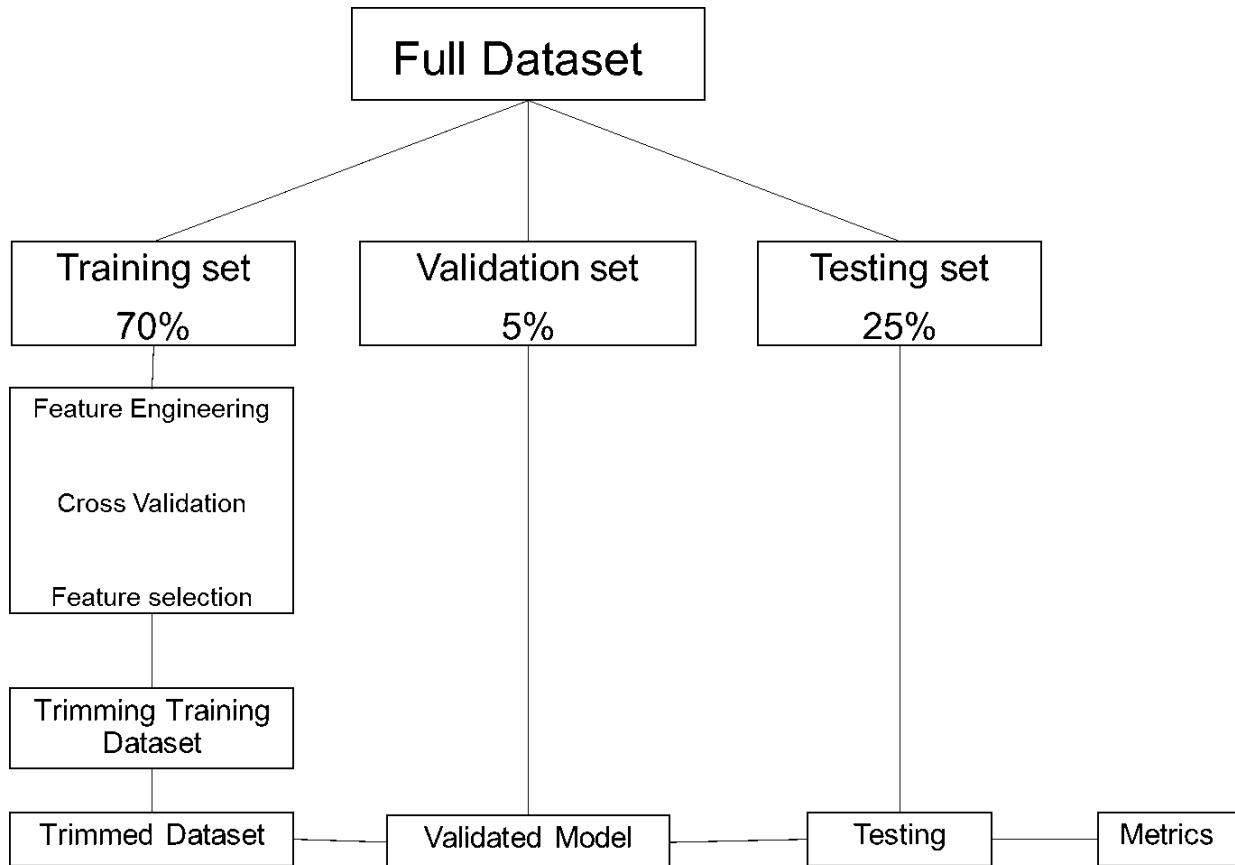
**2.4.3. Prediction task.** Following the framework introduced in the previous chapter, we establish our prediction task first. In the current study, we aim to predict the final performance of each individual using predictive models in two prediction tasks: (a) To demonstrate the utility of the proposed framework to predict WM training behavior, we attempted to predict the performance of each individual on the 15<sup>th</sup> training session. We attempted a classification model as a less challenging problem (than predicting the performance on 15<sup>th</sup> session) since this requires less data and less computational power. The outcome measure for each individual is a categorical “class” variable that divides the sample into two groups, one above median and the other below median based on their performance level on the 15<sup>th</sup> session. Specifically, individuals in the subclass that gain above or equal to median performance will be given a label "1" and the rest will be given a label "0" to create balanced classes. This ‘Class’ variable will then act as the outcome measure for the classification models. An outstanding classifier will have assigned a ‘predicted label’ that will correspond to the assigned ‘target label’. In other words, all below median gainers will receive a prediction label ‘0’ and all above median gainers will receive a prediction label ‘1’. (b) We attempted to predict the performance of each individual on the 15<sup>th</sup> training session on a continuous scale using regression models. Since the training performance are measured on a continuous



scale, no further manipulation was required for this prediction task unlike for the classification task that required artificial categorization.

**2.4.4. Machine Learning Algorithm selection.** Next, although algorithm selection is a vital part of the framework proposed, it is not treated as a separate step for two reasons. First, typically the model selection process depends on the prediction problem (classification vs regression models). Next, once the prediction problem is set, the selection of the model depends on the level of researcher knowledge in determining best fit for the given problem. For instance, researchers may choose either standard machine learning based predictive models, neural network-based models, or custom kernels to fit their own needs. Thus far, the three most popular classification models included Decision Trees, Support Vector Machines, and Logistic Classification models. The two most popular regression approaches included Linear Regressions and Regression Trees. While it is important to establish the value of custom algorithms to boost prediction accuracies, in the current work we focus on most popular approaches rather than pursuing custom approaches. This makes it easier to compare results better and to establish the limits of models that are easier to understand and use for researchers (as they are available off the shelf) with no computer science background (Jasny & Stone, 2017).

In general, the classification of individuals is done by projecting the selected features into a hyper-dimensional feature space which in turn is used to generate a hyperplane to classify the data into required categories, in this case two, 'above median'



*Figure 2.1: Data Splitting Protocol -- A simplified representation of data splitting protocol for our models. We repeated this process 200 times for each model to get a robust measure of accuracies over the 200 iterations to establish the variability of our predictions. Testing dataset was untouched during algorithm selection and tuning process. Cross-validation is only applicable to Study-3 that will be discussed later. Recursive feature elimination was used during the validation stage to evaluate relative values of features. Trimming the training dataset involves removing features that were found to be of little value for solving the prediction task.*

or 'below median' (I H Witten, Frank, & Hall, 2005). Since the data was labeled (as

above or below median), we used supervised ML algorithms for classification and followed the WEKA toolkit protocol (Ian H. Witten, Frank, & Hall, 2011) to compare different algorithms. We compared several popular classification supervised ML algorithms including Logistic Classifier, Support Vector Machines (SVMs), Random Forests, Boosted Trees, and Decision-Making Trees to identify the model with best performance. We used regularized classification methods that punish the models for being extremely incorrect in their predictions. For example, a general equation for the regularized logistic regression uses  $\hat{Y} = 1/(1+\exp(-y_i((w,x_i)+b)))$  instead of  $\max(0,1-y_i((w,x_i)+b))$  where  $\hat{Y}$  is the predicted label,  $y_i$  is the actual class for the  $i^{\text{th}}$  learner with  $x^i$  feature vector,  $w$  is the weight matrix associated with the  $x$  features, and  $b$  is the bias. The optimal models were selected based on a two-step performance criteria that is used to reduce generalization errors: a) the data was divided into 75% training dataset (used to train the models), 25% testing dataset (untouched during training in order to test the model efficacy). In addition, 5% of the data (part of training dataset) is used for validation of the selected model (see Figure 2.1). We used all of the most common metrics (listed below) on the untouched testing dataset to determine the robustness of the models. b) The data was divided at 200 unique seed locations (the point of reference for the division of the three subsets listed above) followed by repeated measurements of each of the metrics 200 times. This two-step approach ensured the predictions' robustness and generalizability to new datasets that are similarly structured (Guyon & Elisseeff, 2011).

There are many different metrics that are used to represent success of a predictive model. For instance, precision is a common metric for rule learning, information gain in decision trees, weighted accuracies for identification of subgroups. However, the most common metrics for evaluating the performance of a model on the testing dataset include Accuracy, F1-score, and area under ROC curves or AUC (Flach, 2003). However, these measures are determined by historic use of metrics within a given field. In the fields of LA and EDM, the most common metrics used are Accuracy and AUC. However, they are insufficient to fully characterize the performance of the machine learning models. In the current work, we use all common metrics in order to promote the idea of using all relevant metrics across settings to better understand the value of predictions.

Table 2.3

*A matrix of predicted and actual class from the predictions can be used to count the true positive (TP), false positive (FP), false negative (FN), and true negative (TN) which can then be used to calculate the performance metrics of the classification prediction models.*

		Actual Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

A confusion matrix (also referred to as a contingency matrix) is commonly used to evaluate the quality of models in a tabular form (Brownlee, 2016). Table 2.3 shows a typical representation of a confusion matrix that includes four components. The components include TP = true positive; TN = true negative; FP = false positive; FN =

false negative. Most models' performance is evaluated based on their successful identification of TP and TN groups (correctly identifying above median and below median learners) and the errors made (FN and FP cases).

Accuracy is the simplest and most intuitive metric derived from the confusion matrix. It counts the total number of correct cases detected across all classes and compares them with all cases. Simply put, it is a proportion of correctly detected cases to all cases. Thus, higher accuracy is better. However, it is an incomplete representation of the performance of the predictive model, especially if the classes are unbalanced (the class sizes are skewed) or if the heuristics that drive the predictions make the predictions extremely easy or hard. Furthermore, accuracies are also not a good measure when detecting either classes (in a binomial classification problem) takes priority. For instance, detecting positive cases of cancer are considered more important and negative cases in criminal cases are considered more important. The formula for calculating accuracy is provided below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision measures the extent of correctly identified positive cases. Precision is a good measure when the cost associated with false positive cases is very high. Recall measures the extent of performance of the model for all cases that were needed to be identified as positive. Recall is a valuable metric when the cost associated with false negatives is very high. However, as evident from their definition, they do not account for the success amongst the negative cases which, incidentally, are accounted by accuracies. The formulae for Precision and Recall are provided below.

$$Precision(P) = \frac{TP}{TP + FP}$$

$$Recall(R) = \frac{TP}{TP + FN}$$

In addition, F1-score is derived as a harmonic mean of Precision as well as Recall, which in turn makes better judgment of the performance of the model for incorrectly identified cases. The formula for F1-score is provided below. Note that 'P' stands for Precision and 'R' stands for Recall. Overall, F1-score acts a better metric in the presence of skewed class distributions or if understanding the incorrectly classified cases takes priority.

$$F1\_score = 2 * \frac{P * R}{P + R}$$

Finally, Area under the ROC curve (AUC) is used as performance metrics for the predictive models. AUC uses two metrics: Recall and False Positive Rate or FPR (FP/N). ROC curves are plotted using the False Positive Rate on the x-axis and Recall on the y-axis. Intuitively, the model that best performs is located simultaneously close to 0 on the x-axis (indicating a low FPR) and close to 100% on the y-axis (indicating a high Recall).

Next, in the regression models, we estimate the relationship between the target and explanatory variables by fitting a curve to the data points so that the distances between the curve and target data are minimized. Specifically, the goal of the regression models is to map  $\hat{Y} = f(X)$ , where  $\hat{Y}$  is the predicted approximate for the outcome measure Y. The function  $f(X)$  varies depending on the model selected. For instance, the general linear regression model for prediction solves the equation  $\hat{Y} = W^T X + b_0$  where  $W$  contains the weight vectors for the features and  $b_0$  is the bias that

can compensate for the offset in predictions. Similar to the logistic classification models, we regularized the models by minimizing the error functions (Lee, Lee, Abbeel, & Ng, 2006; Xu, Caramanis, & Mannor, 2009). Furthermore, optimization of the model was performed using the `fmin_slsqp.optimize` in the `scipy` python package that uses sequential least squares programming to minimize the least square errors (Engel, Mannor, & Meir, 2004).

Similar to classification models, there are many different metrics that can be used to evaluate the performance of the regression models. Besides the  $R^2$  and the adjusted  $R^2$  metrics, the measurements of errors are typically scale-dependent (e.g., Mean Squared Error – MSE; Root Mean Squared Error – RMSE; Mean Absolute Error – MAE...etc.). There are other less used error metrics such as errors based on relative errors, relative measures, scaled errors, percentage errors (Botchkarev, 2019). However, since within the context of our regression models used for predicting learning, we limit our metrics to the scaled measures since having the errors in the same scale as the outcome measures make it easier to compare performance across models. MSE is usually used as the loss function for regression models because of its differential property and large penalties on small errors due to squaring the errors. However, it is not on true scale as the outcome measure. RMSE and MAE both have the same scale as the target values but RMSE puts more penalties on large errors than MAE. Therefore, we used RMSE as a preferred metric instead of MAE. Furthermore,  $R^2$  reflects the percentage of variance explained by the model, but as it is prone to artificial inflation with the increasing number of explanatory variables. Therefore, we used adjusted  $R^2$

instead of  $R^2$  since our models have varying number of features. The general equations for the metrics are given below.

$$RMSE = \sqrt{\frac{1}{N} \sum (y - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

$$adjusted R^2 = 1 - \frac{N - 1}{N - k - 1} * (1 - R^2)$$

where  $N$  is the number of observations,  $k$  is the number of explanatory variables,  $y$  is the target value,  $\hat{y}$  is the predicted value, and  $\bar{y}$  is the mean target value.

#### **2.4.5. Models of interest.**

Following the proposed framework for predictive modeling, we identified four models of interest to solve the prediction tasks that will address our specific research question. As a brief reminder, the research question aims to understand the extent of predictive accuracy achievable from early training performance, age, and training details. This RQ consisted of two subcomponents specifically to evaluate the quality of predictions on a classification as well as a regression task and the identification of features that are most predictive of later learning. The four models identified will be used to solve the classification and regression tasks separately. Note that all four models were provided with increasingly more training session performance in order to make predictions of later learning. This provided a means to understand the three dimensions of interest across the classification and regression models (how much information necessary for



predictions? how early we can predict later learning? and what is the quality of predictions?)

- a. **Baseline model** - We derived a baseline model for predictions using dummy data that is generated using “Random” package in python. Essentially, we used a function within the package to derive a Gaussian distribution around the mean and standard deviation of each week’s performance of learners in the training data. Each learner was assigned a random number from the Gaussian distribution around the true mean and standard deviations. These artificially derived data were then used to train and test the baseline model performance.
- b. **Model with only performance trends** – Next, we used the true performance of the learners to train and test our models. We hypothesized that these models would perform better than the baseline models since models that predict later learning from true learning performance are expected to have a significant advantage.
- c. **Model with performance trends and age** – Next, we used features that are not malleable within the context of our learning setting. The only feature we identified that cannot be manipulated was age which was included in the model along with the training performance.
- d. **Models with performance trends, age, task details, and extrinsic features** – Finally, we used features that were malleable within the context of our learning setting. Specifically, alongside training performance and age, we included the three features identified – Stimulus type, Supervision, and Compensation – into

model d. An overview of all features used for each model are provided in Table 2.4.

Table 2.4.

*Models tested and list of features included in each model variant.*

<b>model</b>	<b>RQ 1 - model a (Baseline)</b>	<b>RQ 1 - model b</b>	<b>RQ 1 - model c</b>	<b>RQ 1 - model d</b>
<b>overview</b>	(random data generated using average and standard deviations of each session's training performance)	(Learners' performance on each training session)	(model a + age)	(model b + task details and extrinsic features)
<b>features</b>	random noise	average weekly lecture quiz performance	average weekly lecture quiz performance  Age	average weekly lecture quiz performance  Age  Stimulus type  Supervision  Compensation

## 2.5. Results

Results from the validation dataset tested on our best models identified multiple linear regression as the best regression model with a validation-RMSE of 0.21. Regression tree-based modeling yield very poor validation-RMSE in comparison (0.93).

On the other hand, within our classification models, both logistic classifier and random forests (classification tree-based models) did well with validation accuracies of 92% and

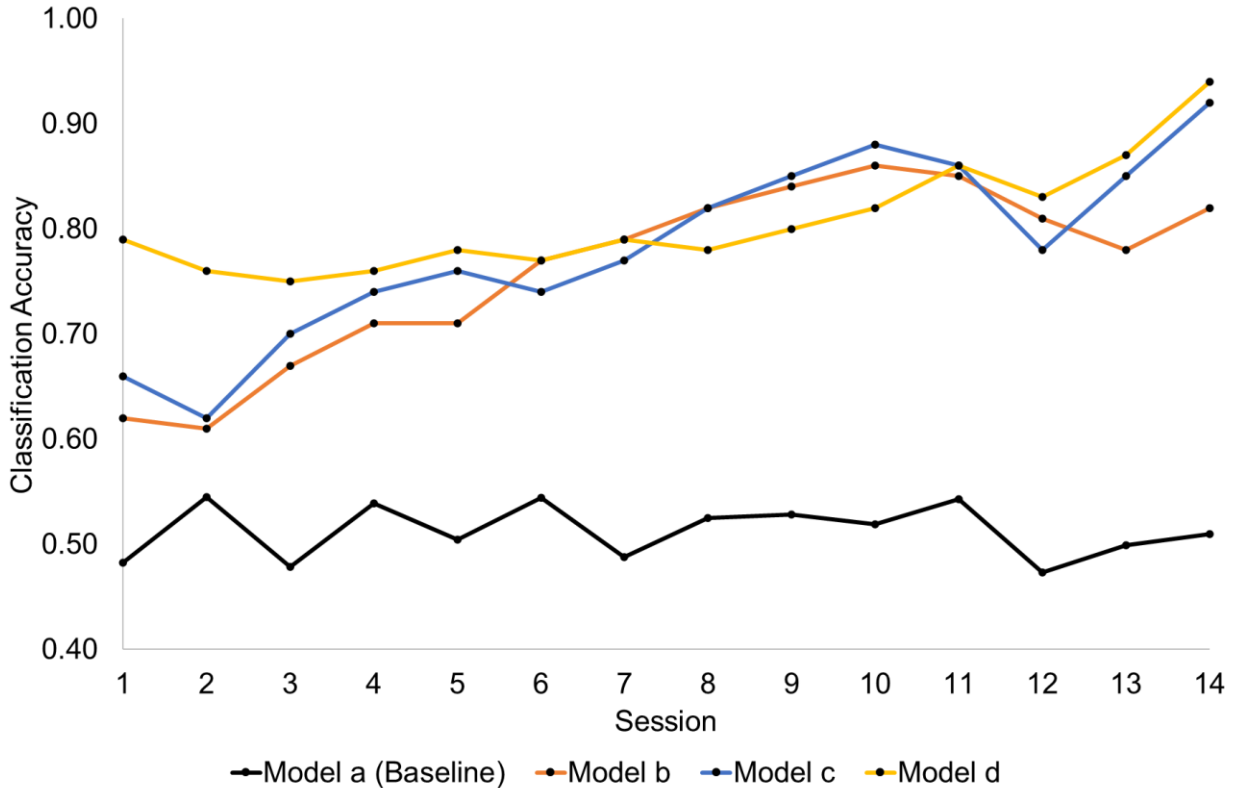


Figure 2.2: Prediction Accuracy with Logistic Regression -- Features included in Models a through d are listed in Table 2.4.

85% respectively. Training log loss and validation log loss (negative log-likelihood of the true labels) were 0.28 and 0.39 respectively for the logistic classifier which were relatively better compared to 0.53 and 0.77 achieved by random forest models. Thus, here we restrict our results to multiple linear regressions and logistic classification models to illustrate the utility of the framework proposed. Prediction performance for all 4 models are shown in Figures 2.2 and Figure 2.3 for classification and regression task, respectively.

Model a (Baseline model) is shown in black in both figures showing the least prediction performance across as expected. The average prediction accuracy for baseline classification model ranged between 0.55 and 0.45 around the expected chance prediction (0.50) given that our data classes were equally distributed. For the regression task, baseline model performed around an adjusted R-squared of 0.10, setting the lower bound for our prediction models of interest. Model b (which was trained

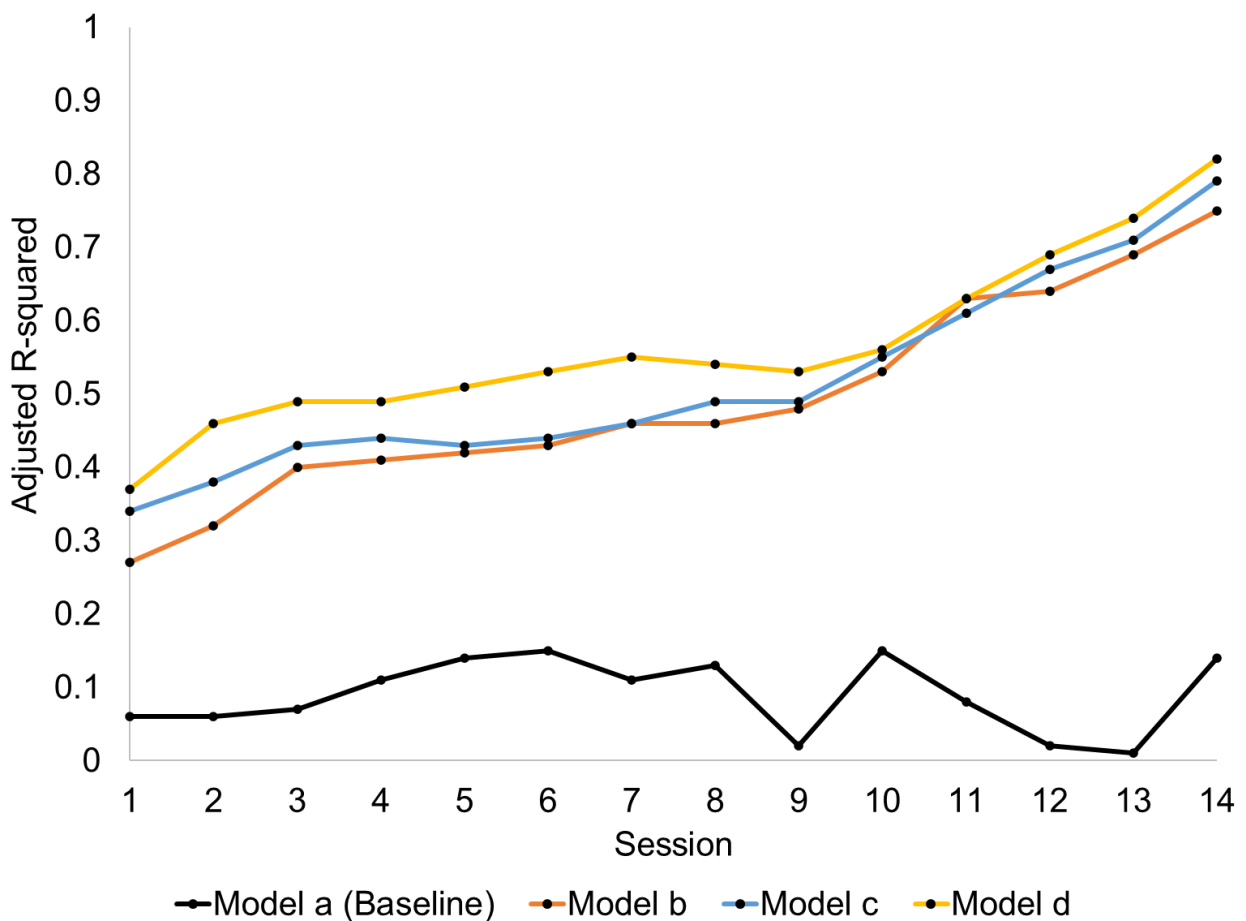


Figure 2.3: Adjusted R-squared -- different models predicting performance on 15<sup>th</sup> session of the WM training.

on performance data alone) is shown in orange in both figures. This model showed a prediction accuracy of 0.62 (14% gain over the baseline model) at the outset. The

prediction accuracies increased as more and more information of the learners' performance trends were incorporated into the model. Similarly, Model b outperformed Model a consistently across the board on the regression tasks. Model c (which included age as an additional feature) is shown in blue in both figures. This model showed a prediction accuracy of 0.72 at the outset with a steady increase overtime. Similarly, Model c quantitatively outperformed Model b in the regression task. Finally, Model d (which included stimulus type, compensation, and supervision as additional features) is shown in Yellow. This model showed a prediction accuracy of 80% at the outset on the classification task. Furthermore, Model d outperformed every other model on the regression task. The prediction accuracies peaked by session 10, followed by a relatively small change in classification model performance. On the other hand, the performance of the regression models did not indicate a plateauing. The overall training accuracy achieved with Model d across all training sessions 81% (+/- .05). The overall average precision for model was 83% (+/-0.03) and average recall was 85% (+/-0.03) with an F1-score of 0.84 (+/-0.04). In contrast, the average prediction accuracy for Models a, b, and c were 0.51 (0.03), 0.76 (0.08), and 0.78 (0.09) respectively. On the other hand, average performance for the regression task were 0.49 (0.14), 0.52 (0.13), and 0.57 (0.12) respectively for Models b, c, and d with corresponding average RMSE scores of 0.93 (0.54), 0.82 (0.46), and 0.68 (0.39). The fluctuations in the accuracies over the 200 iterations of the analysis with random seeding showed a prediction accuracy range of 0.67 – 0.90 for Model d. The overall prediction was significantly different from chance prediction ( $p < .001$ ) for all models at Session 1. The average ROC Area under the curve (AUC) for Model d was 0.91 (95% CI: 0.93 - 0.88) indicating that

the overall predictions are very good. In addition, the model evaluation indicated the features with most predictive value for Model d at Session 10 (session at which the model performed above average for the first time) were early performance (sessions 1-3), age, and stimulus type respectively.

## **2.6. Discussion.**

Our RQ aimed to understand the extent of predictive accuracy achievable from early training performance, age, and training details, and extrinsic features. Furthermore, this RQ was broken down into two subcomponents specifically to evaluate the quality of predictions on a classification as well as a regression task and the identification of features that are most predictive of later learning. Our models have shown that baseline models performed the least across both our prediction tasks. When two models (Model b and Model c) was trained, one with and one without demographic information, and provided with increasingly more training performance information, the model without demographic information started with a better than chance prediction accuracy (~0.60) and the model with demographic information performed slightly better at an accuracy of 66%. However, when the model was provided with details related to the training task (Stimulus type) alongside the extrinsic features (supervision and compensation), the performance was significantly larger at 0.80. However, this difference between the three models only continued to remain until session 6-8. This indicated that predictive models require an understanding of age, task details, and extrinsic features to perform well during the early learning phase. However, once the predictive models are able to learn from the actual performance of each individual over the early learning phase, the ability of the models to predict later learning is similar to

that of the other two models. This indicates that the models understanding of the variances within training performance fairly well and the non-training characteristics do not add any more predictive value at the group level. Furthermore, we were able to identify that early learning, age, and stimulus type had the highest predictive values for our best model at Session 10. Similarly, regression models performed better when provided with more data to learn from during the early phases. However, unlike in the classification models, convergence of the accuracies only occurred at Session 11. The results of our regression models show that our Model d can predict the amount of n-back gains of individuals with an average error rate of half an n-back level ( $\sim 0.68$  RMSE) at Session 10. However, these models that sought to predict the actual final performance of individuals on the 15<sup>th</sup> session did not have sufficient success rates during the early phases with relatively higher error rates for the first few sessions (1.34 - 1.69 RMSE). This indicates that the models are relying on the significantly more training performance data to determine how much change is expected in their learning on average based on training trajectories. However, these results are understandable given that the performance on regression tasks, are typically lower than on a binomial classification task due to the higher error rates associated with determining exact performance of each individual.

The results from our work show that Model d (on classification task) can reliably predict above and below median performers nearly 82 out of 100 times on average indicating that performance on training task can be predicted by including features that go beyond the training performance alone. The results support the findings from an earlier study showing that factors such as baseline abilities, age, and features such as

compensation and supervision could be related to individuals' learning. This was evident due to the sequential modeling approach we have taken following the proposed framework. Critically, this framework helped us illustrate the idea that more data leads to better prediction at the early learning phases, whereas less data that is more directly related to on-task performance such as historic performance, are sufficient to make reliable predictions. To answer the critical question of how soon we can make reliable predictions, the answer differs for each model. Here, we have shown that Models b, c, and d were able to achieve average performance at Sessions 6, 8, and 10 respectively. However, this answer would have made more sense if our models were consistently doing better with more training information provided. Unfortunately, as evident from Model d (on the classification task) starts out with a high prediction accuracy of 80% and continues to hover around 80% with very little change until Session 13. Thus, answering the question of how soon meaningful predictions are attained should be determined by more specific questions posed by the researchers or policy makers. For instance, if a researcher is interested in achieving, say at least 75% prediction accuracy, Models b and c will require at least 6 sessions worth of training information. Whereas, Model d can achieve this accuracy at Session 1. Thus, the researcher can determine the level of accuracy they are interested in and then determine what data are required to achieve this threshold and how early they want to intervene. Overall, the results from the current analysis shows promise in the ability to predict the individual's training outcomes at a relatively early phase during the training process that may allow for interventions (if any).



Interestingly enough, non-training related features add a significant boost to predictions when the training performance information of only 1-3 training sessions is provided to our models. However, there is a general decline dependence on non-training related features (age, stimulus type, compensation, and supervision) as more and more information related to the training is provided to the Models b and c. In general, by session 8, this convergence of model accuracies occurred indicating that once a sufficient learning related to training performance occurred, non-training related information has diminishing value of returns. However, this finding was only confined to our classification task indicating that the classification tasks' prediction accuracies reaches an upper limit sooner with more features and later without these features. On the other hand, the regression tasks' predictions show a continual improvement for all three models. Thus, in the absence of sufficient training dynamics related information, non-training related information boosts predictions. This behavior of our models fits well with the general notion that an individual's initial performance levels when encountering a novel task are often determined by their baseline ability, which are often a function of their age. However, once the advantages and disadvantages of age and baseline performance are accounted for, learning can be predicted more efficiently by understanding how well an individual is doing on the actual task. Furthermore, the upper bound of the classification task predictions are, perhaps, linked to the extrinsic features that we included in model d, given that the predictions reached ceiling values very early and stayed relatively stable.

Overall, we have utilized the stepwise framework described in the previous chapter to establish the value of such an approach in determining how much information

is needed, how soon we can make predictions above a threshold determined by researchers, and determined the lower and upper bound estimates for the predictive modeling within the WM training context. The stepwise modeling approach is valuable in its flexibility that it offers to researchers in making decisions related to what features are critical to their own context and what level of thresholds they are interested in. Furthermore, fully reporting the metrics available also makes it easier for comparisons of model performances across learning contexts. In our data, the accuracies, precisions, and F1-scores were closely related to each other for the classification models because of the artificially created balanced class. However, in many real-life scenarios (for instance letter grade distributions within an online classroom setting), the distributions of the classes are not necessarily balanced. Thus, non-accuracy metrics could be more valuable.

## **2.7. Limitations**

The current results have limitations, specifically, since the data comes from different studies, with different populations who are tested and trained under slightly different conditions. This may yield differing training performance across the studies. For instance, the number of participants from Jaeggi et al. (2008) included in Dataset-1 are only 15 whereas those from Pahor et al., (in prep) are 242. In theory, If these two datasets follow different Gaussian distributions of performances, the predictive models will lead to predictions that are biased against the dataset with lower sample size since the probability of a learner being a part of the latter study is more likely than a learner being part of former study. Thus, the learning achieved by our models that are used to predict future learning might be driven by those participants that were part of Pahor and

her colleagues' study. Furthermore, the performance of participants within each of the studies might follow their own distributions which may make it difficult to predict the actual end-of-the-training performance. In theory, we can circumvent this limitation by using "study" as a feature that uniquely identifies each of those subsets within our data. However, we chose not to do this to limit our discussions on the value of our stepwise framework. Furthermore, using 'study' as a feature acts as a proxy for the differences in the study settings, which is a latent variable that does not add further theoretical value or a means to explain why this feature boosted our predictions and the differential effect of this latent variable on each study. Furthermore, other limitations such as unequal representation of age (in particular, missing the 40-60 age range) are present in our data. Furthermore, our classification models depended on an arbitrary binomial classification of data (a median split) while the training gains are on a continuous distribution. The additional problem with this approach is that our model might be struggling with determining the subtle difference in performance of those individuals that eventually gained 0.85 (labeled as below median) vs 0.87 (labeled as above median). This might have led to a prediction accuracy cap of 88% of our best models. Furthermore, our models are missing information that might be critical for the interpretation of the results, for example, an individual's level of engagement in training, which were not included in our data. We posit that inclusion of these missing features may have potentially explained low probability predictions of certain individuals (non-confident predictions where class labels were assigned at a near chance level). Furthermore, while our method retrospectively demonstrates that machine learning

based models can predict future learning, it is not a real-time application of such modeling to identify the individuals who might need the most support to succeed.

Nonetheless, the current analysis is a crucial step to understand the fate of the participants depending on their individual differences in training dynamics. Further research is required to identify and understand how crucial training relevant parameters (that are directly mutable) such as training difficulty level, speed, length of training per day, total training period, target sensory modality for training material (e.g., auditory and spatial), implementation of motivational features (such as adding a storyline, themes, and unlocking achievements) are required to fully tailor training experience based on the individual needs. With an improved understanding of the features that lead to better training quality at an individual level we may expect more pronounced transfer effects. Furthermore, the ability to predict learning on a WM training task, creates an opportunity to intervene as needed to improve training outcomes in real-time if a robust prediction model can parse individual variances at the earliest possible time. Specifically, the extraction of features that promote training efficacy could ultimately contribute to development and implementation of personalized approaches that further enhance training success. In theory, a successful WM training intervention will account for individual differences and personalize training to match the needs of individuals. Such successful learning on-task of every individual within each study may then lead to broader learning consistent with theoretical expectations. However, these discussions are beyond the scope of the current thesis.

## **2.8. Conclusions**

Overall, our work demonstrated that predictive modeling approaches are good at identifying good and poor learners at an early stage during the training. We have also demonstrated that, while early predictions rely on participants' age, stimulus type, compensation, and supervision, these features added diminishing returns as the training progresses and more and more training related information is available to our models. Consistent application of data mining and machine learning framework can be used to compare results across models as well as setting. One must be cautious when applying machine learning modeling, however, since the results obtained are specific to the context and dataset. Once researchers come together to promote availability and use of large-scale open source datasets, results of these predictive models can be replicated, improved upon, and applied in real-time applications. However, this would require a comprehensive understanding of factors that affect training performances. Thus, we posit that solving the problem of enhancing cognitive capacities should be tackled using both explanatory modeling approaches as well as predictive modeling approaches.

### **CHAPTER 3: FROM INTENTIONS TO ACTIONS: UNDERSTANDING STUDENTS' SELF-REPORTED STUDY PLANS, ADHERENCE TO STUDY PLANS, CLICK BEHAVIORS, AND THEIR RELATION TO LEARNING OUTCOMES**

As a summary of the previous chapter, I have demonstrated that predictive modeling can be utilized to make predictions of later learning from early learning in a WM training dataset using the proposed framework. This framework enabled deriving insights that may not be readily available without stepwise inclusion of features. The primary goal of the current chapter is built upon the results from the last chapter. In Chapter 2, I have explored the usage of predictive modeling to evaluate a learning scenario in order to understand how predictive modeling can benefit from step-wise inclusion of features, specifically, to predict later learning, and what to expect from such an approach in short-burst time spans. Specifically, I explored if within a short-burst time span of learning, predictive models improve as more and more information is made available. Furthermore, I also investigated whether predictive models need any information about the learner or the learning context at all, in making robust predictions. The results show that the models employed were able to approach near 85% prediction accuracies by session 7 of the WM training. Furthermore, the results also show that the models are able to learn better if age and features related to learning context (and any relevant data beyond the training data itself) are provided during the earlier phases of learning. However, this advantage from additional data beyond the actual learning itself, showed diminishing returns. This is perhaps expected given the definition of later learning (learning on 15<sup>th</sup> WM training session), is quite narrow and highly correlated

with performances of sessions that are immediately closer to sessions 15 (say sessions 13 and 14) than the training sessions that occurred earlier (sessions 1, 2, and 3). However, within a more complex learning environment such as in a typical classroom experience, performance on later learning, such as in case of a final exam, the relationship of early learning and later learning are not simply a function of chronological co-occurrence. Rather, performance in final exam (and the final grade the students receive) are driven by factors such as students' knowledge, learning practices, and motivations. Thus, let us consider applying the predictive model to real-world scenarios moving forward (i.e., a blended learning environment and a fully online learning environment) to replicate the results from Chapter 2 and include relevant features for each of these scenarios. While doing so, in addition to the demographics and performances of the students, let us establish the value of self-directed learning of students, in hopes of making quantitatively better predictions, since students' learning intentions and the actions they take to accomplish the learning intentions are known to influence learning with classroom settings. The key point I explore in this chapter is the idea that understanding the learners' self-reported study plans lead to better predictions of later learning.

In addition, I use two features, the overall quantity of the click behaviors and the frequency of the click behaviors to specific assignments/tasks, to differentiate the learners' grades beyond demographics and early learning quality. A few studies that will be discussed later use predictive modeling in hybrid classrooms to investigate the idea of looking towards click behaviors and autogenerated interaction logs within LMS to make predictions of later learning and dropout rates. Specifically, blended learning

classrooms (those that include face-to-face as well as online learning components), have demonstrated a positive association with learning outcomes compared to fully online learning scenarios. Blended learning incorporates support structures and guidance from instructors directly since critical aspects of learning such as lectures and discusses are held face-to-face. Since, meetings and lectures held via remote instruction tools and technology lack engagement or teacher-directed regulation of learning, blended learning has been touted as the best possible middle ground for bridging the advantages of both learning scenarios (offline and online). However, key factors that influence learning, specifically, self-regulated learning are as applicable to blended learning environments as they are to any other classroom learning experiences. Specifically, in the current chapter, I focus on students' intentions of spacing their work and how, if at all, click behaviors within LMS can capture the adherence of students to such intentions.

### **3.1 Learning in Blended learning environments**

The term blended learning defined earlier as any course that involves 30% of academic activities via internet and 21% of the content is taught face-to-face. Blended learning has been pitched as a middle-ground between fully face-to-face learning environments and full online-learning environments, sharing the advantages of both. In fact, blended learning has been one of the top ten trending technologies in the knowledge delivery industry (Boelens, Van Laer, De Wever, & Elen, 2015; Graham, 2006). Blended learning offers two significant advantages over face-to-face learning by providing increased accessibility as well as flexibility and reduced cost. However, given the nature of “blending” two different approaches of learning without a particular



standard or model, blended learning has taken up a few different approaches which are broadly classified into one of the four groups: activity level blending, course level blending, program level blending, and institution level blending. As evident from the naming scheme, each of these variants use approaches to blend online learning into the traditional face-to-face classroom as a function of instructors practices (activity level and course level) or as a function of policies determined at the program or institution levels. Each of these blending approaches could differ in their goals, including for the purposes of enabling learning (for access and convenience), enhancing learning (incremental changes to instructional practices), and transforming learning (learners transform from passive listeners to active constructors of learning). Specifically, factors such as access to knowledge, social interactions, personal agency and study-habits...etc., are known to play a critical role in blended learning environments (Nickel & Overbaugh, 2012; So & Brush, 2008). When it comes to predictive models, learning analytics are employed to examine the frequency of students' engagement with the learning material in relation to the quality and quantity of learning of individuals. The existing literature that seeks to predict behavior in blended learning environments, typically, utilize attributes such as skill level and demographics, participation in online course activities , student engagement with optional online forums, click behavior, and time spent on resources as markers of learning (Anderson, Huttenlocher, Kleinberg, & Leskovec, 2014; Bayer, Bydzovská, & Géryk, 2012; Chaturvedi, Goldwasser, & Daumé Iii, 2014; Guo & Reinecke, 2014; Hershkovitz, Baker, Gowda, & Corbett, 2013; Huang, Dasgupta, Ghosh, Manning, & Sanders, 2014; Ramesh, Goldwasser, Huang, Daume, & Getoor, 2014; Seaton, Bergner, Chuang, Mitros, & Pritchard, 2014; Wilkowski, Deutsch, &

Russell, 2014). Many of the existing models, often, do not take individuals' motivations and study habits into account.

Demographics are often used as a proxy measure for general abilities and cognitive skill levels of each individual within the existing models. However, an individual's motivation is known to play a key role in learning, more so in the online learning context, due to the sustained need for self-efficacy throughout the learning phase (Chen & Jang, 2010; Lim, 2004; Miller, Deci, & Ryan, 1988). Bjork (2017) recently reviewed the cognitive theories of learning that pertains to spacing and learning which informed recent work (Jeffrey and Roediger, 2008; Hartwig & Dunlosky, 2012) which showed that students' intentions to spacing or cramming have a positive effect on the quality of learning. The higher the spacing, the more forgetting that is expected to occur, and better memory formations that attribute towards greater long-term learning retention.

One such behavior that was recently explored by Rodriguez et al. (2019) is students' study intentions pertaining to a STEM course in a blended learning environment. Specifically, this work looked at students' spacing intentions to understand the performance of the students -- spacing (splitting study sessions across multiple days) and cramming (doing most of the studying on the day before the exams). Rodriguez and his colleagues utilized a combination of self-reported study intentions ('spacing' or 'cramming') and students' clickstream data that reflects an individual's engagement with course material to understand learning and how intentions might reflect in students' performance in a course. The results of this work showed that students' self-reported study patterns, i.e., self-reported intentions and implementation

of spacing strategies, showed significantly better learning than those students who crammed during the learning period. Results from this study suggested that those who claimed to have maintained spacing across the learning period showed better grades overall. However, this work did not explore the potential differences between individuals that reported maintaining spacing but did not get good grades. This limitation relevant to differentiating the learners who did and did not receive a good grade despite intentions of spacing is due to the lack of clear connection of students' intentions to adherence to such intentions at the individual level.

Rodriguez and colleagues used a combination of students' pre and post survey responses to categorize the students into those who 'maintained cramming', 'stopped spacing', 'started spacing', and 'maintained spacing' (e.g., students that reported cramming before and after the course are categorized as maintained cramming) and restricted their analyses to those who 'maintained spacing' and those who 'maintained cramming.' In doing so, however, this approach overlooked examining students' initial intentions—what their study intentions were prior to studying for the course.

It is important to account for adherence and implementation of spacing intentions for a richer and more detailed understanding of behavior and how it relates to learning. For example, what differentiates those students who identified as spaced learners during the pre-survey that followed through and those who did not? Furthermore, a few questions are not fully explored in their work, such as - What demographics, if any, have a problem with implementing the spacing intentions from the pre-survey? Which individuals failed to succeed despite implementing the spacing of the learning? How early can one detect the students' performance trajectories based on the click-behavior

and intended study strategies. Perhaps, an answer to these questions might be revealed by an approach that not only looks at students' self-reported study intentions, but parses out the nature of behavior of those students who intended to space their learning (i.e., did those students adhere to their intended study plans).

### **3.2 Specific Research Questions.**

Gollwitzer (1999) did an extensive amount of research showing the importance of understanding not just the self-reported intentions towards learning, but also to what extent each student is diligent to accomplishing these intentions. This ties into the Bandura's self-efficacy theory (1982) - the ability to create realistic goals based on their understanding of their own capabilities and the diligence necessary to accomplish those tasks. To the best of my knowledge, there is no existing work that utilizes individuals' self-reported intentions and planned behavior to predict students' performances at the individual level. This area can benefit from utilizing predictive models that understands the extent to which students' intentions and ability to adhere to the study intentions predicts learning. Furthermore, it is important to understand if the intentions, and implementation can predict outcomes of later learning during the early phases of the learning (i.e., as early as the first three weeks) that provides a window of opportunity for the teachers to make adaptive changes to the curriculum and the pedagogy. To fill these gaps in the literature, the current work aims to address two specific research questions (RQ) -

**RQ 1** – To what extent can we predict students' performance on review quizzes and final grade using (a) reading and weekly quiz scores, (b) demographics and study intentions?

**RQ 2** – Can click behaviors improve predictions of models from RQ 1? If so, which of the click behavior features has the highest predictive value?

Addressing these RQs have several purposes. First, since we seek to validate our results from the earlier study, our focus in RQ 1 is to understand if we can predict later learning from early learning trends on a weekly basis as well as use the weekly trends to predict the final grades. As before, we will use two different models, one without (RQ 1 Model a) and one with demographics (RQ 1 Model b), to understand if demographics have predictive value beyond actual performances in the course. In addition, we included students' study intentions along with demographics as these intentions are innate to the students and are not mutable *ex ante*. We hypothesize that demographics and study intentions carry high predictive value during early learning predictions, but later learning can be predicted from early learning trends alone (i.e., without the need for demographic information or study intentions) since these features carried diminishing returns for predictions in the previous study. Next, our focus in RQ 2 is to understand if click behaviors can be used to boost our predictions from RQ 1. In this model, we included students' clicks behaviors that hoped to measure the intentions of study spacing along with all features in RQ 1 Model b (i.e., model which includes demographics, past performances, and in addition, measures of click behavior.) We hypothesize that click behaviors have very little value in prediction accuracies since the measures of clicks, as we have argued earlier, do not provide any context to students' learning nor do they act as a good behavioral indicator of students' changing motivational, social-emotional, and affective needs and demands. Following our framework, we evaluate the relative predictive value of each of these predictive models.

Specifically using this stepwise approach, our final models are going to include all the features of interest, which provides an opportunity to understand their relative importance in predicting their learning on a weekly basis and in predicting their overall grades. A summary of all models tested are presented in Table 3.1.

Table 3.1

*List of models tested, and features used in each model variant.*

model	Baseline model	RQ 1 - model a	RQ 1 - model b	RQ 2
overview	(random data generated using average quiz scores and standard deviations)	(students' performance on weekly quizzes)	(RQ 1 model a + demographics)	(RQ 1 model b + spacing intentions and click behaviors)
features	random noise	average of pre-lecture quiz performances  average of homework performances	average of pre-lecture quiz performances  average of homework performances  age gender low-income status part-time status minority status first-generation status high school GPA	average of pre-lecture quiz performances  average of homework performances  age gender low-income status part-time status minority status first-generation status high school GPA

SAT scores

SAT scores

study spacing  
intention

study spacing intention

change in study plan

change in study plan

click-data total course  
activity till date

click-data frequency of  
course activity per quiz

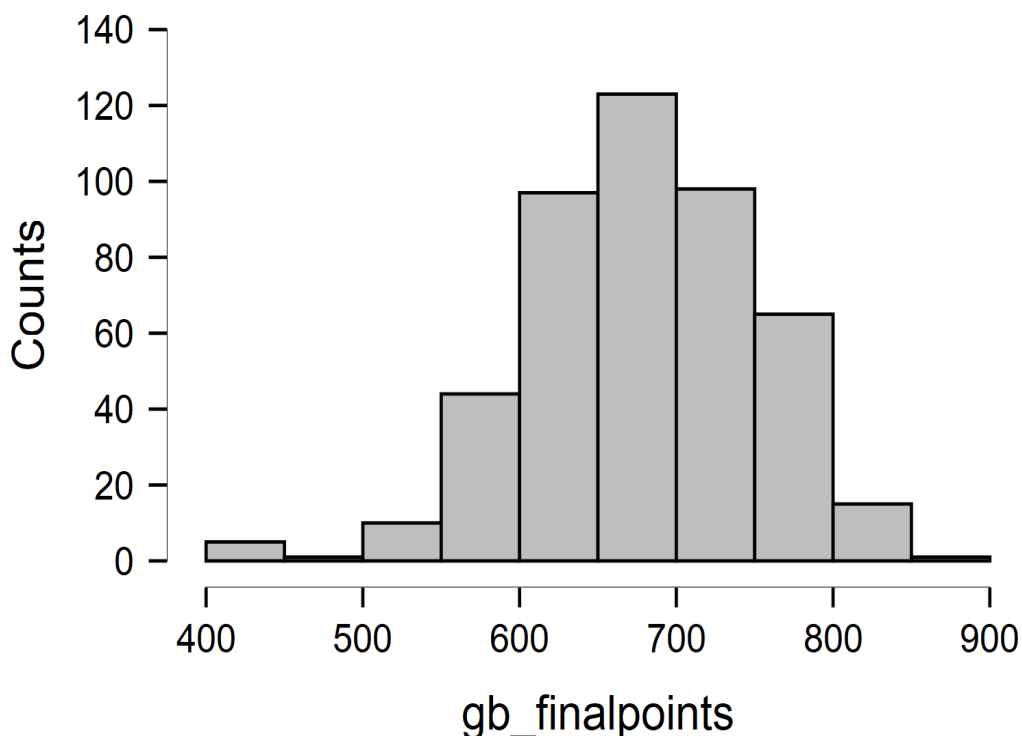
click-data spacing

---



### 3.3. Methods

**Dataset (face-to-face lower division Molecular Biology Course).** We analyzed learning behavior occurred in a 10-week long face-to-face lower division Molecular Biology course implemented at University of California, Irvine during two separate years -- in 2016 (course period of 74 days total) and 2017 (course period of 73 days total). Data were collected over two consecutive years with consistent format, syllabus, and method of instruction (Rodriguez et al., 2019). This course consisted mostly sophomore students and is structured in a way that nearly 50 percent each of the course period was dedicated to lectures and interactive peer group work. This course had three cumulative midterm examinations, one cumulative final exam, pre-lecture quizzes, and weekly



*Figure 3.1.* Grade point distributions of the students (shown as `gb_finalpoints` on the x-axis) -- Scores are cumulative for all individual quizzes, midterms, and final exams.

homework assignments which were conducted via Canvas Learning Management System (Canvas LMS). All of the course material, including the lecture videos, and slides are made available for the students via course’s Canvas LMS.

**Participants.** Year 1 consisted 224 students and Year 2 consisted of 422 students. However, since the predictive models required the students’ responses on the survey data (used to assess study strategies – such as spacing), we excluded students who did not complete the surveys. The final dataset consisted of 132 students from Year 1 (58.92% response rate) and 327 students from Year 2 (77.48% response rate). There were significantly more female students (n=137, p<0.05) in the year-1 dataset compared to male students. There were no statistically significant differences in other demographic features of the data. Therefore, we combined the data from both years for our predictive analysis. A detailed description of Students’ demographics data is provided in Table 3.2.

Table 3.2  
*Descriptive statistics of students’ demographic information*

<b>Feature</b>	<b>Mean</b>	<b>SD</b>
	20.51	1.50
Age (range 19-29)		
College GPA	3.22	0.46
High School GPA	3.41	1.51
SAT Score	1586.23	597.44
	<b><i>proportion</i></b>	
Low Income	0.37	
First Generation	0.49	
Relevant Major	0.75	
<b><i>Gender</i></b>		
female	0.61	

---

male	0.39
<b><i>Ethnicity</i></b>	
Asian	0.52
Latino	0.21
White	0.07
International	0.06
Other	0.04

---

Clickstream data were collected from Canvas LMS for all the students in the dataset. LMS data included time-stamped records of all the clicks made by each of the students, and the web URL specific to each click while using the LMS. Institutional records for all students are collected to identify underrepresented minority status (UMS - African American, Native American, or Latino/a).

**Prediction Task.** Students' final letter grade in the course was used as the outcome measure for our classification task, whereas their final exam score was used for regression task. To avoid issues with prediction biases, the outcome measure was simplified to a binomial prediction problem to differentiate learners that received "At least a 'B-' grade" and "Below 'B-' grade". Equal distribution of the outcome measure is a critical step to avoid overfitting of prediction models, and reduction of false positive predictions due to the differences in the grade distributions of the course (for example, if 70% of students received a grade worse than A, the prediction model would have a 70% accuracy if it assigns a grade lower than A to all individuals within the dataset leading to 100% false negative rate for A grades skewing the predictions). Overall, 117 students achieved at least a 'B-' grade, whereas 107 students received 'Less than B-' grade. The final grade point distributions are shown in Figure 3.1. The distribution of the

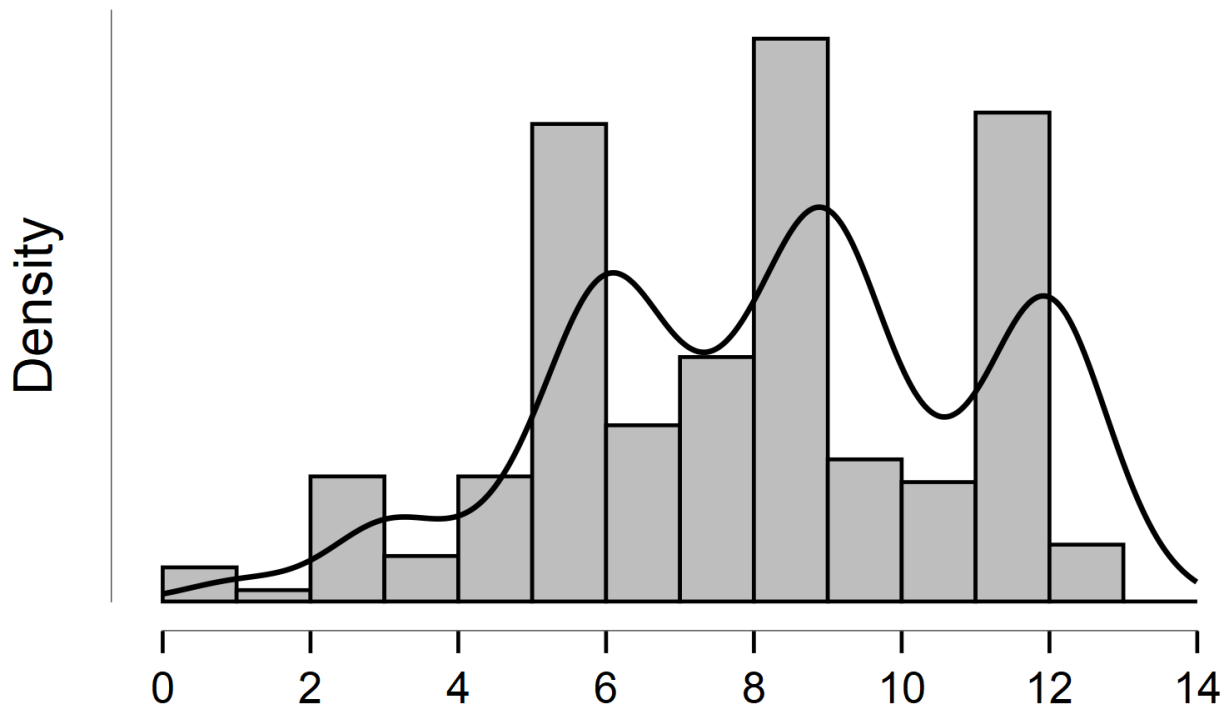


Figure 3.2. Final letter grade distributions of the students -- Numerical values corresponding to letter grades are as follows: 13 = A+, 12 = A, 11 = A-.....0 = F

corresponding final letter grades is presented in Figure 3.2. Additionally, we used students' performance on the in-class comprehensive exams that were conducted during week 3, 6, and 9 and the final exam conducted in the final's week (see Table 3.3 for descriptive statistics) with a median split (for classification task) or a continuous scale (for regression task) so as to mimic the session by session performances predicted in the WM training context earlier. Note that we did not seek to predict the weekly homework assignments or pre-lecture quizzes since their variability was very low rendering them biased prediction tasks.

Table 3.3  
Descriptive statistics of class examination information

Characteristics	Mean	SD
Quiz 1 (range 0-80)	51.69	13.27

Quiz 2 (range 0-80)	52.11	13.22
Quiz 3 (range 0-80)	51.72	15.13
Final Exam (range 0-100)	58.19	12.32

---

### 3.4. Measures

**Demographic variables.** Demographic variables were collected from the UCI's institutional records once the course was completed (after post-test). We collected features such as age, gender, low-income status, part-time status, first-generation status, minority-status, SAT scores, and high school GPAs.

**Grades.** Students' performances were recorded from all the graded assignments which were provided by the instructor. These included pre-lecture quizzes (29 which were conducted before a lecture began - conducted online), take-away homework assignments (9 which students complete on their own). Following the protocol from Study-1, we used scores of all quizzes and homework assignments leading up to weeks 3, 6, and 9 review quizzes for predicting performances of the students on comprehensive quizzes. All grades were collected ex post.

**Click data – total course activity per day.** Students' total clicks per day were measured by summing up the total clicks each student made on the Canvas course space.

**Click data – frequency of course activity per quiz.** We also measured the frequency of clicks per assignment by counting the total number of clicks a student made on each web page specific to the quizzes.

**Click data – spacing.** We measured an estimate of "spacing" of each student's clicks by measuring the total number of clicks made each day of the week for each corresponding quiz. This acted as a proxy objective measure of adherence to spacing

intentions (clicks are spread throughout the week before taking up the assignment) or cramming (clicks are focused just the day before the assignment). We also included an alternate measure of spacing, by calculating the days between the assignment due date and first attempt to submitting that assignment. This measure also acted as a proxy for procrastination behaviors since, procrastinators (planned or otherwise) tend to submit the assignments in the last minute/hour (McPartlan, 2020).

Study-2 followed a similar protocol for data analysis as discussed in Study-1. Due to the nature of the target variables discussed earlier, we used: 1) regression models to predict students' comprehensive quiz performances and final scores, and 2) classification models to classify students as "above" or "below" median performers for the comprehensive quiz performances and "above B-" or "below B-" for students' final grades. Regression models estimate the relationship between the target and explanatory variables by fitting a curve to the data points so that the distances between the curve and target data are minimized. Specifically, we started with a multiple linear regression model and compared the results with a 2<sup>nd</sup> order polynomial regression to account for variable interactions. In addition, we also tested the performance of regression trees given that our models are more complex compared to those in the previous chapter. Next, classification models were used to project the selected features into a hyper-dimensional feature space, generating a hyperplane to classify the data into required categories of interest. We tested two binomial classification models - Logistic Classifier (LC) and Random Forests (RFs) following the WEKA toolkit protocol discussed earlier to compare these algorithms (Ian H Witten, Frank, & Hall, 2011). The optimal regression/classification models were selected based on a two-step

performance criteria: a) data splitting into three groups (70% training + 5% validation + 25% testing) and b) repeated training and testing after data splitting using 200 unique seed locations for the best algorithm selected. We used the average of several metrics described earlier (i.e., RMSE, adjusted  $R^2$  for regression models and Accuracy, Precision, Recall, F1-score, and AUC for classification models) on the testing set to determine the robustness of the models. This two-step approach, consistent with Study-1 was used to ensure the predictions' robustness and to minimize generalization errors. (Igyon & Elisseeff, 2003; Tang, Alelyani, & Liu, 2014). Using the validation dataset, we determined that Random Forest model outperformed Logistic classification model significantly (validation accuracies of 83% and 71% respectively), whereas Multiple

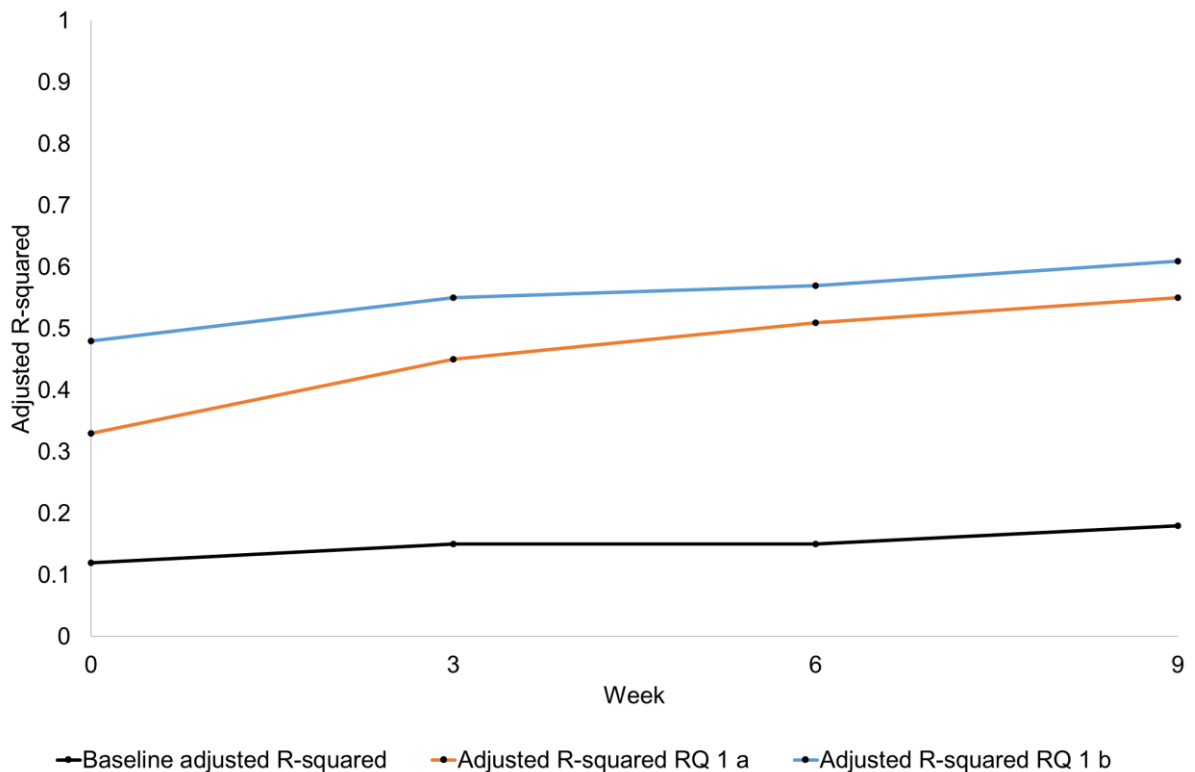
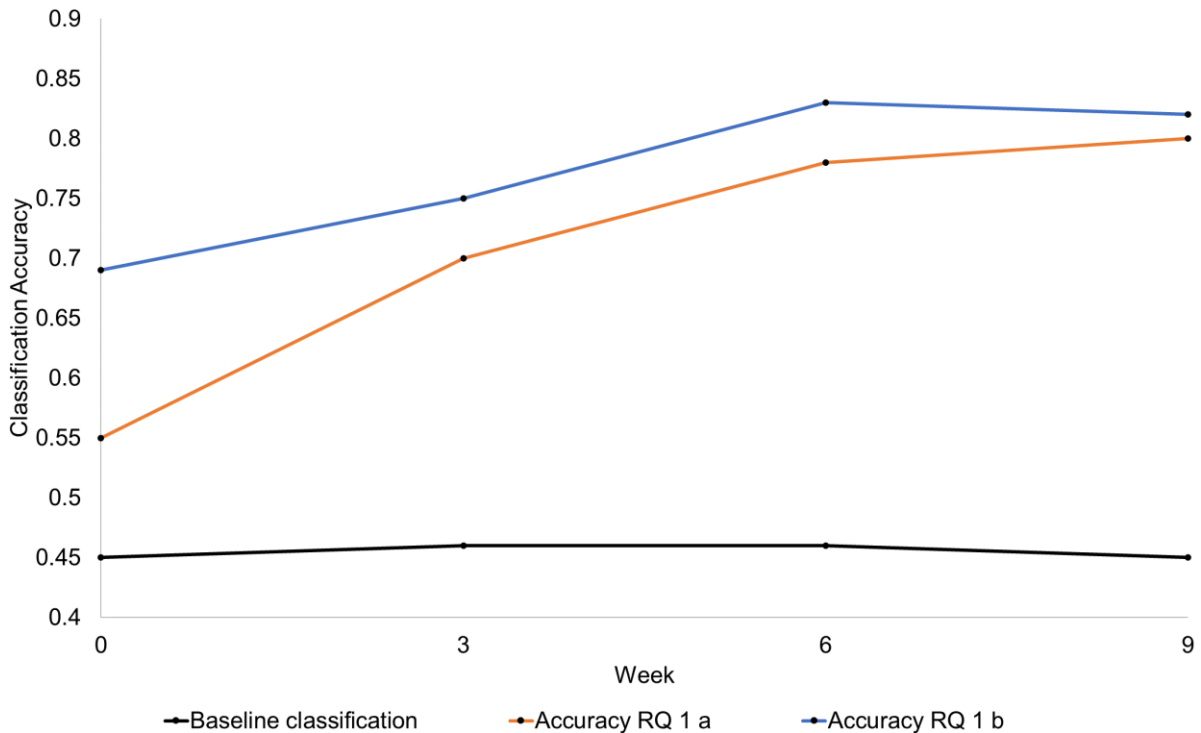


Figure 3.3. Results from the regression models for study-2 RQ1 -- predicting the performance on comprehensive quizzes held during weeks 3, 6, and 9 using three models Baseline (Black), RQ 1a (Orange), and RQ 1b (Blue).

linear regression performed better compared to 2<sup>nd</sup> order polynomial and regression trees at the regression task (adjusted-R<sup>2</sup> of 0.55 to ~0.36). Thus, the results provided here are for Logistic classification model and Multiple linear regression model, respectively.

### 3.5. Results

**RQ 1** – To what extent can we predict students’ performance on review quizzes and final grade using (a) reading and weekly quiz scores, (b) demographics and study intentions?



*Figure 3.4.* Results from the classification models for study-2 RQ1 -- predicting the performance on comprehensive quizzes held during weeks 3, 6, and 9 using three models Baseline (Black), RQ 1a (Orange), and RQ 1b (Blue).

Figure 3.3. shows the results of the three multiple linear regression models that is used to answer RQ 1 (regression task). Baseline model is presented in Black. As



hypothesized, the baseline model has the worst performance, close to 0.1 across all predictions. RQ 1 Model b (shown in Blue) performed better compared to RQ 1 Model a (shown in Orange). Recall that model 'a' only included the learners' performance on the course and model 'b' included demographic data as well as students' self-reported study plans and adherence. The maximum adjusted R-squared achieved by baseline model was 0.13, whereas the maximum adjusted R-squared achieved by RQ 1 Model a and RQ 1 Model b were 0.43 and 0.52, respectively for week 9 predictions.

Figure 3.4 shows the results of the logistic classification model that is used to answer RQ 1 (classification task). As before, Baseline model is shown in Black which performed poor compared to the other two models with prediction accuracies close to chance around 45% across. Classification accuracy for model 'a' is shown in Orange and model 'b' is shown in Blue. Similar to the regression models, model 'b' outperformed model 'a' and baseline models across all predictions. The maximum classification accuracy achieved by baseline model was 0.46, whereas the maximum classification accuracy achieved by model 'a' and model 'b' were 0.75 and 0.81, respectively for week 9 predictions. In addition, the final grade classification accuracies (which used data from posttest survey data and average performance on all review quizzes) over the 200 iterations of our model for model 'a' and model 'b' were 0.81 (+/- 0.06) and 0.86 (+/- 0.05), respectively and the adjusted R-squared was 0.69 (+/- 0.09) with RMSE of 7.59 (+/- 2.34). The overall average precision (0.84), recall (0.81), F1-score (0.82), and AUC (0.86) were very close to the levels of accuracy indicating that the models performed well.

Overall, the results from these three models shed similar light as the previous study. Baseline models that do not use any actual student learning data tend to do poorly compared to the other two models. This is an expected behavior of the baseline model and required to evaluate if the other two models' performance are better in comparison. If the results of the Baseline model are (abnormally) high, it would mean

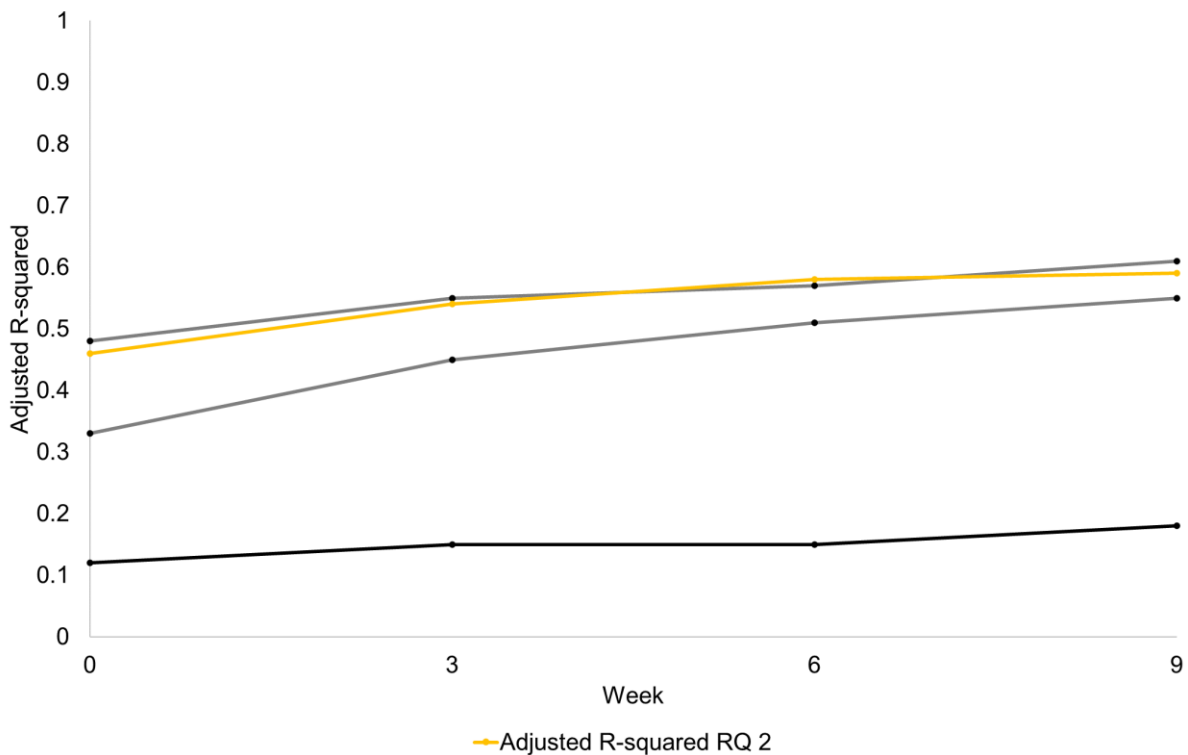
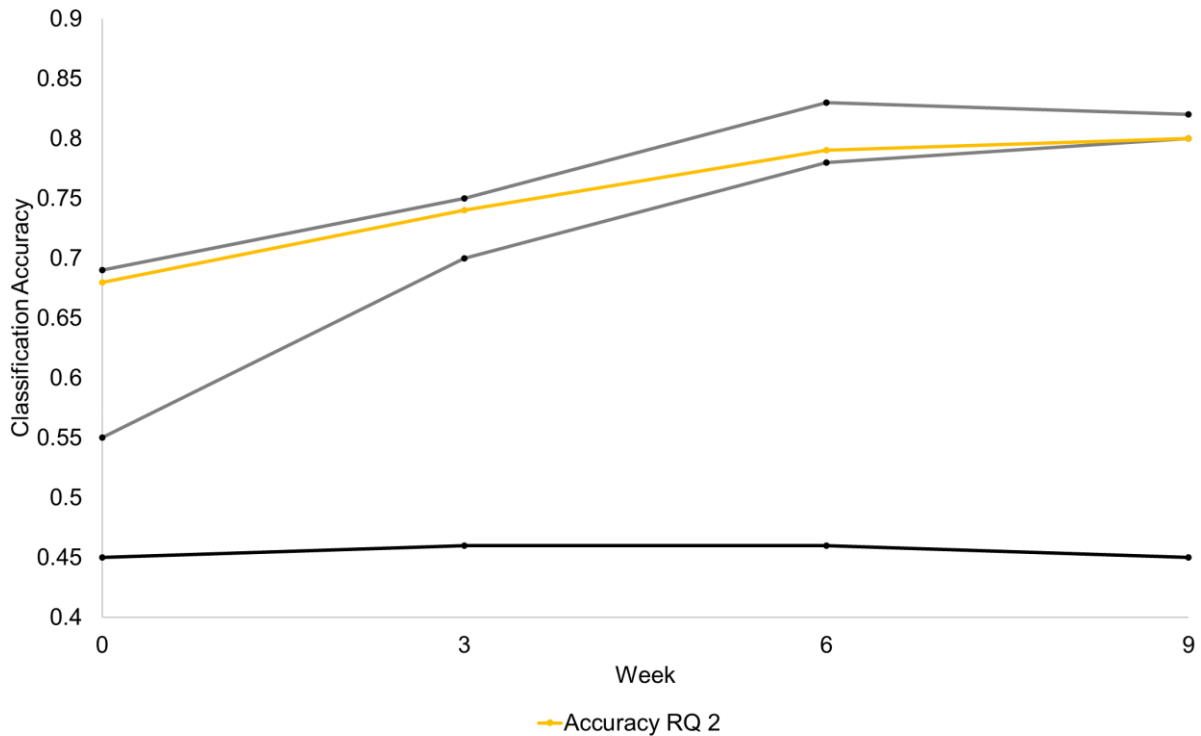


Figure 3.5. Results from the regression models for RQ2 -- predicting the comprehensive quiz scores during weeks 0, 3, 6, and 9 using the features listed for RQ 2 (shown in Yellow color). Models from RQ 1 discussed earlier are faded out but left in for comparison.

that the models are learning to predict outcomes from random noise. This would have indicated spurious results. However, the results shown from our baseline model are consistent with our expectation. Next, results suggest that the prediction trends we have



*Figure 3.6.* Results from the classification models for RQ2 -- predicting the comprehensive quiz scores during weeks 0, 3, 6, and 9 using the features listed for RQ 2 (shown in Yellow color). Models from RQ 1 discussed earlier are faded out but left in for comparison.

seen from the last study partially hold true. During the early learning phase, since the predictive models cannot learn from the students' performances are insufficient for making good predictions. However, once the model (b) is provided with demographics and the students' study intentions, the predictions are quantitatively better. This trend continued to hold true across our predictions. However, convergence did not occur, unlike in the previous study for both regression as well as classification tasks. This is perhaps expected since later learning in a classroom setting is not a function of simple

repetition and mastery as I have argued to be the case for WM training task. Furthermore, perhaps the number of predictions that we made compared to the previous study are also very limited (4 vs 14). Perhaps, this provided our model ‘a’ an insufficient understanding of the trends in students’ learning performances for the model to converge with the performance level of model ‘b’. The advantages gained from using non-performance features in model ‘b’ continued to hold across all predictions. This indicates that the extra features introduced in model ‘b’ are important for making early predictions which then continue to help our prediction models to better classify and quantify learners.

**RQ 2** – Can click behaviors improve predictions of models from RQ 1? If so, which of the click behavior features has the highest predictive value?

Let us now consider the results of models for RQ 2. Figures 3.6 and 3.7 show the results for regression task and the classification tasks, respectively. The baseline model (shown in Black), and the two models from RQ 1 are left in for comparison in both figures. Model RQ 2 (shown in Yellow) show that the results for this model have not improved in comparison with model ‘b’ form RQ 1. Recollect that model RQ 2 incorporated data from click-behaviors in addition to the features used in RQ 1 Model b. These results suggest that using click behaviors, as they were derived and utilized in the current context, do not provide any predictive value. This could be due to the nature of the click-behaviors that were derived within Canvas LMS. Specifically, since Canvas

Table 3.4

*Feature importance derived for comprehensive quiz predictions*

<b>Feature</b>	<b>Feature importance - Week 0</b>
highschool GPA	0.30

spacing intentions	0.12
SAT scores	0.10
part-time status	0.08
low-income status	0.07
<b>Feature importance - Week 3</b>	
average of pre-lecture quiz performances	0.35
average of homework performances	0.14
highschool GPA	0.11
SAT scores	0.09
study spacing intention	0.06
<b>Feature importance - Week 6</b>	
average of pre-lecture quiz performances	0.39
average of homework performances	0.20
highschool GPA	0.11
SAT scores	0.06
study spacing intention	0.06
<b>Feature importance - Week 9</b>	
average of pre-lecture quiz performances	0.26
average of homework performances	0.18
highschool GPA	0.09
change in study plan	0.09
study spacing intention	0.06

LMS is not a fully integrated learning environment and since the resulting clicks do not speak for students' learning out of the LMS, it is difficult to explain learning performances using click-behaviors. Furthermore, in comparison to the best model from RQ 1 (i.e., model 'b'), the results can be considered slightly worse, given that the additional features have a high variability and noise. This finding is in line with our previously reviewed literature where click-behaviors and log activities of students yielded varying degrees of value for understanding learning.

Next, to investigate which features contributed the highest predictive value, we used the feature importance derived from the logistic classification model to inspect the top 5 features for each prediction made from RQ 1 Model b (see Table 3.4). Results indicated that the most important features during week 1 were highschool GPA, spacing

intentions, SAT scores, part-time status, and low-income status. Specifically, students past academic records were the most valuable to making predictions of later learning during week 0 (without any data of the students' performances within the course). Specifically, the students' self-reported spacing intentions were relatively more important than that of SAT scores, the students' part-time status, and low-income status, indicating that it is important to ask the students of their study plans for the course. It is also possible that just by asking the students to discuss their study plans may have somehow played a role in their later self-regulation which determines their learning in the course. From week 3, the students' average performances on the pre-lecture quizzes and average homework grades were more predictive of their learning than Highschool GPA, SAT scores, and study spacing intentions. These features continued to be important throughout the predictions made. Notice that SAT scores were replaced by students' self-reported change in study plan which were collected after the course was completed. These findings make theoretical sense, given that early predictions are typically based on students' past achievements and perhaps immutable demographic features. Later, once the models start learning from the students' performance within the course, the predictions become more accurate and highly dependent on the students' performance during the course itself. One interesting thing to notice is that age and gender were not in the important features listed. In this context, this makes sense, given that the age range is very narrow and there was no reason for us to suspect any gender differences in the course (biochemistry). Overall, results across all models were consistent with our hypothesis and showed that students'

learning predictions were driven by highschool GPA, SAT scores, their study spacing intentions, and their performances during the course itself.

### **3.6. Discussion**

In the current study, we attempted to predict students' learning performances using the proposed stepwise predictive modeling framework from earlier. Specifically, we used features that are important for a blended learning course with both online and face-to-face components of pedagogy. Predictive models that were employed in the context of blended learning reviewed earlier, suggested that the prediction accuracies typically range between 65% – 83%. However, answers to the three fundamental questions – “How soon can we predict?”, “How well can we predict?”, and “How much information is needed for predictions?” have not been thoroughly answered. Results from this work, indicated that answers to these questions are typically dependent on one another. For instance, during week 0 (the earliest possible time for the predictions), the accuracy of our best models was around 70% which required all of the available features (except the click-behaviors). However, this was not the best possible model. If a researcher is willing to wait a few more weeks, say until week 6, then the prediction accuracy increased to 80% with the same amount of data as well as the students' performances through week 6. Thus, the current work provides a means for researchers to evaluate the *urgency – accuracy tradeoffs* when making relevant predictions. Furthermore, results from our model with only the data from the students' learning performances (RQ 1 Model a) has provided us insights into the limitations of predictive modeling efforts when researchers do not know much information about the learners. In this model, gaining prediction accuracies of over 70% would require the researchers to

use data until week 6 or beyond. Thus, this result can be considered synonymous with our findings from Study-1. Specifically, predicting later learning from early learning performances within a given context (or from on-task performances) yield subpar results and perhaps not ideal for making very early predictions. Furthermore, the results indicate that more information is generally better for robust early predictions. However, this is not always true.

In this specific study, using click-behaviors that the students displayed within the course LMS, did not improve the results. In fact, the results were worse compared with the model without this data. There are two potential explanations for this result: a) click-behaviors are not a good measure of students' learning behavior. As we have noted earlier, it is possible that context-agnostic click behaviors are not a good enough indicator of students' learning behaviors at all. This makes theoretical sense because believing that click behaviors or students activity log from the LMS provides insights into their learning and behavior would mean that we are attesting click behaviors to students' behaviors such as attention, engagement, interest, and attitude towards the course. Unfortunately, there is no clear way to make those connections between artificially generated metrics with students' goal-directed behaviors. One way to make those connections is by simply asking students about their intentions for making every potential click on the LMS. However, this is implausible given that on average, students within this course made over 800 clicks over the course of 11 weeks. Keeping track of every single one of clicks would be unrealistic, cumbersome, and error-prone. b) Perhaps another potential reason for why click behaviors did not predict learning could be due to the ways in which we measured the importance of click behaviors.



Specifically, the current work used three different click-behavior metrics: click-data for total course activity, click-data frequency of course activity per a given quiz, and click-data spacing. While none of these features added any predictive value to our models here, perhaps there might be more robust or innovative ways to remove noise from the click-data. For instance, perhaps identifying regular and irregular patterns in clicks of each student and using these patterns to making predictions might be a better approach. Perhaps making using of highly extreme click behaviors to analyze black swan cases might have yielded better predictions. However, for the sake of this thesis, we limited our analysis to the three features that we identified. More work is necessary to better understand and evaluate click-behaviors in the context of blended learning environments.

Overall, as hypothesized the models showed that including information about the students' study intentions during the pre-survey is a good predictor of overall grades and later performances. Specifically, the current best algorithm from RQ 1 (Model b) predicts students' learning accurately predict whether a student achieved at least a 'B-' grade nearly 80 out of 100 times as early as 3rd week of the course. This prediction accuracy is slightly lower than in a skill refinement context, such as on a cognitive training (working memory training) context seen in the previous chapter that has 86% accuracy on average by 1/3<sup>rd</sup> of the training period. The reason could be that the variances on working memory dataset are slightly lower because of the task being straightforward and increasingly difficult unlike in a typical classroom setting where knowledge building is known to be a more complex operation than skill acquisition (Michalski, Carbonell, & Mitchell, 1983).

The efforts of the current work are crucial to understand the relationship between study intentions and prediction of learning. Additionally, understanding features that affect various learning outcomes are not only a critical step for integrating predictive modeling, but also to understand the fundamentals of students' motivations that relate to learning. Current results show a promise in detecting, early in the learning phase, whether an individual learner may receive at least a B+ grade. This information may help inform pedagogical practices that can alter these projected learner trajectories by providing additional support to those students that might need more help.

The limitations of the current work pertain to its limited prediction capability. The current models can only differentiate those individuals that are above or below B+ grade level. This is due to the unequal distributions of grades of the current dataset. With the limited data we have, we did not seek to make predictions of grossly unbalanced letter grades. Further limitations related to bias in data selection exist with the current work. Specifically, those students that responded to the surveys might be intrinsically different than those students who did not respond to the surveys. Further work is necessary to overcome these two limitations.

## **CHAPTER 4: DYNAMIC OF MOTIVATIONS, PRIORITIES, AND STUDENT LEARNING: LOOKING BEYOND CLICK BEHAVIOR TO PREDICT STUDENT LEARNING IN AN ONLINE COURSE**

As a summary of the two previous chapters (2 & 3), I have demonstrated that predictive modeling can be utilized to make predictions of later learning from early learning in two different scenarios. The primary goal of the current chapter is built upon the results from the last two chapters. In chapter 2, I have explored the idea of using predictive modeling to evaluate a learning scenario in order to understand how predictive modeling can be used, specifically, to predict later learning, and what to expect from such an approach in short-burst time spans. The models employed were able to approach near 90% prediction accuracies by session 7 of the WM training in the short-burst time span of interest. In chapter 3, I applied this model to a real-world scenario (i.e., blended learning environment) and I was able to replicate the results from chapter 2. In doing so, I established the value of self-directed learning of students in making quantitatively better predictions over the models that learned only from the students' early learning behavior. The results suggested that we can predict later learning as early as half-way through the short-burst timespan of interest (i.e., an academic quarter or a fixed duration WM training) with accuracies consistently above 80%. I have also demonstrated that click behavior as a feature has very little to no predictive value in a blended learning environment. Furthermore, neither the overall quantity of the click behaviors, nor the frequency of the click behaviors to specific assignments/tasks were able to differentiate the learners' grades beyond demographics

and early learning quality. Let us now consider the possibility of click behaviors adding value to the learners' behavior in a fully online classroom experience.

The key point I explore in this chapter is the idea that understanding the learners' motivational, social-emotional, and affective processes lead to better predictions over click behavior-based models. This is not a controversial idea. A few studies discussed below from the existing literature of predictive modeling in online learning contexts investigated the idea of looking towards more learner-centric metrics (as opposed to using click behaviors alone) to make predictions of later learning and dropout rates. Most of the studies that did not make use of learner-centric metrics in building predictive models acknowledge the need to include these key factors that go unaccounted for to develop stronger predictive models. Therefore, the need for learner-centric metrics to empower predictive modeling of learning is not foreign or naïve. Note that I use 'learner-centric' as a word to encompass the learners' motivational, social-emotional, and affective processes that are at the heart of the goals of the current chapter. The issue of understanding context of students' actions is increasingly critical due to several reasons. In online learning, there is a lack of support and guidance from instructors directly since there is no face-to-face learning time. Meetings and lectures held via remote instruction tools and technology do not provide the same levels of engagement or teacher-directed regulation of learning. Thus, it falls up on the student to define, process, and regulate the learning. The lack of structured classes, fixed schedules, and emphasis on learning as a social process emphasizes self-regulated learning, goal setting, and the ability of the students to make decisions and determine when and how to engage or disengage from learning tasks. Prior research has shown that a majority of

learners in online platforms struggle with effective self-regulation and thus display poor learning (Kizilcec & Halawa, 2015; Tempelaar, Rienties, & Giesbers, 2011; Winne et al., 2006).

The question that remains unanswered is, to what extent are the click-behaviors valuable in making predictions of the students' later learning, their grades, and are learner-centric metrics effective means to empower predictive modeling above and beyond click-behaviors? To answer this question and to understand the value of learner-centric metrics, I explicitly compare predictive models that include or exclude the features of interest. In addition, I also explore the idea of using the dynamics of learner-centric metrics over the period of the course to look for consistency of the predictive relationship of the changes in learner-centric metrics and changes in performance of learners over the short-burst time span of interest.

#### **4.1 Click Behaviors in Predictive Modeling of Online Learning**

There are many challenges to environments of learning, be it offline, online, or blended learning, and with the process of learning itself, that are yet to be fully understood. One of the major problems that received attention in the recent past is the need for personalized learning. The root of this challenge is the individual differences in the varying factors that influence human behavior and specifically learning behavior differently for individuals. When a curriculum is designed and delivered to an average student with average skill level and average level of acceptable prior subject knowledge, then some students are likely to fall behind while some students feel unchallenged. Both of these issues are known to debilitate learning quality (Larrivee, 2000). While adaptive learning algorithms are said to alleviate this problem to a certain extent, the adaptive

technologies fail to account for the needs such as motivational and state of mind that might change from day-to-day in favor of prior ability, learning rates, and click behaviors (Mangaroska & Giannakos, 2019).

Specifically, most of the extant literature used click behavior within learning management systems (LMS) to understand student behavior. An emphasis has been placed on click behavior within fully online learning environment because of the lack of face-to-face time with teachers (that provides teachers with an intuition about students' learning during a lecture) and because click behavior is a readily available metric. Since teachers do not have the direct ability to observe, adjust, alter, or control the learning process, the reliance on metrics generated via the LMS in the form of learning analytics (LA) are widely promoted (see Aldowah, Al-Samarraie, & Fauzy, 2019 for a recent review). Per Aldowah and colleagues review (Aldowah et al., 2019), there are four important categories that educational data mining (EDM) and LA have been used to understand and promote learning: computer-supported Learning Analytics (CSLA – 120 articles or 30% of the reviewed articles), computer-supported predictive analytics (CSPA – 253 articles or 63.25%), computer-supported behavioral analytics (CSBA – 80 articles or 20%), and computer-supported visualization analytics (CSVA – 38 articles or 9.50%).

In their review, Aldowah and colleagues classified any article that utilized data mining techniques to understand the interactions within the course such as assessing students' click behavior in group activities to identify potential interventions, as CSLA. Typically, these studies used data generated via click behavior in the LMS to associate system-level objects (e.g., course related activities such as discussion forums, content

delivery, assessments) to the students' preferences in order to provide a greater level of support for individuals, specifically to make students more aware and better understand course activities and objectives. Most of the studies categorized as CSLA are focused on enhancing different aspects of learning that promote collaboration, networking, self-learning, self-assessments, and self-regulated learning. The studies classified as CSLA, also offered insights into effective ways to promote usage of task-specific activity logs to understand collaborative learning, associations of students within small social group work, communication habits, self-learning behavior such as self-explanations of complex concepts using the learning tools within LMS. Overall, goal-specific click behaviors and logs from LMS are used to broadly understand learner behavior to provide the teachers with tools to assess and evaluate the effectiveness of material and tasks assigned within online learning (Agudo-Peregrina, Iglesias-Pradas, Conde-González, & Hernández-García, 2014; Nussbaumer, Hillemann, Gütl, & Albert, 2015; Shum & Ferguson, 2012).

Aldowah and colleagues classified all articles that focused on predicting students' performance and retention as CSPA. Much of the reviewed literature fell into this category (63.25%) reflecting the importance of predictive analytics within literature. The overarching goals of articles classified as CSPA involved using user logs (often to act as a proxy for student engagement and participation), achievement, grades and quiz scores, and subject knowledge to discover hidden patterns within large datasets to predict outcomes and behaviors. CSPA were used to predict early dropouts, later learning, and to identify students who needed extra support to master learning material. Furthermore, predictive analytics were applied to understand and enhance quality of

learning material, constructing coursework, planning and scheduling classes, evaluating, assessing, and monitoring student performances (Baradwaj & Pal, 2012; Manek, Vijay, & Kamthania, 2016; Salas, Baldiris, Fabregat, & Graf, 2016).

Another major category of articles, classified as CSBA, were articles where data mining and learning analytics were applied to understand students' learning behavior. Information of students' behaviors within group work were assessed using click behaviors in lieu with aspects of personality and state of mind including motivation, metacognition, and attitudes to promote learning process. For instance, irregularities in students' performances and poor final grade performances were also detected using outlier behaviors in students' activities on LMS without looking at their actual performances in class (McCuaig & Baldwin, 2012).

Finally, Aldowah and colleagues classified any articles that utilized visualization techniques alongside data mining approaches as CSVA. Essentially, visual analytics were used to represent individual behaviors with respect to assigned class activities such as assignments to understand the learners' behavior (Peña-Ayala, 2014; Varun & Chadha, 2011). Some of the work reviewed included studies that used visual analytics to map online discussions and evaluating the quality of the posts using engagement (measured using clicks). Things such as frequency of visiting a resource, time spent on the resource, associations between subsequent clicks, associations/networks within group activities were visually analyzed (i.e., using graphical representations). The overall goal of the CSVA articles was to simplify complex data to track and understand students' behaviors within the LMS.



#### **4.1.1. Closed vs Open Models: Clicks Out of Context in Learning Analytics**

Out of the 402 articles reviewed, regardless of the data mining tools and analytical approaches used (e.g., sequential patterns, text mining, association rules, regression) to represent student learning and behavior, one common aspect was the use of click behaviors within LMS across a majority of the articles to understand student behavior and learning. There are two critical issues with using click behaviors to gauge engagement and course mastery. First, typically, the clicks that arise from students' interactions within LMS do not provide any context to their activities. For instance, every click within the environment are equal regardless of what the student was doing while clicking within LMS. The underlying assumption is that the click represents some form of engagement with the course content or with artifacts created in the LMS and that such engagement translates to learning outcomes. Second, there are a lack of clear guidelines and openness to using the clicks and student engagement within LMS to predicting their behaviors. For instance, the students (often, even educators) are not included in decision making process to tracking and determining student behaviors (Baylor & Ritchie, 2002; Brusilovsky et al., 2014).

In the earlier generations, the models that drive personalization within LMS were hidden from the end users (i.e., teachers and students). This approach was criticized for not providing transparency to the personalization process, not being inclusive, and not accounting for the students' self-efficacy or study strategies. Later, with the introduction of open learner models, the students and teachers were provided with an ability to incorporate self-reflections and self-organized learning elements to enhance the transparency of LMS. The advantage of including and accounting for study strategies

promotes positive value by understanding the potential differences in individuals click behaviors based on their task priorities and resource allocations (i.e, time of day or day of week a student wishes to study for the given course). However, working with open and transparent models does not only mean showing the end users the learners knowledge representations (which are obscure and hard to understand), but rather it means that the LMS is treated as a fully transparent and controllable medium that the students have access to and can interact with at their leisure (Bodily et al., 2018; Bull & Kay, 2013, 2016).

Learning analytics (LA), when used to provide feedback of students' activities, their engagement within LMS, and their interactions with resources of the course in a constructive manner could empower the end users to carefully assess and understand the impact of every step of the learning process. In theory, the open models of LA can help achieve these goals. Specifically, the guidelines provided by empirical research on visual analytics dashboards (Chatti et al., 2014; Verbert et al., 2014) have laid out the ground work to accomplish the greater goal of providing individualized feedback to students. The crucial elements identified by these guidelines are discussed by Vesin and colleagues (2018) in their recent review. These guidelines include data awareness; visualizations of activity data; self-reflection on activities; sensemaking; comparison with other learners; goal-oriented visualizations; open-learner model; and impact and behavior change. The goals of the proposed open model framework were to establish a robust criterion that makes the LA fully transparent for the end users (specifically, students). This will promote the ability of the learners to not only become aware of the data, the flow of the data, and the ways in which the data is used to drive the

recommendations and predictions, but also to help them by impacting the best practices to master the course content. Verbert and colleagues (Verbert et al., 2014) discussed their work centered on such an open LA model to answer questions pertaining to elements within LMS user activity that can be considered relevant user interactions, actions, and the ability of LMS to create opportunities for learners to self-reflect (on their progress, mastery, and overall learning). They discuss the relevance of including artifacts such as blogs, fora, twitter feeds, question responses, help requests, annotations, student-generated artifacts, social interactions, ratings, comments on blogs, time spent on tasks, test scores, and self-assessment results. Most of such artifacts within LMS are tracked via click behaviors of the learners within LMS. While getting information from virtual sensors hidden within LMS (e.g., time stamp trackers, click trackers...etc.) can accomplish the goals to some extent, they do not provide the full context of ongoing learning and the shifts in learning demands and social-emotional and affective processes. They suggest that in addition, use of physical sensors to capture facial expressions of learners alongside virtual sensors that can track actions of learners within the LMS (login behavior, contributions) are crucial to benchmarking and tracking the overall learner behavior. Their review provided some indication (7 out of 24 LA dashboards) that supports evidence for better engagement, higher grades, posttest results (of their choice of metrics such as final exams), and improved self-assessments using a combination of physical as well as virtual sensors.

Unfortunately, however, automated tracking of students via click behaviors within LMS are not fully capable of detecting student behaviors. For instance, evaluation studies reported by Verbert and colleagues (Verbert et al., 2014) also showed that

students rated learning dashboards very poorly due to the incompleteness of their context and scope. For instance, many relevant learning activities happen outside the LMS that are not tracked using clicks generated within LMS. For instance, a student might be accessing one resource within LMS only once and then study the material by themselves offline or they might be learning using resources beyond LMS (e.g., Khan Academy or YouTube videos). Furthermore, lacking comprehensive tracking is noted to be challenging within closed environments such as Canvas LMS since the clicks only represent the “tip of the iceberg” of the entire learning process (Verbert et al., 2014). The issue of inability to track every productive activity of the learner that leads to good grades can be extended to open learning models. A wide array of tools and services are used to drive learning within open learning models without universal standards and comprehensive access to all activities that go beyond the LMS (for instance, third-party tools can be used to promote visualizations of students’ click behaviors). This makes it harder to track everything the students do to gain knowledge and succeed in the course. There are also other issues of student privacy that go beyond the scope of the current work which make it difficult to track every relevant (to the learning) click behavior. This leads us back to the central questions of the current work:

Can click behavior within Canvas LMS be used to predict learning quality beyond demographics and information about the students such as past academic records? If so, to what extent? If not, can we predict the students’ learning better by understanding the learning needs, motivations, cost-value associated with learning the course, and other non-academic obligations?

For instance, the disproportionately high rates of attrition in online learning (an indicator of negative motivation) has been extensively studied and by far remains the pressing issue of online learning as a platform that can compete with face-to-face learning (Bawa, 2016; Gallego & Topaloglu, 2019; Jordan, 2015; Kizilcec & Halawa, 2015). One potential reason for the lack of sufficient work investigating the motivation and subsequent failure to reduce dropout rates could be due to the emphasis on students' cognition while ignoring affective and social-emotional processes (Miltiadou & Savenye, 2003). Thus, we need to shift our focus towards understanding the complex behavioral experience of learning with an emphasis on behavior, early performance trends, demographics, intentions of study spacing, adherence to the study intentions, and expand our modeling efforts to include the dynamics of motivation on a daily basis. The overarching goal is to understand the differential predictive value of these key features in determining the quality of learning and the individual differences in the predictive values of these features beyond the click behavior. Note that the goal of current work is not to determine the factors that lead to the high attrition rates that plague most online learning platforms. Rather, the goal is to understand the predictive value of motivational, affective, and social-emotional processes in a fully online course conducted at a university setting that does not have the issues of high attrition rates typically noticed in private online learning platforms such as Coursera (Glance & Barrett, 2014; Kolowich, 2013). This provides an opportunity to focus on the predictive value of learner-centric factors and click behaviors to the on-going learning itself rather than as a marker for dropout rates.

#### 4.1.2. Learner-Centric Factors

So far, I have discussed the importance of further investigating the validity of click behavior in predicting the quality of learning within fully online learning environments due to its importance, in theory and in practice. Next, I will discuss the need to understand the interactions of click behaviors with other important learner-centric factors (e.g., motivation) that determine quality of learning in making predictions of learning. I investigate factors central to the learners, specifically, their motivations, goals and values, state of mind, and self-efficacy that are known to be important mediators of learning in online environments in predicting the grades alongside the students' click behaviors within the LMS (Aldowah et al., 2019; Daud et al., 2019; Larrabee Sønderlund, Hughes, & Smith, 2019; Pardo, Jovanovic, Dawson, Gašević, & Mirriahi, 2019). There are a wide range of factors that contribute to an individual's learning regardless of the learning medium (online, hybrid, or offline) that simply cannot be captured by click behaviors. Let us now consider the learner-centric factors that are explored in the context of online learning behavior that can empower predictive models tested in the previous chapter.

*Self-efficacy* is defined as “beliefs in one’s capabilities to organize and execute the courses of action required to produce given attainments” (Bandura, 2010). Self-efficacy beliefs are known to drive an individual’s thinking, feeling, motivation, and behavior through cognitive, motivational, affective, and selection processes. Self-efficacy is described as the prime factor that drives human agency. However, self-efficacy is known to be domain specific. High self-efficacy in one domain does not transfer to or guarantee high efficacy in another domain. However, there has been

some evidence to suggest that self-efficacy traits (or sources of confidence or lack thereof about their learning abilities) are shared between offline and online classroom experiences to some extent (Y. C. Lin, Liang, Yang, & Tsai, 2013). Factors such as past performance experiences or enactive mastery experiences (e.g., grades in the past courses), vicarious experiences (e.g., performance of other students in a shared learning experience), verbal persuasion (e.g., encouragement from an authentic constructive feedback), and physiological states (e.g., aversive arousal during learning is debilitating). However, past work specific to online learning and self-efficacy has shown that four key factors drive learners' experiences. Success rates in past online learning experiences, training received before the start of the course, authentic feedback from the instructor, and anxiety related to the learning technology are all considered to be impactful to their journey (Alqurashi, 2016; Bates & Khasawneh, 2007).

In a recent review of the literature of self-efficacy in online learning environments, Alqurashi (Alqurashi, 2016) has discussed the three main categories that have been a focus of all related research conducted between 1997 and 2015:

- (i) computer self-efficacy – learners' confidence in their ability to use a computer and other mobile devices,
- (ii) internet and information-seeking self-efficacy – learners' confidence in their ability to search the internet for relevant information to succeed in a classroom, and
- (iii) LMS self-efficacy – learners' confidence in within LMS and how it affects their performances within the LMS.

Across these three categories, the focus of research so far has been on the factors surrounding the confidence in using technology. For instance, Jan (2015) focused their work on past academic experiences with computer self-efficacy and satisfaction in an online classroom. Similarly, Kuo and colleagues (2014) have focused on internet self-efficacy and its impact on online learning experiences. Research so far has shown that the students' computer self-efficacy had the largest positive and significant relationship with online learning, satisfaction, and likelihood of taking future online courses. Internet and information-seeking self-efficacy on the other hand, had shown a weak to no relationship with classroom performances and satisfaction. LMS self-efficacy also showed a weak relationship with classroom performances in fully online learning contexts. However, LMS self-efficacy did have a significant positive relationship with hybrid learning.

In summary, the existing literature is inconclusive and falls short in two important ways: 1. It does not speak for course-specific self-efficacy and its role in general self-efficacy and online learning and 2. None of the existing studies investigated the role of changes in self-efficacy over the learning period and impact of its dynamics to learning experiences. The former is important to understand the fundamental nature of the relationship of self-efficacy and learning within online learning contexts and the latter is important to understand how this relationship evolves and drives learning over the learning period. Although computer, internet and information-seeking, and LMS self-efficacies are related to online learning experience in some way, more research is needed to develop our understanding of their relationship with successful online learning and in lieu with other important factors that drive learning such as grade



expectations (self-regulated learning), goal-setting, and subjective task-value, and attributions.

*Self-regulated learning* involves the effortful and deliberate actions that an individual takes in order to plan, execute, observe, evaluate, and alter behavior specific to different learning contexts (Nussbaumer et al., 2015). While self-efficacy speaks for the confidence an individual may have towards their own ability to complete a goal (i.e., course), it is only part of the entire story about how to complete challenging online courses (Cho & Heron, 2015; Hodges & Kim, 2010). It is important to maintain sustained motivation and necessary actions to set and accomplish goals throughout the learning process. These effortful and deliberate actions, referred to as regulation and volition, are known to be strong predictors of academic success (Corno et al., 2001; Gabrielle, 2003; Gollwitzer, 1999; Zimmerman, 1989). Self-regulation has been long established as a very important aspect of learning process. According to Zimmerman (1989) classic work, there are three components of self-regulation.

- (i) Behavioral component – For example, alterations to behavior based on observations, judgments, adjustments to performances
- (ii) Environmental component – For example, through social interactions, persuasion by peers, parents, or teachers, indirect peer influences such as class performances
- (iii) Personal component – For example, choices and actions taken to engage or disengage from tasks and persistence through the tasks

Specifically, the personal component of self-regulation is directly related to self-efficacy of an individual since the choices and actions taken to engage with a task (or a choice

not to engage with the task) are directly dependent up on the confidence of an individual in the said task. Therefore, it is implied that while self-efficacy impacts learning at the beginning of the course, self-regulation impacts sustained learning. For instance, studies have shown that self-regulation can become impaired in online mathematics learning, despite high levels of self-efficacy, due to the lack of face-to-face interactions with the instructors and peers (Dennen, Darabi, & Smith, 2007). Despite consistent work showing that low self-regulation impairs online learning, very few studies have inspected the nature of self-regulation over the period of a course and how it alters the performances of individuals. Specifically in online learning contexts, self-regulation is said to be influencing students' performances via various mechanisms including goal-setting, commitments, effort regulation, and persistence (Corno et al., 2001; Kizilcec & Halawa, 2015). Overall, there is a need to understand the reciprocal and dynamic impacts of self-regulation and self-efficacy in online learning contexts.

*Goal-setting*, a part of social learning theory of Bandura (Bandura, 2012; Bandura & Jones, 1962), is a central aspect of self-regulation. According to this theory, learners often set goals such as specific skill acquisition, gaining knowledge, completing work, securing good grades at the beginning of any activity. This is referred to as goal-setting and has been studied alongside the many social and affective processes that influence learning process. Goal-setting plays a key role in the first of the three cyclical phases of self-regulation described by Zimmerman (2000) described as "Forethought phase" where a learner sets goals and plans strategically to achieve a certain goal which are molded by self-motivation beliefs such as self-efficacy, outcome expectations, and intrinsic interests, values, and goal orientations. Therefore, it is important to

understand how a learner within online learning setting sets goals and follows through with specific goals related to the course. It is also important to understand how an individual ranked the importance of current goals with competing goals they are required to accomplish during the learning period (i.e., other courses and non-academic goals). Recent review and meta-analysis work to evaluate goal-setting in online learning has shown that determining goals, time management, and effort regulation surrounding each goal have a strong statistically significant relationship with learners' performances compared to factors such as elaboration, rehearsal, and help seeking (Broadbent & Poon, 2015; Richardson, Abraham, & Bond, 2012). For instance, self-regulated learning improved substantially in an online web-based portfolio assessment system by incorporating a diary where the students can set goals and keep track of these goals over the learning period (Chang, Tseng, Liang, & Liao, 2013). Additionally, Broadbent and Poon's meta-analysis (2015) has shown that self-regulated learning and goal-setting strategies had a weaker relationship with performance in online learning settings compared to face-to-face classroom experiences. They surmise that potential reasons for the smaller effect sizes could be due to a combination of issues such as measurement errors of self-regulation and engagement in online contexts, assumptions that self-regulation strategies work similarly in both offline and online contexts, and that fact that taking online courses by itself does not promote self-regulation. Additionally, it is important to evaluate the effects of self-efficacy, goal-setting, and the overall process of self-regulation during online learning by considering the second (performance and volitional control) and third phase (Self-reflection and self-evaluation) of the three cyclical phases of self-regulated learning proposed by Zimmerman (2000). The second

phase of this cycle is relatively straight forward where the learners engage in a learning activity, controlling their learning processes, and monitor their own performances over the learning period to fine tune their learning, perhaps, via open learning models described earlier (Abrami, Bernard, Bures, Borokhovski, & Tamim, 2011). The third phase, self-reflection, involves self-judgement via evaluations and casual attribution and self-reaction via notions of self-satisfaction, affective, as well as defensive responses. Thus, it is also important to understand the ways in which the learners' attributions mold the learning within the self-regulated learning cycle triad.

*Attributions* for successful academic endeavors are part of the final phase in the self-regulated learning cycle. Attributions refer to the interpretations a learner makes to make judgements about causes of own and others' behaviors and the results of such behaviors (i.e., academic success or failure). In the iterative process of the self-regulated learning, attributions play a central role because it affects as well as predicts several behavioral traits displayed during the learning process (Kitsantas & Zimmerman, 2006; Schunk & Zimmerman, 2012; Zimmerman, 1989). Traits such as procrastination, effort, perceptions of ability, perceptions of context and external influences, and luck are evaluated via students' attributional beliefs. Such beliefs are critical to the learning process since they act as a feedback mechanism within the self-regulated learning process that influence and drive the interpretations students' make about their performances. The attributions are more important in online learning experiences due the bulk of learning and self-regulation lies with the students and peer and teacher interactions have little influence on student beliefs leading to negative tendencies such as procrastination (Klingsieck, Fries, Horz, & Hofer, 2012). To remedy the potential

maladaptive practices and to understand the uncertainty and unpredictability of influences such as novel or recurrent daily life experiences (i.e., entering a new school, taking a new job, babysitting to help parents, moving to a new location, dealing with illness...etc.), it is important to understand the students' attributions and how they change over the learning period (Heckhausen, Wrosch, & Schulz, 2010). These attributions are not only important in learning mechanisms but also important to overall life-span development as well as for success and healthy aging (Haase, Heckhausen, & Wrosch, 2013; Schulz & Heckhausen, 1996). Specifically, at the university level, shifts from high school to university, having to move to a new location, increased frequency in failures, boredom, lack of support from professors, increased financial demands, and unstable social networks are identified as straining objectives that the learners' need to overcome (Hamm, Perry, Chipperfield, Murayama, & Weiner, 2017; Parker, Perry, Chipperfield, Hamm, & Pekrun, 2018). In the current world, having to deal with a global pandemic while taking courses in an online environment, sometimes for the first time, might be considered a factor that leads to negative attributions and poor performances. Thus, it is important to understand the cyclical relationship and the predictive values of these features in an online learning environment. For instance, Perry and Hamm (Perry & Hamm, 2016) have shown that treatments that target negative attributions can alter the self-regulated learning process in a positive way by understanding the negative setbacks from previous unsatisfactory experiences in online learning environments, and promoting deep self-reflections to initiate cognitive processes that facilitate receptiveness and engagement which eventually lead to higher academic performances (Rakes & Dunn, 2015; Rakes, Dunn, & Rakes, 2013).

*Subjective task-values (STV)* of students interact with their self-efficacies to drive the learning and course experiences according to the expectancy-value theory (EVT - Eccles, 2013). According to EVT, an individuals' learning and performance, persistence (or dropout from a task), and individual choices leading up to a completion (or lack thereof) of a task depend on the expected success rates and values attested with the task. In contrast to self-determination theory (Deci et al., 1991) which emphasizes the use of learners' motivations to explain behaviors of students, expectancy-value theory describes the ways in which motivation is represented through mental processes. One thing to note is that expectations of task performances, while analogous to self-efficacy, differ in the way the self-concept of ability (expectations in EVT vs competency beliefs in self-regulated learning). STVs include several sub factors:

- a. Attainment value: the subjective importance of attaining success on a task
- b. Intrinsic value: the perceived enjoyment of doing the task itself
- c. Utility value: the perceived relation of the success at current task with present and future goals
- d. Cost: the negative emotions and feelings related to engaging in the task such as fear of failure, anxiety, remorse, effort required to succeed, lost opportunity by engaging in the current task.

Per Eccles' model, values are a relative worth of commodity, activity, or person and the consequent attraction/repulsion by the object/activity. This value is subjective due to the individual differences in assigned values to each task. The models show that STVs are a function of intrinsic value (enjoyment gained from doing the task), utility value (usefulness in the current and future plans of the individual), attainment (sense of self,

and personal goal achievement), and costs (a subjective opinion about the amount of effort spent on a task and the worth of doing such task). An extensive amount of work has shown the importance of STVs to predicting learners' intentions as well as goal-setting and persistence in traditional classroom environments such as for learning and persistence in mathematics (Meece, Wigfield, & Eccles, 1990) and to attend graduate school, family life, and STEM research careers (Battle & Wigfield, 2003; Tan-Wilson & Stamp, 2015). While not as extensive as in traditional classrooms, some evidence exists for the importance of understanding students' intrinsic value (Bong, 2001, 2004), utility value (Yang, 2018), or the entire spectrum of subjective task values (Chiu & Wang, 2008; T. Lin, Imamiya, & Mao, 2008) in online learning contexts. The results of these students indicated that expectancies, attainment, utility, and intrinsic values, and negative emotions associated with failure and anxiety were moderate to strong predictors of performance in online classrooms along with self-efficacy and goal-setting. Additionally, Eccles and her colleagues' (1983) extensive work showed that 'subjective task value (STV)' is a key determinant of goal attainment. Various studies have shown that factors such as self-determination, satisfaction, ARCS (attention, relevance, confidence, and satisfaction), self-efficacy, task value, self-determination theory (SDT) have a positive predictive value towards students' performance and success rates in online courses and the performance is moderated by self-regulated learning process (Doménech-Betoret, Abellán-Roselló, & Gómez-Artiga, 2017; Gabrielle, 2003; Roca & Gagné, 2008; Vallerand & Blissonette, 1992).

## 4.2. Current Study

With the primary conclusion that the click behaviors yielded no predictive value to learning of students within a hybrid classroom environment, here, I look to reproduce the results of Study-2 in a fully online course. While a plausible outcome is that the results are, indeed, replicable, it is important to understand the validity of the results within the context of a fully online course due to the reasons explained earlier. Furthermore, I will investigate five important factors (Table 4.1) and affordances that are known to be of theoretical importance for learning as discussed earlier: self-efficacy, grade expectations, subjective task value, goal-setting (academic and non-academic), and the emotions associated with the goals (regret/satisfaction). I hypothesize that these learner-centric affordance measures yield more predictive value than the click behaviors since click behaviors do not provide context of learning and behavior and might be inaccurate indicators for engagement and learning.

Table 4.1.

*An adaptation of MACM: theories and variables of interest and their definitions (McGrew, 2007).*

---

Theory of motivation	Variables of interest	Definition
Self-efficacy theory	Self-Efficacy	“A person’s belief about the perceived causes (internal vs. External) for their success or failure. An internal attribution orientation is present when a person perceives their success or failure as contingent on their own

---



---

		behavior and due to relatively unchanging personal characteristics. An external orientation is present when success or failure is perceived as being under the control of others, unpredictable, and the result of luck, chance, or fate.”
Grade expectations  (Self-regulated learning theory)	Online Self-Regulation, effort regulation, online expectations  grade expectations	A learner’s use of “self-regulated learning strategies, responsiveness to self-oriented feedback about learning effectiveness, and their interdependent motivational processes. Self-regulated students select and use self-regulated learning strategies to achieve desired academic outcomes on the basis of feedback about learning effectiveness and skill”
Subjective task value  (Expectancy-value theory)	Utility Value, interest value, attainment value, cost value	“An individual’s behavior is a function of the expectancies of utility, interest, attainment and the cost consequences and the value of the goal the individual is working towards”
Goal-setting theory	Goal-setting – Helpful, list of other courses, other activities, completion	“A person’s ability to set, prioritize and monitor progress towards appropriate and realistic short-(proximal) and long-

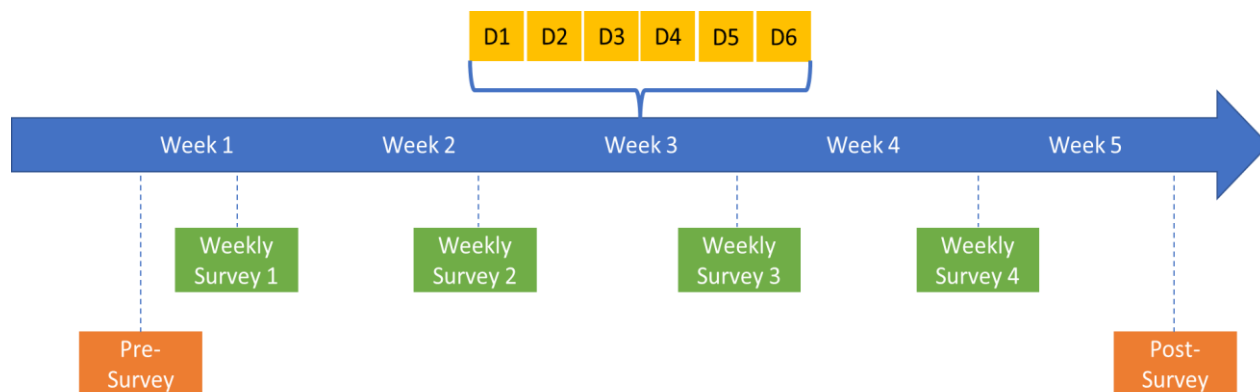
---

	rates and ranks of activities	term (distal) academic goals that serve to direct attention, effort, energy, and persistence toward goal-relevant activities (and away from goal-irrelevant activities)”
Attribution theory	Consciousness	“A person’s beliefs, self-evaluation, and self-awareness regarding their academic-related skills and abilities.”

*Note.* Reproduced from the Model of Academic Competence - McGrew (2017). The definitions are copied verbatim to retain the meaning intended by the original author. In the public domain. (c) Institute for Applied Psychometrics, llc 01-07-08

#### 4.2.1. Dataset (Online lower division Biology and Chemistry of Cooking Course)

I analyzed learning behavior of 147 students, enrolled in a 5-week long online lower division course, Biology and Chemistry of Cooking, held at UCI in Summer 2018. This was an elective sophomore course that fulfills breadth requirement for non-biology majors. This course was implemented in a fully online environment with no face-to-face elements. The course consisted of 17 lectures with post-lecture quizzes that were worth 4 possible points each, 18 reading quizzes worth 6 possible points each, two cumulative midterm exams, and one cumulative final exam. Additionally, the course consisted of 5 weekly quizzes with 20 maximum points possible for each. All exams, quizzes, and course materials were completed online via course Canvas LMS. The aim of this study was to understand the predictive value of self-reported study strategy, and self-reported motivational dispositions listed in Table 4.1 in addition to in-class performances of students on their quizzes and tests, demographics, and prior performances measured as SATs and GPAs. The self-reported study intentions were measured using the same



*Figure 4.1* Timeline of data collection -- Data were collected using pre-course survey, post-course survey, weekly surveys at the end of each weeks' learning (4 weekly surveys), and 6 daily surveys at the end of each day during week 3 (excluding Sunday).

protocol detailed in chapter 3. The motivation of each student and the dynamics of the motivations over the learning period were captured using questionnaires that targeted specific motivational dimension listed in Table 4.1. A list of sample questions that are used to measure motivation across constructs are listed in Table 4.2. Each construct was measured using standardized questionnaires listed. Each questionnaire consisted of 1-7 questions varying depending on the questionnaire. While many of the questionnaires involved measures on Likert scales, some of the questions also included fields for open-ended responses from the learners (see Table 4.2 for examples).

#### **4.2.2. Participants and study context.**

Institutional records for all 147 students are collected to identify demographics including age, gender, SES status, underrepresented minority status (African American, Native American, or Latino/a), as well as GPAs, and SAT scores (Table 1). 98% of the enrolled students are sophomores with an age range of 18-25 with a mean age of 19. 40 students declined to complete the surveys and are excluded from further analysis. 8 students dropped the course within the first 2 weeks and were not included in the final dataset. The final dataset consisted of 99 participants. The final dataset included slightly

more female students (56%) than male students and 60 (41%) students reported minority status. 59% of the students reported that they were taking an online course for the first time. Students were requested to respond to surveys before and after the course (Pre-and post), Once every week for the weekly surveys, and once every day for the daily surveys during week 3 (Figure 4.1).

Table 4.2

*A list of sample questions by construct used as predictors for the full predictive model in Study-3*

<b>Name of the variable</b>	<b>Pr e</b>	<b>Pos t</b>	<b>Construct</b>	<b>Sample questions</b>	<b>Response values</b>	<b>Brief Description</b>
<b>Course Goals</b>	x		Motivation for Taking Course	Please select all your goals for taking this course	1=Get a high grade, 2=Just pass the class, 3=Learn the material thoroughly, 4=Learn specific parts of the material that I'm especially interested in, 5=Get better at cooking, 6=Broaden my interests, 7=Other	goals for this course
				What is the goal you wish to accomplish by taking this course?	open-ended	goals for this course – open ended question
<b>Study Days</b>	x	x	Course Plan – Study Intentions	How many days of each week will you work on this course?	1 to 7	days per week you will work on course
<b>Study Plans</b>	x		Course Plan – Study Intentions	Would you say you have a study plan for the course?	0=no, 1=yes	have a study plan for the course
				What are your study plans for this course?	open-ended	
<b>Change in</b>		x	Course Plan –	Think about the study plan that	0=No, I stuck to my study	changed study plan

<b>Name of the variable</b>	<b>Pre</b>	<b>Post</b>	<b>Construct</b>	<b>Sample questions</b>	<b>Response values</b>	<b>Brief Description</b>
<b>study plans</b>			Adherence to Intentions	you had at the beginning of the course. Did you end up changing your study plan?	plan, 1=I never had a study plan, 2=Yes, I changed my study plan a bit, 3=Yes, I changed my study plan a lot	for the course
<b>Study hours</b>	x	x	Course Plan – Study Hours	On average, how many hours per week do you plan to spend on all aspects of this course?	1 to 40	hours spent per week?
<b>Grade wanted</b>	x	x	Grade Expectations	What grade do you want to get in this course?	13=A+, 12=A, 11=A-, 10=B+, 9=B, 8=B-, 7=C+, 6=C, 5=C-, 4=D+, 3=D, 2=D-, 1=F	wanted grade
<b>Grade expected</b>	x	x	Grade Expectations	What grade do you expect to get in this course?	13=A+, 12=A, 11=A-, 10=B+, 9=B, 8=B-, 7=C+, 6=C, 5=C-, 4=D+, 3=D, 2=D-, 1=F	expected grade
<b>Other courses</b>	x	x	Other Courses – Course load	How many other courses are you taking this summer?	1=0, 2=1, 3=2, 4=3+	number of other courses taking
<b>Importance of course</b>	x	x	Other Courses – Relative Course Importance	Compared to other courses you are currently taking, how important is this course?	4=Most important, 3=Second-most important, 2=Third-most important, 1=Fourth-most important	importance compared to other courses
<b>Other activities</b>	x		Other Activities List	What other important activities do you plan on doing in July while completing this course? (e.g., working for pay, caring for family members, taking another	open-ended	other activity 1-15

<b>Name of the variable</b>	<b>Pre</b>	<b>Post</b>	<b>Construct</b>	<b>Sample questions</b>	<b>Response values</b>	<b>Brief Description</b>
				course, playing sports, completing home projects, etc.)		
<b>Other activities completion rate</b>		x	Other Activities Completion	At the beginning of the course, you said you planned on doing the activities below. Did you end up doing them?	0=No, 1=Yes	actual completion of other activity 1-15
<b>Course rank</b>	x	x	Other Activities Rank	Please drag and drop your responsibilities during this course in order from most important to least important	1 to 16	importance rank of course responsibilities
<b>Other activities rank</b>	x	x	Other Activities Rank	Please drag and drop your responsibilities during this course in order from most important to least important	1 to 16	importance rank of other activity 1-15
<b>Hours spent on other activities</b>	x	x	Other Activities Time	On average, how many hours per week will you spend on each of these activities in the month of July?	0 to 40	hours per week on other activity 1-15
<b>Online self-regulation</b>	x	x	Online Self-Regulation	how often do you work in a place where you can read and work on assignments without distractions?	slider: 1=Never, 5=All the time	work where there are no distractions (study habits)
<b>Effort regulation</b>	x	x	Effort Regulation	I often feel so lazy or bored when I study for this class that I	slider: 1=Strongly disagree, 5=Strongly Agree	often feel lazy or bored studying

<b>Name of the variable</b>	<b>Pre</b>	<b>Post</b>	<b>Construct</b>	<b>Sample questions</b>	<b>Response values</b>	<b>Brief Description</b>
				quit before I finish what I planned to do		
<b>Online expectations</b>	x		Online Expectations	When taking an online course, I expect to perform...	slider: 1=Not at all well, 7 = Very well	expect to perform well in ol course
<b>Self-efficacy</b>	x	x	Self-Efficacy	I'm certain I can master the skills taught in this course	slider: 1=Not true at all, 5 = Very true	can master skills in this course
<b>Utility value</b>	x	x	Utility Value	How beneficial for your daily life is understanding the biology and chemistry of cooking?	slider: 1=Not beneficial at all, 7 = Very beneficial	course beneficial for daily life
<b>Interest value</b>	x	x	Interest Value	How often do you wonder about the science behind cooking?	slider: 1 = Never, 7 = Very often	wonder about science of cooking
<b>Attainment value</b>	x	x	Attainment Value	How important to you, personally, is it to be a person who understands the science behind cooking?	slider: 1=Not at all important, 7 = Very important	important to be a cooking science person
<b>Cost value</b>	x	x	Cost Value (Emotional)	How stressful will this class be?	slider: 1=Not at all stressful, 7 = Very stressful	class will be stressful
<b>Attributions</b>		x	Consciousness	I see myself as someone who ... does a thorough job	1 = Strongly disagree, 3 = Neither agree nor disagree, 5 = Strongly agree	I do a thorough job
<b>Attributions</b>	x		Persistence	How likely are you to stay	slider: 1=Not at all likely, 7 =	likely to stay in this



Name of the variable	Pre	Post	Construct	Sample questions	Response values	Brief Description
				enrolled in this course?	Very likely	course
<b>Adherence to Intentions</b>		x	Activities	Preparing for classes (studying, reading, writing, homework, lab work, etc.)	(0) 0 hours per week ... (1) 1-5 ... (2) 6-10 ... (3) 11-15 ... (4) 16-20 (5) 21-25 ... (6) 26-20 ... (7) 31+	hours spent preparing for classes each week
<b>Goal setting</b>		x	Goal Setting Helpful	In the paid surveys during the course, you were asked to list your planned activities and goals for each day or week. Do you remember doing this?	0=no, 1=yes	remember goal setting
<b>Goal setting</b>		x	Goal Setting Helpful	To what extent did you find planning daily or weekly academic activities was helpful? - Planning daily academic activities	slider: 1=not helpful at all, 7=absolutely helpful	planning daily activities helpful

*Note.* A full set of construct items are listed in Appendix A, Table A.1.

#### 4.2.2. Survey Design

**Pre-survey** was assigned three days before the course began. Students were requested to complete the *pre-survey* in four days. In the *pre-survey*, students were asked whether they were interested in participating in the study. The **post-survey**, however, was activated on the day after the last lecture and was due before the final exam (thus obtaining final grade expectations before taking the final exams). A sample list of questions asked during pre- and post-survey are listed in Table 4.2.

**Weekly surveys** were given four times in total throughout the class. Each weekly survey was due two days after the due date of their first four weekly reviews. Survey questions mainly collected information on students' activity plans for the incoming week and their change of grade expectations.

**Daily surveys** were assigned to students in week 3, every day from Monday through Saturday. Students were asked about their reflections of the activities and priorities for the past day, to what extent those activities were accomplished, the associated emotions with each of the task completion rates, and their activity plans for the next day.

Most of the questionnaires were implemented using Qualtrics™ (Qualtrics, 2016) since the adaptive algorithms on the platform can be used to generate formative questions. In other words, responses from one question can be used to generate subsequent questions that are used to follow-up the next day. For instance, one of the questions in the daily surveys asked each student “what activities are you going to do today (relevant to this course)?” If, say, a student responded to the question with “*watch videos*”, or “*office hours*”, the follow up questions asked subsequently included the

student's own responses as part of the question. For example, a sample follow up question was "How likely is it that you will do this task: *watch videos*?" Similarly, a follow up question asked on the next day's daily survey included, "Yesterday, you said you were going to *watch videos*. To what extent did you complete it?". Furthermore, the list of responses on the activity questions were also integrated into follow up questions that required the students to drag and drop the list items into their respective rank order based on the importance of the activity. This approach allowed us to understand the students' academic needs, other academic (i.e., other courses) and non-academic commitments, and intentions of academic activities in detail and an opportunity to follow-up on the implementation rates of each activity.

Survey data consisted of quantitative and qualitative (open-ended) questionnaires. The survey data included an aggregate of 1343 survey variables (due to the repeated measures of variances of motivation and affective processes using the pre-post, weekly, and daily surveys) that are based on standardized questionnaires (Appendix A, Table A.1). The surveys measured a range of variables that included expectancy-value, self-efficacy, self-regulation, intrinsic and extrinsic motivation, goal-setting and implementation intentions, planned behavior relevant to the academic and non-academic activities, achievement goal orientation, task autonomy, task challenge, and emotions related to the course performance. The current analysis included survey data that measured the five constructs of interest that I have discussed earlier, due to their strong purported predictive value in online learning contexts.

### 4.2.3. Measures

**Demographic variables.** A wide range of variables were collected from the UCI's institutional records once the course was completed (after post-test). These measures included age, gender, low-income status, part-time status, first-generation status, race/ethnicity, SAT scores, and high school GPAs.

**Grades.** Students' performances were recorded from all the graded assignments which were provided by the instructor. These included lecture quizzes (18 which were part of lectures conducted online), reading quizzes (18 which were conducted as part of the reading materials or lessons assigned), weekly review quizzes (5 – comprehensive quizzes conducted each week), two midterms, and one final exams. Following the protocol from Study-2, we used weekly review quizzes for predicting weekly performances of the students, while reading and lecture quizzes were used as features. These were collected after post-test.

**Self-efficacy.** Students' self-efficacy was operationalized as a measure of confidence in taking the course using five items. Statements such as “I am certain I can master the skills taught in this course” and “I can do almost all the work in this course if I don't give up” were provided to the students who were then required to rate those statements on a scale of 1 (Not true at all) to 5 (Very true). Higher ratings implied better self-efficacy. These were measured the pre- and post-test surveys. This was included as a feature for predictions of performance on first quiz as well as for final exam. This measure was not included in the weekly quiz performance predictions.

**Self-regulation.** Students' self-regulation was measured using four sets of items: grade expectations, online expectations, online self-regulation, and online regulation.

Students' expectations pertaining to grades were obtained by 3 survey items. We asked students to report the grade they wanted, grade they expected, and the worst grade they could get that is considered acceptable. All grade expectations were obtained in the form of letter grades. Students' online learning expectations were obtained with two items. We asked students to rate their online learning performance expectations on a scale of 1 to 7, higher being better. Students' online-regulation was obtained from seven items where the students were asked to rate their self-regulation within online environments such as maintaining distraction free learning environment, tracking assignments on Canvas...etc., on a scale of 1 to 5, higher being better. Students' effort regulation was measured using 6 items, on a scale of 1 to 5, higher being better. We asked students regarding their effort and persistence through course work (see Appendix A, Table A.1). These measures were conducted during pre-, post-, weekly- and daily-surveys. Average of each weeks' scores were used to predict the review quiz performance for that week. All of the self-regulation measures were included in pre-, post-, weekly-, and daily-surveys.

**Grade achievement.** Students' final grades were subtracted from their expected grade to derive grade achievement measures.

**Motivation.** Following the Expectancy-Value Theory of motivation (Eccles, 2013), motivation was operationalized using the expectancies of success within the course and the values attached to the course. We measured utility value using five items, interest value using four items, attainment value using 3 items, and cost value using 3 items each for emotions, loss of valued alternatives, and outside effort. All of these items were adapted from Gaspard et al. (2015) and have been described by

McPartlan (2020). These measures were collected during pre-, post-, weekly-, and daily-surveys.

**Goal-setting.** Students' goal-setting were measured with items that asked about their study plan for the course (spacing or cramming intentions similar to Study-2), number of days in each week do the students intend to study for the course, total number of courses taken in that academic quarter, course importance compared to other courses being taken, other planned activities, time spent of academic activities, and time spent on non-academic activities. In addition, the students were asked to rank their course activities in relation to other courses and non-academic activities planned. We also asked the relative interest in the current course compared to other courses. We asked the students the amount of time they intend to spend on all other activities in a given week on a scale of 0 to 40 hours. Finally, we asked students about the completion rates of each task and used an aggregate measure based on these responses to understand task-completion rates as well as implementation of intentions. This acted as a proxy quantitative measure of students' self-efficacy and diligence towards tasks. This measure was collected at pre- and post-tests.

**Attributions.** Students' attributions were measured using 6 items at post-test. Statements such as "I see myself as someone who... does a thorough job/is a reliable worker/ tends to be lazy" were provided. Students rated themselves on a scale of 1 (strongly disagree) to 5 (strongly agree). All negatively-coded items' scales were reversed for consistency in the final analysis. For instance, 5 on a negatively-coded item was changed to 1. These measures were collected during pre-, post-, weekly-, and daily-surveys.

**Click data – total course activity per day.** Students' total clicks per day were measured by summing up the total clicks the students made on the Canvas course space.

**Click data – frequency of course activity per quiz.** We measured the frequency of clicks per assignment by counting the total number of clicks a student made on each web page specific to the review quiz each week.

**Click data – spacing.** We measured an estimate of “spacing” of each student's clicks by measuring the total number of clicks made each day of the week (Monday through Sunday). This acts as a proxy objective measure of spacing (clicks are spread throughout the week) or cramming (clicks are focused on a single day, typically on Tuesday since the weekly review quiz was due on Wednesday). We also included an alternate measure of spacing, by calculating the days between the assignment due date and first attempt to submitting that assignment. This measure also acted as a proxy for procrastination behaviors since, procrastinators (planned or otherwise) tend to submit the assignments in the last minute/hour (McPartlan, 2020).

**Other activities – open ended questions.** Apart from quantitative variables, the surveys also contained open-ended questions. Specifically, we asked the students to report course-related activities and other important non-course-related activities they planned to do. Since these open-ended questions lacked conclusive answers, we attempted to group free responses into several major categories through unsupervised clustering algorithms. Once the categories were found, we could label students' answers correspondingly, turning indeterminate variables into categorical variables that would replace each open-ended response with a category that the open-ended

response would represent. These categories were used for training the predictive models. To accomplish this goal, we adapted two popular unsupervised clustering methods: 1) KMeans and 2) Latent Dirichlet Allocation.

*KMeans* is a classic distance-based unsupervised clustering method, first introduced by Lloyd (Lloyd, 1982). Based on the pre-selected K number, KMeans will randomly assign K partitions each with a random centroid, which are iteratively reset by reassigning each point to the nearest center (by calculating distances between each data entry to the centroids), reassigning data to its nearest centroid, and then updating the centroid locations until the centers do not move any further. Once KMeans converges, each data will be allocated to a group respective to its assigned centroid. The algorithm is considered “straightforward” but selecting the K number of centroids and validating the clusters is necessary. In order to find the optimal K, we fit the KMeans model on the preprocessed (see below) open-ended questions using K number ranging from 2 to 8, as we did not expect more than 8 major activity categories after reading the open-ended responses. We derived sum of squared errors for each K and picked four clusters based on the “elbow-effect”, similar to a scree-plot in PCA. We validated the clusters by generating word cloud plots of the four identified clusters for to check whether the words in those clusters are human-interpretable.





the best coherence measure performance,  $C_v$ , as discussed by Roder et al (2015). After deriving the coherence scores for each of the  $K$  topics, the  $K$  number associated with the highest coherence score was selected.

Before fitting the models on the raw students' responses, we employed several NLP preprocessing techniques on the open-ended response text data. For the sake of consistency, we lower-cased and lemmatized every word (e.g. "Better" will be converted to "good"). To avoid extremely low- and high-frequency words, we only considered words that appeared at least 5 times and at most 500 times among all responses following the protocol described by Alteras and Stevenson (2013). We also removed commonly used non-informative words or "stop words", such as "a/an" and "s/he/it." Lastly, we took advantage of Tf-idf (term frequency-inverse document frequency) to weigh each word according to their relevance in each students' response, so that a common word would weigh less than a rare word in one sentence (Ramos, 2003). After pre-processing the open-ended question responses, we followed the procedures of

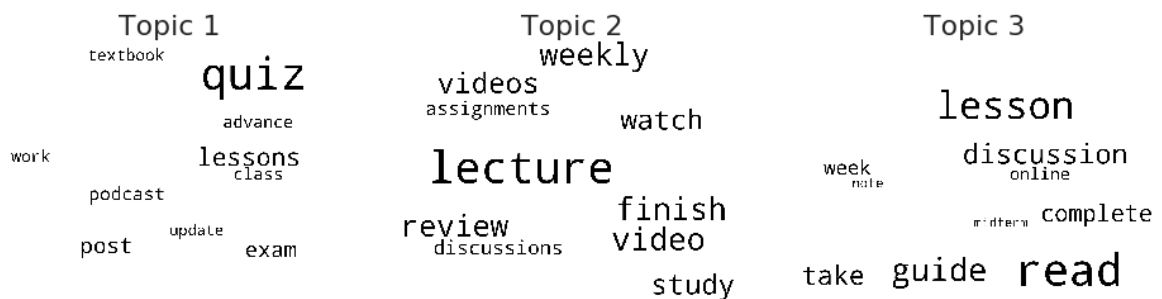


Figure 4.3. LDA generated word clouds -- for students' course-related activities. LDA found three optimal topics.

each unsupervised clustering method discussed above and obtained four clusters using LDA, similar to Kmeans. Results of both approaches are shown below - Figure 4.2.(a) for KMeans clusters and 4.2.(b) for LDA topics.

For “other activities” open-responses, of the four clusters KMeans produced, cluster 1 was significantly larger (~1500 responses) than the other three clusters (~250 responses each). Though there were overlaps, we were able to reliably label cluster 1 as “Class/Friend”, cluster 2 as “Work”, cluster 3 as “Personal/Family”, cluster 4 as “Another Course”. In contrast, LDA generated much cleaner word clouds because only important words in each topic were plotted. Similar to KMeans, we generated 4 labels/categories for the topics in the same order as the clusters. While both approaches derived the same number of clusters and same topic models, since LDA provided relatively less noisy data, we used the results from LDA to label our data.

Following a similar approach, we were able to generate a variable with three categorical labels for “course related activities” of students using the LDA method. These clusters were labeled, “Quizzes”, “Lecture videos”, and “discussion/reading guide” (Figure 4.3).

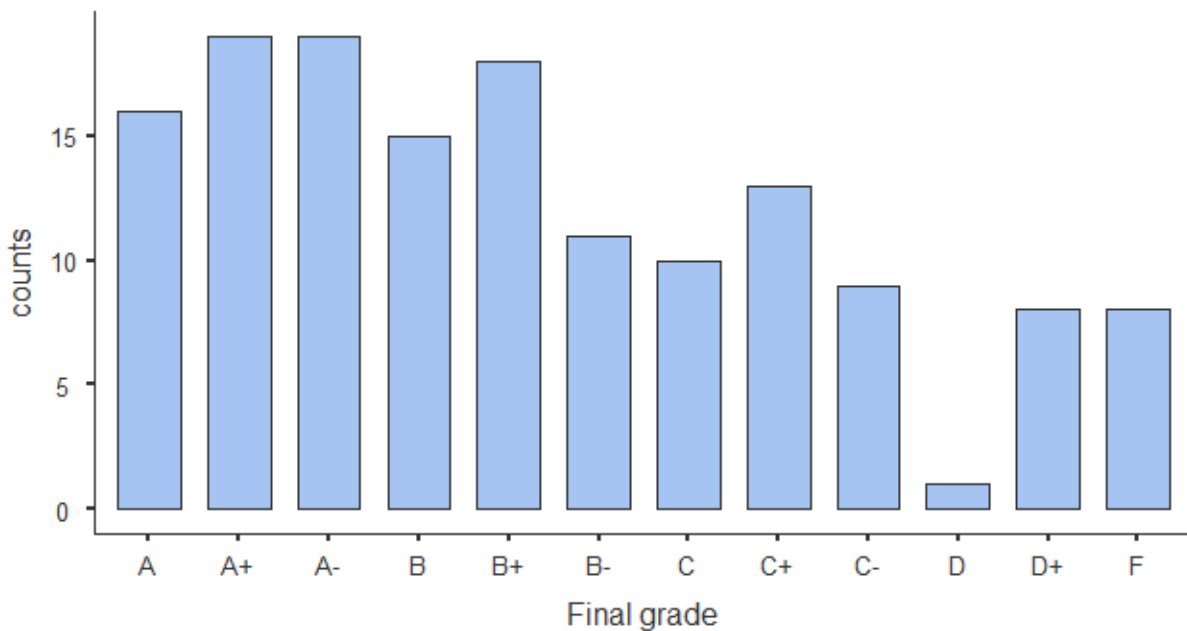


Figure 4.4. Final grade distributions for Study-3

#### 4.2.4 Outcome Measures or Target Variables

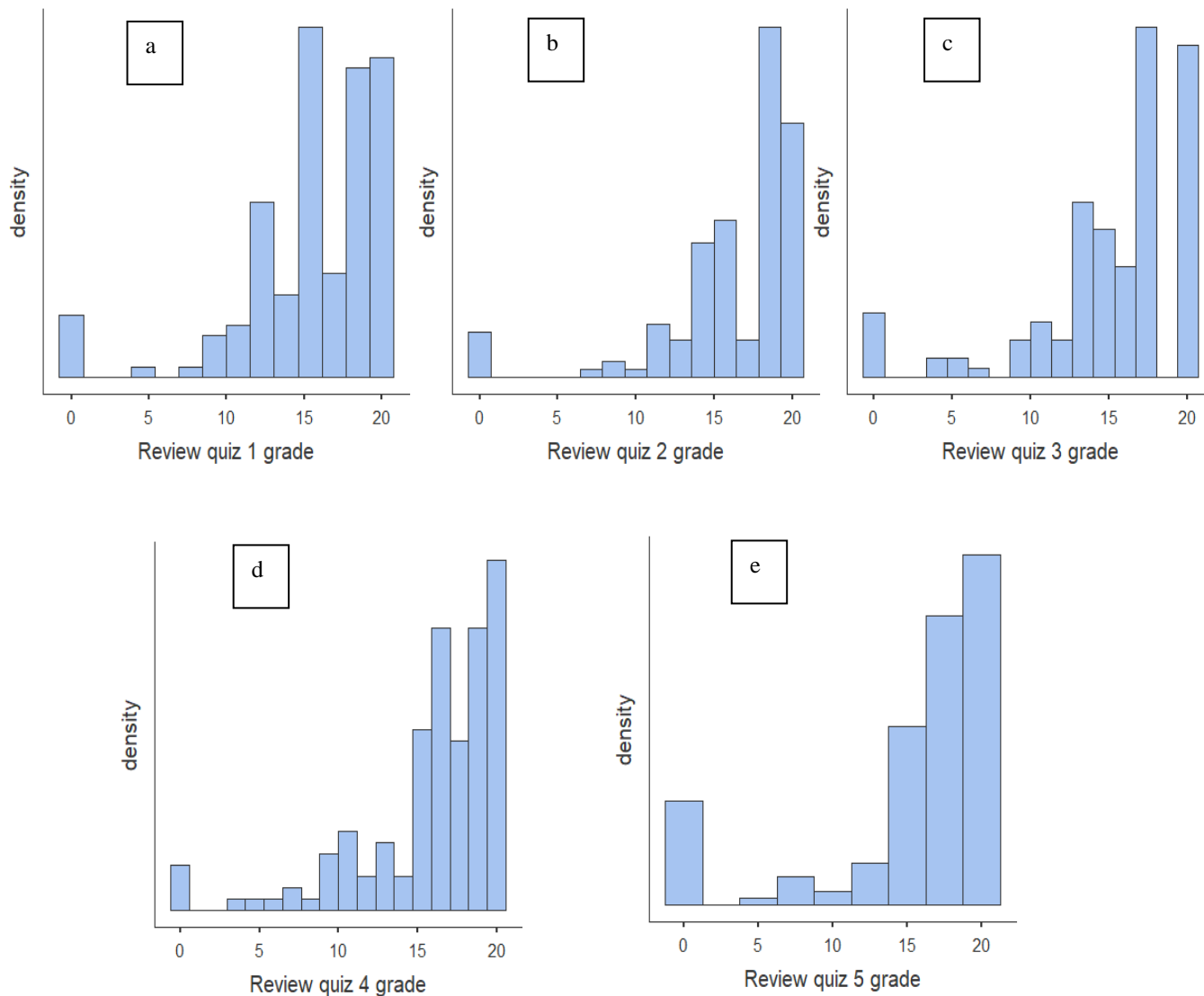


Figure 4.5. Weekly review quiz score distributions -- from week 1 through week 5 (Panels 'a' through 'e').

Similar to Study-2, there were four outcome measures of interest (two each for regression and classification models): (a) Students' performances on weekly review quizzes. The review quizzes were held on Wednesdays each week (5 in total) with a maximum possible score of 20 points each. This acted as our intermediary prediction

targets for regression models, much like predicting session by session performances in WM training from Study-1, (b) above or below median performance on weekly review quizzes. Students were divided into two equal groups for our classification modeling of each week’s performance, (c) Final scores – which were a weighted sum of performances across the course with 100 being the maximum possible points, and (d) above or below ‘B+’ final grade consistent with Study-2. The distributions of final scores are shown in Figure 4.4. Individual review quiz score distributions are shown in Figure 4.5 panels (a) through (e) respectively for each week’s review quiz. Table 4.3 shows the mean, median, standard deviation, and ranges for all 5 review quizzes, averages for the average reading quiz scores, average review quiz scores, and average lecture quiz scores.

Table 4.3.

*Descriptive statistics for the review quizzes, average reading quiz and lecture quiz scores.*

	Review quiz 1 grade	Review quiz 2 grade	Review quiz 3 grade	Review quiz 4 grade	Review quiz 5 grade	Average reading quiz score	Average review quiz grade	Average grade on lecture quizzes
N	147	147	147	146	147	146	146	147
Missing	22	22	22	23	22	23	23	22
Mean	15.6	16.3	15.4	16.0	15.7	5.10	15.9	3.68
Median	16	18	17	17.0	18	5.39	16.6	4.00
Standard deviation	4.43	4.33	4.94	4.51	6.06	0.94	3.19	0.69
Minimum	0	0	0	0	0	1.51	3.60	0.00
Maximum	20	20	20	20	20	6.00	20.0	4.00

### 4.3. Research Questions and Hypotheses

As discussed earlier, studies have shown that dynamics of motivation and mood are linked to processes of learning, decision making, and changes in behavior (Madigan

& Bollenbach, 1986; Eldar, Rutledge, Dolan, & Niv, 2016). To understand the integrative behavior of learning as a dynamic process, it is important to understand the nature of the dynamic processes that are associated with performances of learners (Cattell & Child, 1975), and to predict the trajectories of learners accurately. In the review of the literature relevant to the effects of all the motivation social-emotional, and affective dimensions listed in Table 4.1, I discussed the two important shortcomings of the existing work – the prevalent use of click behaviors to predict learning outcomes disregarding the value of context-agnostic click behaviors, and the lack of studies that sought to understand multiple dimensions of dynamics of the discussed factors together, over the learning period, to predict learning over the course (Mayer, 2014). Unlike human personality, motivations and moods of individuals vary by situation, by day, by experience, which in-turn effect the learning and performance of individuals (Dayan & Daw, 2008). Therefore, the current work will focus on understanding the dynamic nature of these dimensions to predict the learning outcomes throughout the learning period. In the end, the goal of the current work will be to identify learners who have relatively stable performances overtime despite the variances in motivations which might indicate that these types of learners might be rigid to changes in motivations and vice versa (to detect highly labile learners whose performances may change with change in motivation and affective processes). In Study-3, I will build on the existing models from Study-2 to understand the predictive value of the intraindividual variabilities in these dimensions of interest that may facilitate better predictions. To accomplish this goal, Study-3 is driven by three research questions that explore the predictive value of I seek to answer the following research questions -

**RQ 1** – To what extent can we predict students' performance on weekly review quizzes and final grade using demographics and reading and weekly quiz scores?

**RQ 2** – Can click behaviors improve predictions of models from RQ (1)? If so, which of the click behavior features has the highest predictive value?

**RQ 3** – Can motivational and affective measures and their dynamics improve predictions of models from RQ 1 and RQ (2)? If so, which of the motivational and affective features carry the highest predictive value?

Addressing these RQs have several purposes. First, since we seek to validate our results from the earlier two studies, our focus in RQ 1 is to understand if we can predict later learning from early learning trends on a weekly basis as well as use the weekly trends to predict the final grades. As before, we will use two different models, one with and one without demographics, to understand if demographics have predictive value beyond actual performances in the course. I hypothesize that demographics carry high predictive value during early learning predictions, but later learning can be predicted from early learning trends alone (i.e., without the need for demographic information) since demographics carry diminishing returns for predictions in our studies-1 and 2. Next, our focus in RQ 2 is to understand if click behaviors can be used to boost our predictions from RQ (1). This model will be similar to the third model from Study-2 (i.e., model which includes demographics, past performances, and in addition, measures of click behavior). I hypothesize that click behaviors have very little value in prediction accuracies since the measures of clicks, as I have argued earlier, do not provide any context to students' learning nor do they act as a good behavioral indicator of students' changing motivational, social-emotional, and affective needs and demands.

Finally, our focus in RQ 3 is to understand if the survey metrics that directly ask the students about their own motivational, social-emotional, and affective processes will add predictive value. This fourth model will illuminate the value of collecting survey data from students throughout the learning process, specifically, about their learning habits, other priorities, task values, self-efficacy, and emotions associated with success in course towards predicting their learning. Since our predictive models are going to include all these features, we have the opportunity to understand their relative importance in predicting their learning on a weekly basis and in predicting their overall grades.

#### **4.4. Analytical approach.**

Study-3 follows a similar protocol for data analysis as discussed in Study-1 and Study-2. Due to the nature of the target variables discussed earlier, we used: 1) regression models to predict students' weekly review quiz performances and final scores, and 2) classification models to classify students as "above" or "below" median performers for the weekly quiz performances and "above B+" or "below B+" for students' final grades. In general, the regression models estimate the relationship between the target and explanatory variables by fitting a curve to the data points so that the distances between the curve and target data are minimized. Specifically, we started with a multiple linear regression model and compared the results with a 2<sup>nd</sup> order polynomial regression to account for variable interactions. On the other hand, classification models project the selected features into a hyper-dimensional feature space, generating a hyperplane to classify the data into required categories, specific to the problem (i.e., "below B+" vs "above B+" or "below" vs "above.") We tested the two identified binomial



classification models - Logistic Classifier (LC) and Random Forests (RFs) following the WEKA toolkit protocol discussed earlier to compare these algorithms (Ian H Witten et al., 2011). The optimal regression/classification models were selected based on a two-step performance criteria: a) three-fold cross-validation: the data were randomly split into three equally-sized groups (33 students each), where one group was held out as the testing set and the other two were used as the training set at a time repeated until each group had become the testing set once. Then, we used the average of several metrics described earlier (i.e., RMSE, adjusted  $R^2$  for regression models and Accuracy, Precision, Recall, and F1-score for classification models) on the testing set to determine the robustness of the models. b) The data were shuffled randomly at 50 unique seed locations (the point of reference for the division of the two subsets listed above) followed by the repeated measurement of each of the metrics 50 times. The two-step approach was used to ensure the predictions' robustness and generalizability to new datasets that are similarly structured (Iguyon & Elisseeff, 2003; Tang et al., 2014).

First, we collected raw, declassified data into csv files and preprocess the data (i.e., cleaning the formatting inconsistencies, and dropping missing data that do not have results on key outcome variables listed below). Then we extracted the preliminary features of interest (such as categorizing and making dummy variables of the outcome measure into "above B+" or "below B+" performers.) Next, we sampled the data by dividing the data using the 3-fold approach discussed earlier. Cleaning of data was performed using a three-step approach discussed in the previous chapter. In summary, all features with more than 10% missing values were dropped, features with low variance were dropped, features with high correlations were collapsed into a single

variable (for instance, since all five items in the self-efficacy questionnaire showed very high bivariate correlations (>80%) we used the average self-efficacy as a feature instead of the individual features). In addition, the dimensionality of the datasets is reduced by dropping features with no predictive value and those that have no mutually exclusive information (for instance, dropping standardized English language test scores which were only available for 2 students). This further reduced the number of features used to build the prediction models for the listed RQs. Furthermore, we standardized all the performance measures for pre-lecture quizzes, reading quizzes, weekly review quizzes, mid-terms, and final exams into percentages for consistency since the scales for these variables were different. The final steps of analysis included training, hyperparameter optimization (correcting for false positives and false negatives using a 5% validation dataset from the training dataset), and post-processing (final model selection and calculating performance metrics for the selected model) and then testing and evaluating the model separately for RQ (1), RQ (2), and RQ (3). All of the features used to answer the three RQs are provided in Table 4.4. There were a total of 5 models that were used to answer the three RQs. First model was a Baseline model that predicted learning on each week's review quiz using randomly generated noise around the true mean and standard deviation of the average quiz performances in week 1. Since there is no real data to predict from, similar to Study-1 and Study-2, the resulting accuracies and Adjusted R-squared were poor. Next, to answer RQ 1 we used two models – 'a' and 'b' – similar to the two previous studies. RQ 1 – model a used only the quiz scores to make predictions. RQ 1 – model b used quiz scores as well as demographic information. We expected that the model b would significantly outperform

model a first the first three weeks of the predictions. Models a and b predictions were expected to be similar after converging at third week. Next, to answer RQ 2, we used a model that included everything from RQ 1 – model b and study intentions such as spacing and click behaviors, similar to Study-2's final model. We expected the results to be similar or slightly better (if the click behavioral measures positively predicted the performance unlike in a hybrid classroom). Finally, to answer RQ 3, we used the model from RQ 2 and included the dynamics of learner-centric measures. We expected that this model will yield the highest predictive accuracies compared to the rest of the models since understanding learner-centric features are known to be directly relevant to their learning outcomes. All five models used data from week 1 quiz scores, survey responses, and pretest-only measures for predicting week 1 performances in the first step. The 2<sup>nd</sup> step included data used for 1<sup>st</sup> step as well as week 2 quiz performances and survey responses, and so forth. During week 3, in addition we included the daily survey responses for each of our learner-centric features of interest. Validation accuracy of Random Forest model was higher (96%) than that of Logistic Classifier (89%). Therefore, we used RFs for all further classification analyses. Multiple linear regression resulted in slightly better results overall compared to the non-linear model. Thus, all results of regressions presented below are from the multiple linear regression. All of the analysis and results are also available in the Github repository link provided at the end of this document. All results reported below are for the testing dataset only (since the training dataset prediction accuracies and R-squared are not a good indicator of model performances). All testing dataset predictions were consistently within 10% of the training dataset predictions, implying no overfitting of models occurred.

Table 4.4

*Models tested and list of features included in each model variant.*

model	Baseline model	RQ 1 - model a	RQ 1 - model b	RQ 2	RQ 3
overview	(random data generated using average quiz scores and standard deviations)	(students' performance on tests)	(RQ 1 model a + demographics)	(RQ 1 model b + spacing intentions and click behaviors)	(RQ 2 + dynamics of learner centric metrics)
features	random noise	average weekly lecture quiz performance	average weekly lecture quiz performance	average weekly lecture quiz performance	average weekly lecture quiz performance
		average weekly reading quiz performance	average weekly reading quiz performance	average weekly reading quiz performance	average weekly reading quiz performance
		midterm performance	midterm performance	midterm performance	midterm performance
			age	age	age
			gender	gender	gender
			low-income status	low-income status	low-income status
			part-time status	part-time status	part-time status

model	Baseline model	RQ 1 - model a	RQ 1 - model b	RQ 2	RQ 3
			first-generation status	first-generation status	first-generation status
			high school GPA	high school GPA	high school GPA
			SAT scores	SAT scores	SAT scores
				study spacing intention	study spacing intention
				change in study plan	change in study plan
				click-data total course activity till date	click-data total course activity till date
				click-data frequency of course activity per quiz	click-data frequency of course activity per quiz
				click-data spacing	click-data spacing
					grade expected
					grade expectation achievement
					number of other courses
					course importance rank

model	Baseline model	RQ 1 - model a	RQ 1 - model b	RQ 2	RQ 3
					course interest rank
					time spent on course
					time spent on other activities
					activity completion rates
					average self-efficacy
					average online self-regulation
					average effort regulation
					average utility value
					average interest value
					average attainment value
					average cost value
					average attributions

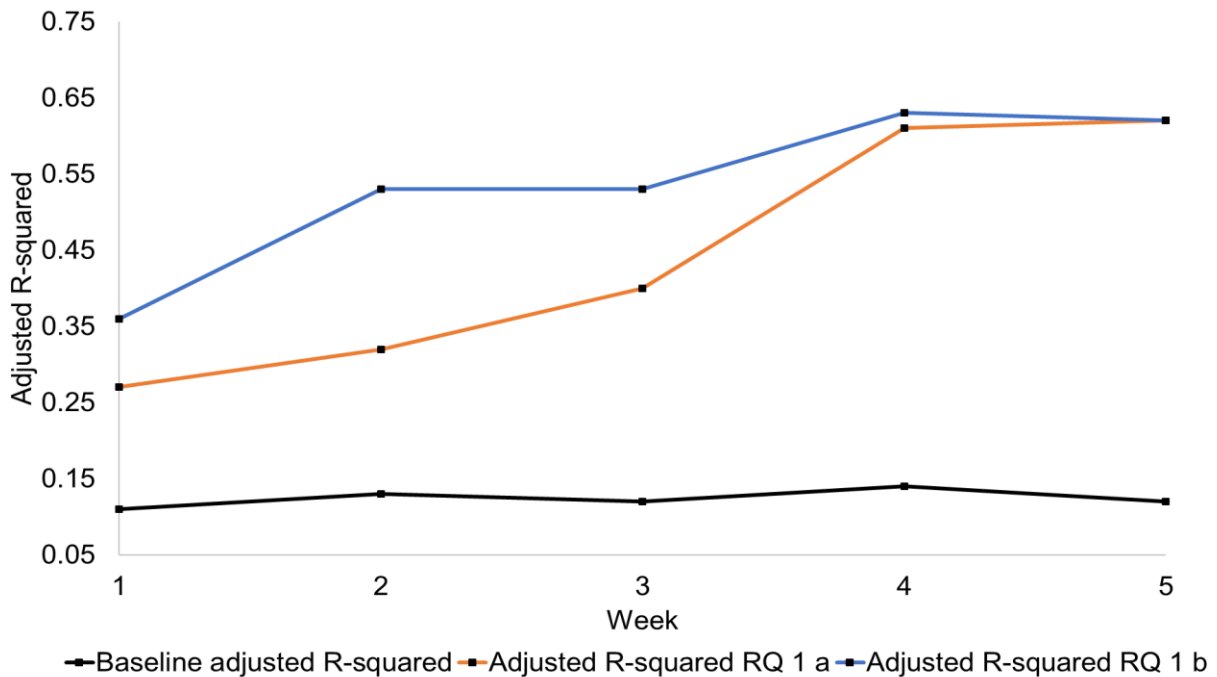


Figure 4.6. Results from the regression models for RQ1 -- predicting the weekly review quiz scores using the features listed for RQ 1 - model a and RQ 1 - model b in Table 4.4.

## 4.5. Results

**RQ 1 –** To what extent can we predict students’ performance on weekly review quizzes and final grade using demographics and reading and weekly quiz scores?

Figure 4.6 shows the results of the multiple linear regression model that is used to answer RQ 1 (regression problem). Baseline model is shown in black. As expected, the baseline model performed poor compared to the other two models (RQ 1a shown in orange and RQ 2 b shown in blue). As expected, RQ 1 b performed better than both RQ 1a and Baseline models during the first three weeks. However, convergence did not occur until week 5. The maximum adjusted R-squared achieved by baseline model was

0.14, whereas the maximum adjusted R-squared achieved by RQ 1a and RQ 2 b were 0.62 and 0.63, respectively for week 4 predictions.

Figure 4.7 shows the results of the RF model that is used to answer RQ 1 (classification problem). As before, Baseline model is shown in black. As before, the baseline model performed poor compared to the other two models with prediction accuracies close to chance around 48% for all 5 weeks. Classification accuracy for RQ 1a is shown in orange and RQ 2 b is shown in blue. Akin to the regression problem, RQ 1b outperformed both RQ 1a and Baseline models during the first three weeks. However, convergence did occur at week 4 after which there was no difference between the two models' performance. The maximum classification accuracy achieved by baseline model was 0.48, whereas the maximum classification accuracy achieved by RQ 1a and RQ 2 b were 0.78 and 0.79, respectively for week 5 predictions. In addition, the final grade classification accuracies (which used data from posttest survey data and average performance on all 5 review quizzes) for RQ 1a and RQ 1b were 0.80 each and the adjusted R-squared was 0.65.

Overall, the results from these three models shed similar light as the two studies discussed earlier. Baseline models that do not use any actual learner performance data tend to do poorly compared to the rest of the models. This is expected and required for the approach being used to be valid for further modeling. If the results of the Baseline model are (abnormally) high, there would be no validity for the results from the rest of the results. Next, results suggest that the trends we have seen from the other two models persist. During the early learning phase, since the predictive models cannot learn from the students' performances (for instance, the variance from the performance



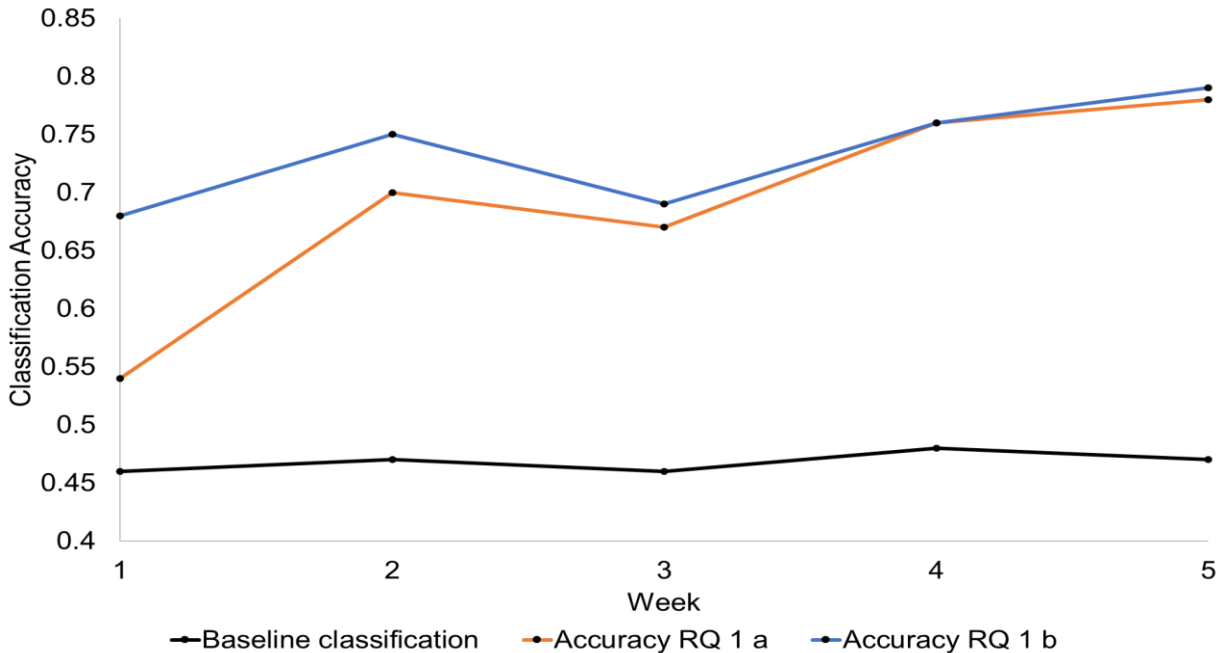


Figure 4.7. Results from the classification models for RQ1 -- predicting the weekly review quiz scores using the features listed for RQ1 model a and RQ1 model b in Table 4.4.

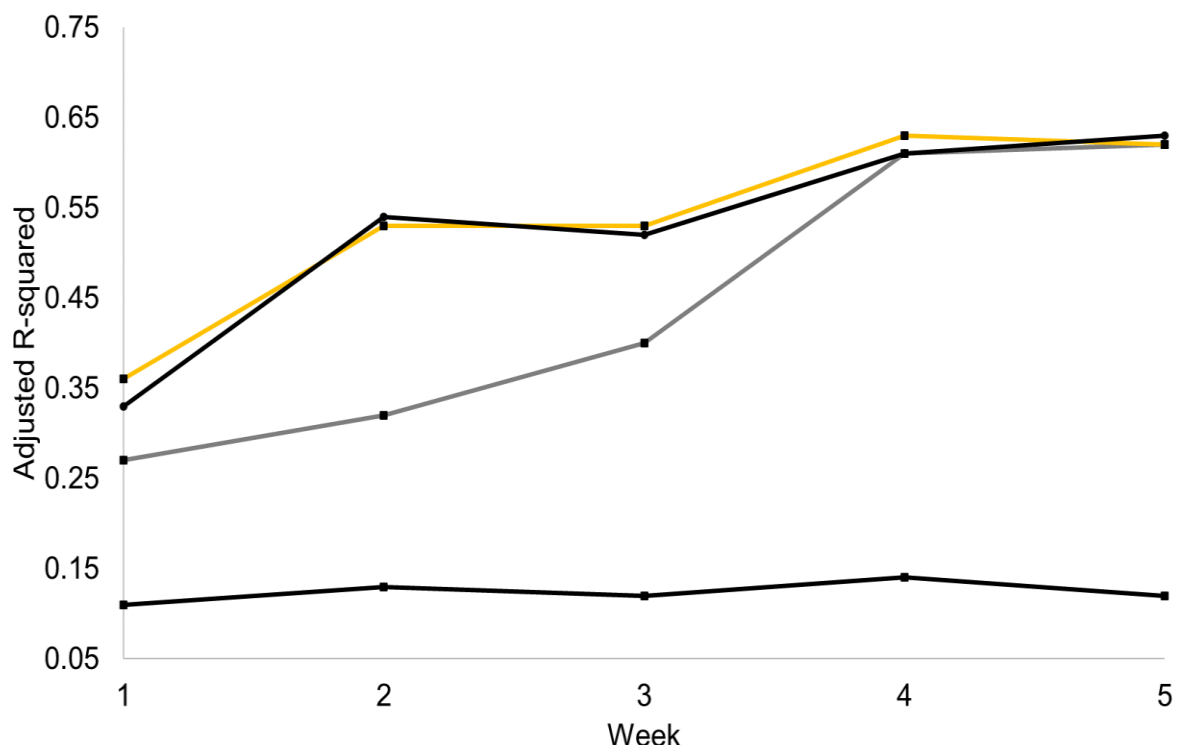
of week 1 alone are insufficient to reliably predict later learning), predictive model RQ 1b seeks to find associations between the demographic features to the performances in order to better predict the students' quiz scores or to classify them as "above" or "below" median performers. Once the models tend to achieve a better understanding of each student's performance, the results of the model without these demographic data tend to keep up with the model with demographic data. If we assume that the students' early learning is a result of the demographics and the amount of subject knowledge that they bring to the table, it is reasonable to hypothesize that the model with this critical information tends to outperform the model without these data. However, once the predictive models learn and relearn their classroom performances, the predictions become as reliable as the models with the demographic data. These results suggest

that, knowing the background of a learner is critical to making reliable and better predictions of their later learning.

Results show that the maximum prediction accuracy achieved by the two models is highest towards the end of the 5-week learning period at around 79%. This is significantly lower than the prediction accuracies on the WM training data (90%) halfway through the learning period. This supports our hypothesis that making predictions of learning within a skill-acquisition context is relatively easier compared to classroom performance due to constraints imposed by the classroom learning scenario (such as prior knowledge, subject mastery, relevance of course to career, social structure, emotional, and affective processes being different.) However, the prediction accuracies are also slightly lower compared to a hybrid classroom (82%) when our predictive model was given similar information. This could be due to two different reasons - (a) the hybrid classroom performances predicted included a larger dataset (460 compared to 99 students in this dataset), and (b) the course for which the analysis was conducted in Study-2 was held for 10 weeks as opposed to 5-week long course in this dataset. Both affect the criteria for optimal performance of machine learning approaches where higher amounts of data available to learn tend to yield better predictions. Overall, a theme apparent from the results of these three models across the three studies suggest that demographics or any theoretically relevant information other than the learners' performances alone boost early prediction accuracies. Sometime during the stepwise training approach, typically halfway through the learning period, the apparent benefits of such additional features fade out due to diminishing returns to making predictions of later learning.

**RQ 2** – Can click behaviors improve predictions of models from RQ (1)? If yes, which of the click behavior features has the highest predictive value?

Let us now focus on the results shown in Figures 4.8 and 4.9. First, Figure 4.8 shows the results for the regression model RQ 2 listed in Table 4.4. We left the models used to answer the questions earlier for comparing the results of those three previous models with the results for model RQ 2 (shown in yellow). Recollect that similar to Study-2, here we included measures of study intentions and click behavior metrics generated from the



*Figure 4.8.* Results from the regression models for RQ2 -- predicting the weekly review quiz scores using the features listed for RQ 2 in Table 4.4 (shown in Yellow color). Models from RQ 1 discussed earlier are faded out but left in for comparison.

students' LMS activity. Similarly, the same set of features were used for the classification problem. Results for the classification models for RQ 2 shown in Figure 4.9 (in yellow) also show the previous three models for comparison. Results for model

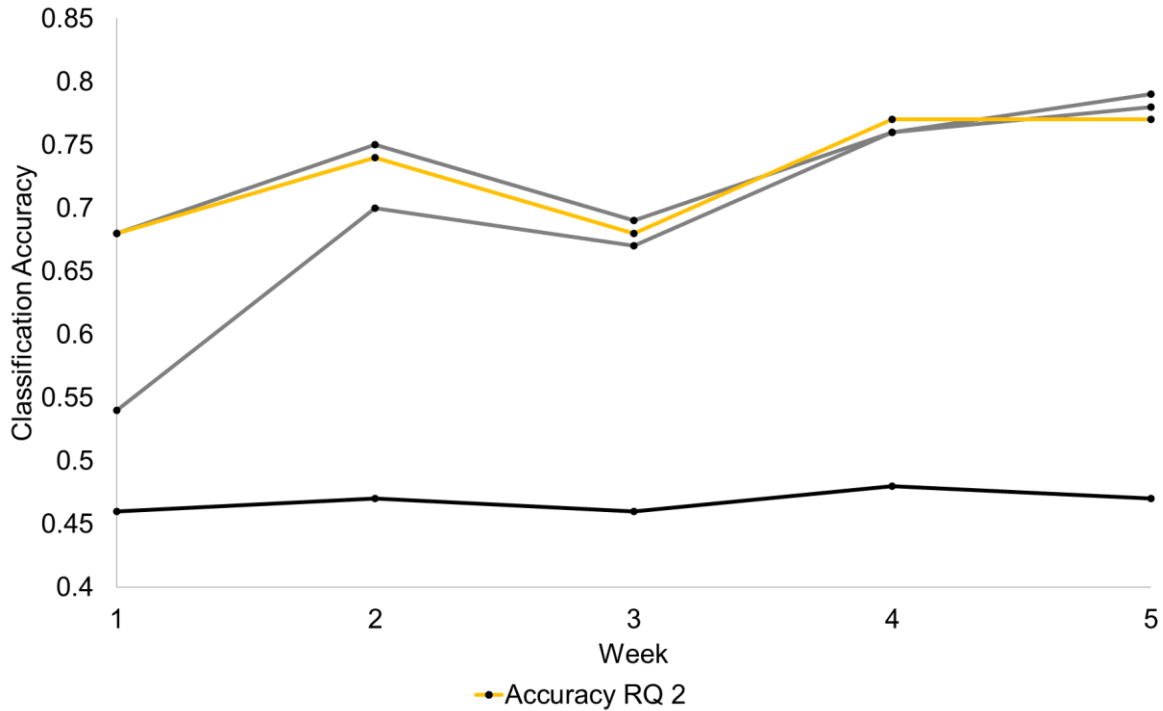


Figure 4.9. Results from the classification models for RQ2 -- predicting the weekly review quiz scores using the features listed for RQ 2 listed in Table 4.4 (shown in Yellow color). Models from RQ 1 discussed earlier are faded out but left in for comparison.

RQ 2 were as predicted. Final grade predictions were the same as week 5 were similar to the predictions of RQ 1 b. Students' study plans and intentions and their click behaviors within the LMS do not improve our prediction accuracies over the RQ 1b model. The results of RQ 2 follow a similar pattern for both classification as well as for regression models. However, note that the results are not worse than that of RQ 1b model. The approach taken to aggressively punish false positive and false negative predictions (L2 regression-based corrections) that we conducted in Study-2 were also utilized here to keep the models from overfitting on the training data. Overall, the model accuracies neither improved, nor significantly worsened compared to our predictions with demographics and learners' weekly performances. These results suggest that using click behaviors in an online classroom without the context of knowing what the

click might mean will yield poor prediction accuracies. This also supports previous works that have shown that there are no significant predictive values for click behaviors (see for e.g., McPartlan, 2020).

**RQ 3** – Can motivational and affective measures and their dynamics improve predictions of models from RQ 1 and RQ 2? If so, which of the motivational and affective features carry the highest predictive value?

Next, to answer RQ 3, we used all the features for RQ 2 and included the survey metrics for learner-centric measures listed in table 4.4 to predict the students' weekly review quiz scores. The results for regression model are shown in Figure 4.10 (in Green color) and results for classification model are shown in Figure 4.11 (in Green color).

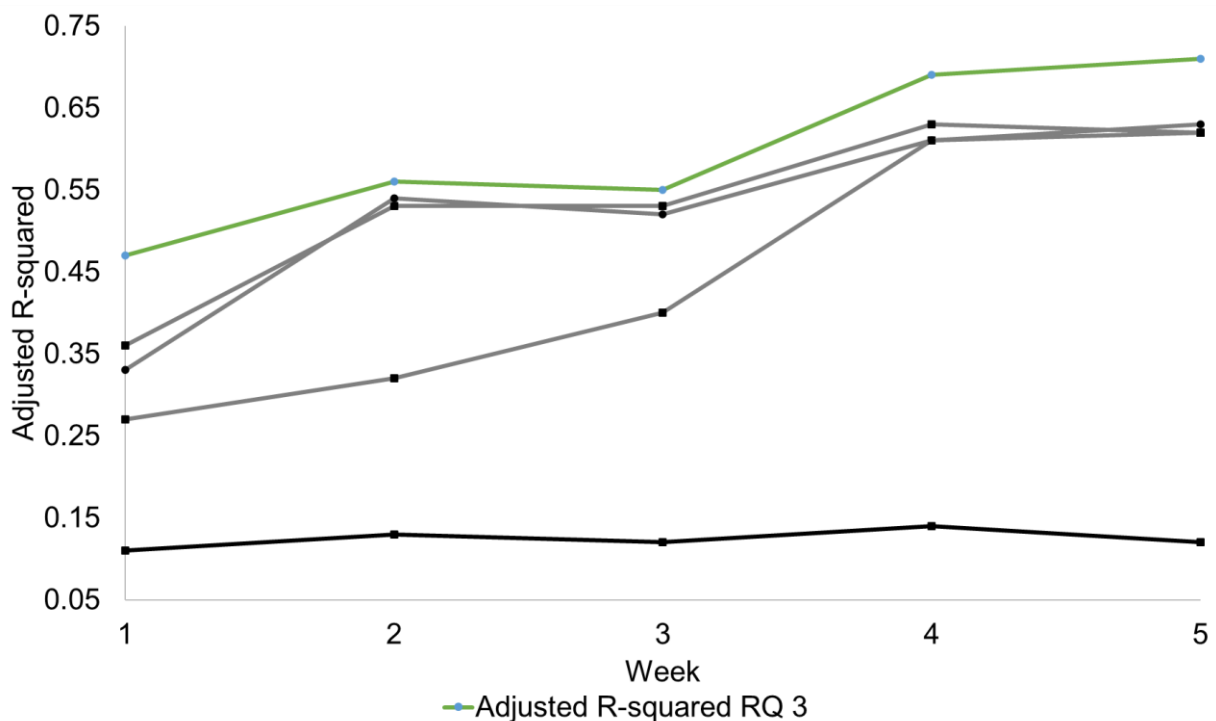


Figure 4.10. Results from the regression models predicting the weekly review quiz scores using the features listed for RQ 3 in Table 4.4 (shown in Green color). Models from RQ 1 and 2 discussed earlier are faded out but left in for comparison.

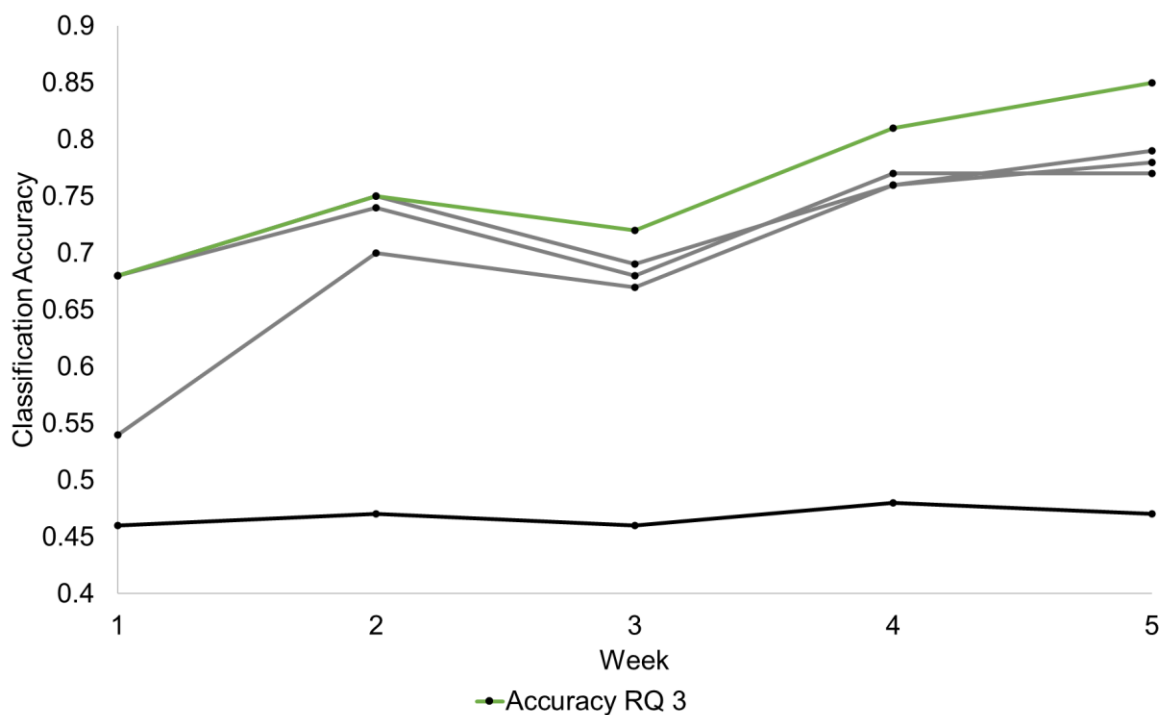


Figure 4.11. Results from the classification models predicting the weekly review quiz scores using the features listed for RQ3. Results for models RQ1 model a and RQ1 model b in Table 4.4 are left in but faded out for comparison.

Note that all four models from RQ 1 and 2 are left in and faded for comparison. Models for RQ 3 performed the best overall.

At week 1, the model had the highest performances for both regression as well as classification models with 0.47 adjusted R-squared and 68% accuracy, respectively. This trend continued for all five weeks review quiz predictions. Notice that there was a dip in performance of our models during week 3. This could be due to the higher number of features which could lead to increased noise during week 3 due to the inclusion of daily survey responses in the prediction models. The prediction accuracies bounced back during week 4 since the extra features were only included during week 3 predictions. However, RQ 3 model outperformed all other models even during week 3,

albeit, by a very small amount (3% better than RQ 2 model). Overall, the results shown follow our hypothesized results with 81% prediction accuracy by week 4. Including the survey responses as features improved our prediction accuracies significantly. In addition, the final grade classification accuracies (which used data from posttest survey data and average performance on all quizzes) over the 50 iterations for model 3 were 0.88 (+/- 0.07), respectively and the adjusted R-squared was 0.81 (+/- 0.06) with RMSE of 6.61 (+/- 3.12). The overall average precision (0.89), recall (0.91), F1-score (0.90), and AUC (0.89) were close to the levels of accuracy indicating that the overall model performance was acceptable.

Table 4.5

*Feature importance derived for weekly review quiz predictions and final grades.*

<b>Feature</b>	<b>Feature importance - Week 1</b>
Average Self-Efficacy	0.21
Average Attainment Value	0.15
Highschool GPA	0.11
Average Cost Value	0.07
Course Rank	0.06
	<b>Feature importance - Week 2</b>
Review Quiz Week 1 Grade	0.31
Number of Other Activities	0.19
Average Cost Value	0.13
Grade Expectations	0.11
Goal Completion Rate	0.09
	<b>Feature importance - Week 3</b>
Review Quiz Week 2 Grade	0.32
Review Quiz Week 1 Grade	0.15
Average Cost Value	0.09
Grade Expectations	0.08
Goal Completion Rate	0.08
	<b>Feature importance - Week 4</b>
Review Quiz Week 3 Grade	0.32

---

Review Quiz Week 2 Grade	0.19
Grade Expectations	0.11
Average Cost Value	0.09
Course Rank	0.09
<b>Feature importance - Week 5</b>	
Review Quiz Week 3 Grade	0.29
Review Quiz Week 4 Grade	0.19
Grade Expectations	0.13
Review Quiz Week 2 Grade	0.11
Goal Completion Rate	0.09
<b>Feature importance - Final Grade</b>	
Average Review Quiz Score	0.34
Grade Expectations	0.24
Goal Completion Rate	0.16
Average Utility Value	0.06
Average Cost Value (Emotions, Effort, Attainment Combined)	0.02

---

Next, to investigate which features contributed most to our predictions, we used the feature importance derived from the RF models for each week and retrieved the top 5 features for each week as well as for the final grade predictions (Table 4.5). Note that importance of all variables listed are for the classification model. Self-efficacy carried the highest variable importance during week 1 followed by attainment value, high school GPA, and cost value with rank assigned to the course being the fifth most important feature. Self-efficacy, a measure for how confident the student is in their ability to tackle the task was identified as the most important feature. This follows the theoretical expectations given that self-efficacy is demonstrated to be a strong predictor of students early learning and dropout rates (Kuo et al., 2014). Attainment value and high school GPA were the two other features that carried the highest variable importance. For week 2, the topmost feature was the students' performance on review quiz-1 indicating that relevant past performances carry the highest predictive value as we have seen in Study-1. Total number of other activities, a proxy for other academic and non-academic



responsibilities the students have, carried the second highest variable importance, followed by Cost associated with course (emotional, outside effort, and loss of valued alternatives combined). Grade expectations (grades expected for the course) and goal completion rate (how many of the activities from past weeks the students were able to successfully complete), took the next two spots. During week 3, as before, past performance on review quizzes (week 2 followed by week 1) carried the highest feature importance followed by the cost values, grade expectations, and goal completion rate. During week 4, performance on review quizzes (week 3 followed by week 2) carried the highest feature importance followed by grade expectations, cost value, and the rank of the current course among other activities. Features of importance for week 5 were similar, where, review quiz scores for week 3, 4, and 2 took the ranks 1, 2, and 4 respectively. Surprisingly, performance for week 3 had a higher feature importance than performance for week 4 (0.29 to 0.19). Grade expectations and goal completion rates were also in the top 5 list.

Overall, for week 1, despite the predictions of performance being lower, self-efficacy and attainment values carried the highest predictive value. High school GPA was only important for week 1 predictions. None of the background or demographic features made it to the list. Perhaps, the boost in performance of model RQ 1b over RQ 1a might be driven by the students' high school GPAs. Results indicated that cost values, grade expectations, and goal completion rates were consistently important for predictions across all 5 weeks indicating that it is very important to understand the values and cost to predict learning rather than using click behaviors, which did not make the top 5 (or top 10 – not shown) list across the predictions. Furthermore, it is critical to

establish the expectations that the students have. Perhaps, the expected grades tie into students' behaviors such as planned procrastination or diligence in task completions, the latter also being one of the top 5 features consistently after week 1. Finally, the top 5 features of importance for the overall final grade classification model used in RQ 3 were Average review quiz scores, grade expectations, goal completion rates, average utility value, and cost values. Yet again, the predictions were driven by learner-centric measures rather than the click behaviors. Furthermore, students' learning predictions were not driven by features such as SATs or high school GPAs indicating that relying on past performances alone while making future predictions might lead to poor predictions.

#### **4.6 Discussion**

In this study, we attempted to predict learning using a stepwise approach (predicting performance on weekly quizzes) using features that are learner-centric in addition to click behaviors. Research reviewed earlier has suggested that predictive models employed for future predictions in learning scenarios including in online learning settings, typically yield binomial classification accuracies between 65%-77%. Except for a very small number of the reviewed studies, most of the existing literature used click behaviors as a proxy for many human behavioral traits such as engagement, motivation, procrastination. While these endeavors are important to improve the LMS platforms (e.g., for making it easier to understand the students' engagement within the learning platform), results indicated that these artificial sensors are woefully short of capturing the true learning behaviors and engagement (Aldowah et al., 2019; McPartlan, 2020). The results of the current work yielded 81% accuracy roughly mid-way through the course timeline (just before the second midterm and 2 weeks before

the final exam was due). While the results from the current study yielded a small to moderately higher prediction accuracies compared to the existing literature, there are a few differences in our modeling approaches. Our results suggest that personalizing learning to the needs of the learners', their learning goals, and expectations should incorporate learner-centric measures such as self-efficacy, goals, expectancies and values, measures to understand students' task-completion rates, and academic diligence. More importantly, we were able to replicate the results from the other Study-1 and Study-2 to show that demographics boost early predictions, which have a lower predictive value later. Next, click behaviors were not a good indicator for students' performance predictions. Finally, we were able to surpass the prediction accuracies of our model by incorporating learner-centric measures on a weekly basis. Let us first revisit the importance given to click behaviors and how to navigate its advantages and disadvantages.

First, click and activity logs are readily available since most of the modern LMS platforms already have mechanisms in place to capture the clicks and activity logs. From a managerial perspective, it makes sense to want to explore the value of these readily available measures. The reasons for wanting to use these measures could vary based on literature review. However, intuitively, the simplest reasoning could be to make use of what is available that, in theory, could act as a proxy for students' learning behaviors without having to pour in additional resources. However, while gathering click behavior data from an established system is simply a matter of storage space, typically, making sense of such data involves a substantial amount of work, specifically, to recognize the patterns in users' behavior if LMS activity logs without any context. For

instance, the amount of resources consumed in terms of computation power, hours taken to clean, and process click data within our 5-week long course was a huge challenge, to derive metrics such as “click data – spacing” feature generated for the current study. The lack of clear guidelines in determining most useful click behaviors, the differences across LMS platforms used, and the limited involvement of students and in determining how and in what ways click behaviors are relevant to learning, the issues of overwhelming the students within open learner models, and the ethical concerns of tracking the students’ learning beyond LMS activity, all play a role in limiting the ability of using click behaviors to making accurate predictions. Thus, it is important to look beyond clicks to predict students’ learning.

Our results show that students’ expectancies and values, attainment and cost, task completion rates, and grade expectations consistently yielded the highest predictive value throughout the course. This resonates with the previous works where comprehensive educational data mining can learn from students’ motivational and affective processes which are correlated with classroom performances (McPartlan, 2020; Park et al., 2018). Specifically, this indicates that expectancies and values are not only positively associated with learning behaviors but can also be utilized for predicting learning behaviors in an online environment. While we did not specifically look to explore spacing behavior of the students in terms of procrastination, our spacing measures obtained from click behavior did not predict learning. Task completion rates of the students was a better predictor of learning. The fundamental difference between these two measures is at the level of the source. While spacing behavior was artificially derived from the clicks, number of completed tasks was derived from the self-reported

task-completion rates. This emphasizes the critical role of asking the learners about their behaviors rather than extrapolating their intentions from clicks. Furthermore, asking questions related to confidence of students in tackling an online course (via self-efficacy), and self-reported motivations were fruitful in our predictions. Therefore, the patterns of prediction accuracies of our stepwise modeling also underscore the importance of knowing demographics about the learners to get more accurate early predictions. Furthermore, using self-reported motivations on a weekly basis, specific to each weeks' tasks, importance non-academic activities and priorities, our models were able to learn students learning patterns better as was recently proposed by McPartlan (2020).

There are several limitations of the current work. First, our models were geared towards predicting the learning of students. This implies that we did not explore any associations between specific learner-centric traits to their actual performances. This implies that while we can see that self-efficacy was important to predicting the early learning, we do not know what level of self-efficacy is related to what level of student performance. Thus, it is difficult to determine the relationship of these learner-centric traits to the students' outcomes from the current work alone. Fortunately, recent work related to motivational traits and potential Utility Value Intervention (UVI) studies conducted speak for the mechanisms of these relationships. For instance, work done by Hulleman and colleagues (2017) investigating the value of UVI to increase student learning was possible due to the increased confidence of students in tackling learning which led to higher performance than the control group, and especially useful for lowest performing students since it improves the students' interest in the course. In our case, it

is reasonable to hypothesize that the students' performances were highly correlated with their motivational and perceptions, specifically their confidence coming into the course (self-efficacy), utility and the costs associated with taking the course, as well as the priority placed on course. In addition, the task completion rate (a measure of students' diligence and self-regulation) was highly correlated with GPAs and learning outcomes (Cochran, Campbell, Baker, & Leeds, 2014; Gore, 2006; Komarraju et al., 2013). Furthermore, learner-centric measures such as academic expectations, costs, and task completion rates were consistently important for predicting weekly performances. Therefore, our work reinforces the importance of measuring the dynamics of changes of these learner-centric traits to boost performances.

Another limitation of our work is in its lack of ability to determine the interplay between various features used. The approach we have used to demonstrate our classification results are black box models that look for hidden patterns and associations among variables. While there are human-interpretable models or ways in which we can turn the black box models into white box models, we did not choose to do so since the step-wise inclusion of features into our models were theoretically driven. Such an approach can only fix the lack of transparency to a certain degree. For instance, we know that including learner-centric features boost predictions sufficiently enough for us to care about them. However, we cannot elaborate on the ways in which these measures interplay with the students' learning directly. Another way to inspect these black box models would be to derive decision paths for each individual student (Guidotti et al., 2018). White box models would approach the predictive problems slightly differently. For instance, a decision path can be derived for each student to

inspect what features were considered relevant to enable the decision. While there are many other approaches similar to decision paths, most of these approaches sacrifice a certain level of accuracy in favor of interpretability of the model as detailed in the review by Guidotti and colleagues. There are two fundamental reasons for our choices to not proceed with a white box model. First, we did not seek to understand the decision made for each individual separately. The aim of our models was to establish the relative importance of our features, which was accomplished by the two steps we have taken in our modeling (i.e., theory driven stepwise feature inclusion and making use of robust random forest models that subset the data features for each iteration of predictions to generate overall features of importance). Next, we decided to forego individualized modeling in favor of individualized survey items by using our survey question responses to derive follow up questions for each student separately while withholding the overall survey items and structure. This goal was also driven by the fact that our models were conducted after the course was completed. Had we approached this prediction problem as a real-time solution to predicting each student's learning, we would have made use of a more transparent and interpretable modeling approach that would make it easier for the teachers and students to understand how to improve. Another potential issue of using decision paths approach would be in the varying amounts of data across models, and across the 5 weeks of data collection. Take the data from week 3 for instance: instead of data for the average learner-centric metrics per week, we would have six average learner-centric metrics. This would mean that the decision paths would have had a higher complexity and would have to be interpreted differently. However, these limits can be overcome in the future work by incorporating explanatory models such as

hierarchical linear models or structured equation modeling to determine the statistical significance of each predictor for each target metric in parallel to weekly predictions. This would help illustrate the significance of each weeks' predictions and the value of the significant (and non-significant) features towards predictions of students' learning.

Further limitations related to bias in data selection exist with the current work since measures of intentions and motivations are self-reports. Specifically, those students that responded to the surveys might be intrinsically different than those students who did not respond to the surveys. However, this issue might not be hindering of our results since the results presented here are on a held-out testing subset after learning from the training subset. Next, measures from self-reported data may also have validity issues (Chan, 2008). Schwarz (1999) also showed that self-reported measures are imperfect at measuring behavior and maybe affected by the wording and specific details within the questionnaires used. However, all of our survey metrics included more than one item to measure each learner-centric metric. The questionnaires were standardized, and the results of our survey responses demonstrated high reliability (Cronbach's  $\alpha > 0.85$  – see McPartlan, 2020). An attempt to overcome these limitations is made in the current work by incorporating multiple items of each dimension being measured and by including questions that require open-ended responses that may partially remove the bias induced by the researcher instrument.

Regardless of the limitations, the key to the success of our predictive models is to learn about user needs and motivations, rather than click behavioral patterns, to understand how and when an individual learning trajectory might show variations. Building predictive models that can capture and learn from the intraindividual variability



of these learner-centric models lead to successful predictions of learners' future behavior. This has several implications and for pedagogical practice. First, predictive modeling that seeks to understand students' learning should emphasize the value of learner-centric measures. While our work lays out the groundwork necessary to approach this problem, it is done after the course was completed (predictive analysis was conducted after the entire data was collected and not in real-time). This limits the potential of our work to making suggestions about interventions. For instance, UVI interventions that specifically target students' motivation and affective processes, might help improve positive learning behaviors. For instance, passive 'nudging' mechanisms built into LMS to promote task completion rates might help improve learning outcomes (Hulleman et al., 2017). In addition, it is reasonable to assume, given the results of our current work, that real-time predictive models will be able to perform better at understanding learners' needs if the models can account for students' motivations immediately prior to a task (task-specific student motivations), to improve student learning predictions. This might provide an opportunity for stakeholders such as administrators and teachers to allow for meaningful modifications of the learning material, speed, or instruction style, and the LMS features to identify those students that are struggling with setting timely goals and achieving them. In the absence of any learner related information, simply assuming that past repeats itself or somehow translates to future learning does a poor job of predicting a user's future learning. However, once the models understand the students' performance patterns on similar course activities (such as review quiz performances in week 3 predicted from week 2 and week 1 review quiz performances) at least after a sufficient amount of learning

trend of the user is available (and learned from). This, however, is not as robust or accurate as we have seen in the WM training task, and perhaps, any potential skill-acquisition task, where the dynamics of subject knowledge and level of understanding do not play the same role in learning. As a result, we were compelled to search for other sources of data that can complement a model based on the shallow history of a learners' performance in the early sessions of training. Our work demonstrated that a robust prediction model will be able to understand the erratic trends in the intraindividual variabilities using multidimensional information that are relevant to the learner, content, and the context. Overall, results suggested that predictions based solely on the learners' history of past actions may suffice to understand the general fate of the learner, as measured by whether a learner is in the upper half of his/her peers. These prediction trends are not limited to the level of binomial classifications.

Our regression models were able to replicate these results and show that we can come to a fairly robust estimate of the students' final scores on a scale of 0-100. Unfortunately, based on the current results, we are not confident in suggesting the use of our regression models without further considerations. For instance, a predicted score of 80 and an actual score of 87 would mean that the students would have a two-grade level difference in predictions (B+ vs B). As a result, the current model recommendations we can confidently make would be limited to understanding a coarse course performance difference (i.e., a median split or quartile splits). Any finer predictive modeling would require a more robust model that is resistant to noise, especially given the high dimensionality. This issue of higher noise can also be noticed in our own models consistently during week 3 performance predictions. As noted earlier, week 3

included survey response data on a daily level rather than, which resulted in poor predictions (often even compared to the predictions in week 2). Thus, a need for optimal predictive model search remains. While we first started our argument that we need a robust predictive model to promote equity by understanding learners' goals, needs, capabilities beyond averages, ironically, we can only recommend a model that can differentiate learners around the class performance centrality.

#### **4.7 Conclusion**

The methods in the current work demonstrate the validity of using learner-centric measures such as initial and evolving grade-expectations, cost value, self-efficacy, and task-completion rates. Our combined models that allowed for theoretical inclusion of features as well as stepwise predictions of performances week-by-week, allows for understanding the dynamics of learning trajectories over the short-burst time-span of interest. In general, our work validates the growing evidence for the lack of predictive value of unrefined and context-free click behaviors and the importance of individuals' explicit input in predictive modeling (Aldowah et al., 2019; Baradwaj & Pal, 2012; McCuaig & Baldwin, 2012; McPartlan, 2020; Salas et al., 2016). This further supports the need for open learner models where students inputs are utilized to improve predictive modeling within LMS (Baylor & Ritchie, 2002; Brusilovsky et al., 2014). When such theory driven predictive modeling is undertaken and thoughtfully implemented within LMS, we may provide an opportunity to identify individuals who are at risk for showing poor learning. Predictive models have very robust practical applications that can be used to make real-time alterations to many dimensions that help improve learners' trajectories. Thus, the hopeful message is that, we can understand learners'

needs and improve the overall learning process, by asking for learners' inputs. Perhaps a weekly check-in with students to understand their responsibility, priorities, and how well they are able to achieve their little goals, are enough to make a difference. This may also reinforce the study habits of each individual that may translate to optimal learning over time beyond the context of a single classroom.

## CHAPTER 5: GENERAL DISCUSSION

In the current state of the world, there is a sudden peak in the interest in online learning and virtual classrooms. The rise in growth of online learning has been noticed over the past decade despite the concerns over the validity, effectiveness, and quality of online learning. Although disparities in learning has been a major concern across learning settings, approaches such as personalization are hopeful to promote access and quality of learning. Predicting learning quality and quantity at the earliest possible time has been a central and recurring theme of these personalization approaches. However, there have been differences in the ways predictive modeling has been applied in predicting learning behavior. While some of these differences are attributed to the exploratory nature of the data mining approaches, a majority of these differences are difficult to understand, largely due to the lack of single framework to test varying features and models. The three fundamental issues that I have discussed continue to prevail in the applications of predictive modeling to online learning settings. These issues arise from the lack of standardizations of predictive modeling, the fundamental limits of predictability of human (learning) behavior, and the lack of consensus amongst researchers regarding the value of predictions. Combine these three issues with the extensive use of context-agnostic click behavioral data in determining policies and practices that determine learning, and the inability of predictive models to provide actionable insights, the fields of EDM and LA have a non-trivial task at hand.

In the current work, I have proposed to use a single framework that may overcome some of these issues. The framework has four critical elements – a) identification of the prediction task and reporting all metrics for the best model

determined by researchers, b) including a baseline model for comparing the performance of the models of interest, c) using data of the learners early learning alone in determining later learning, and d) stepwise inclusion of features of interest, starting with least malleable to most malleable features that are appropriate for the context.

Here are some of the key contributions and a summary of the results of my work.

In the introduction, I discuss the different approaches taken, the complexity of the approaches and the features that are used to understand learning. The existing literature emphasizes using features relevant to learners, the content being taught, and the context in which learning occurs to understand learning. While there are many different approaches to making predictions of learning specific to any given context as we have seen from the extensive literature, these results lack comparability. The current literature has a diversity of features, models, and results. The current work seeks to use this single framework to better ascertain the fundamental value of predictions and to promote ability to compare results. This framework is novel and meaningful for several important reasons.

- First, focusing on a single or a best model and reporting the results from such a model does not provide a meaningful comparability of results across models and settings. The current framework encourages a step-wise utilization of modeling to report results from varying degrees of information provided to the models to compare the value of features.
- Second, the fundamental nature of using accuracies alone to determine value of a classification model is known to lead to many issues such as lack of understanding of specificity and sensitivity of models, overfitting to drive

- higher accuracies from models, and lack of understanding of accuracy-complexity tradeoffs. The current work has shown that accuracies of models are directly related to the two other questions – how soon and how much information which are overlooked often.
- Third, relatedly, the current work hopes to answer the question – how soon can we make good predictions? This question is often overlooked in the context of making predictions. However, as I have demonstrated and discussed across all three studies, the ability to determine the utility of a predictive model should not be restricted to how well. In fact, more often than not, in a learning setting, teachers would want to get an idea of who is likely to fail at the earliest possible time, at the cost of some accuracy. The current framework provides researchers and teachers with an opportunity to understand the accuracy – recency tradeoffs for their own learning setting.
  - Fourth, the current work showed the varying degrees of utility of features in making predictions of learning. For instance, the results have shown that across all three settings, predictive models that hope to learn from students' behavior, without knowing much about their demographics or motivations seem to underperform early. However, we have seen that it becomes relatively easier to make predictions of students' learning behavior if the models are provided with information about the students' behavior. This approach was useful in determining prediction values of features such as click behaviors and learners' dynamic motivations. Specifically, results indicated that click behaviors do not offer good predictive value, whereas, asking the

learners about their goals, costs, and other motivational and affective traits over the learning period is more valuable to making predictions.

These four key contributions are accompanied by a few separate key points that I highlighted across the thesis. First, using a single metric variant (i.e., Accuracy or RMSE) might not suffice given that model performances assessed by a single metric are rarely meaningful. As such, optimally solving a prediction task requires that the researchers make careful and valid choices for data sources and features measured ex ante, data preprocessing and feature selection ex post, and evaluation metrics that are used to communicate their findings. This implies that subjective choices made by researchers, often influence the results. Furthermore, results across predictive models are difficult to replicate without sandbox environments due to constantly changing software and core implementation of off the shelf algorithms. Even without using novel approaches to optimize models (which might arguably be beyond the ability of social scientists that are drawn to the novelty and applicability of machine learning approaches), one way to ensure that results are better understood is by reporting all the decisions that were taken in determining the model choices. Furthermore, providing full sets of metrics (accuracy, precision, recall, F1-score, and AUC for classification tasks and adjusted  $R^2$  for regression tasks) will help understand the varying results across settings. For instance, in Study-1, we have seen that the adjusted  $R^2$  obtained on the best model by Session 10 of the WM data was  $\sim 0.60$ , indicating that the regression model was able to explain 60% of the variance. However, the classification model on the same dataset when inspecting using accuracies as a metric yielded an 83% accuracy concluding that these predictions are “very accurate”. In theory, even if the



data sources and models are held constant, every decision a researcher makes could lead to fundamentally differing results and conclusions. Each branch that arises from these decisions can be reported as an independent study by itself. For instance, one study could be focused on optimal depth and tree size selections (pruning) in a random forest model to maximize prediction accuracies and minimizing model complexity. Another study could be conducted on the same data to compare different classification models. Each of these studies could be rationalized based on what the researchers are hoping to accomplish. Thus, it is important to account for all choices being made and to report all the results.

Thus, across all three studies in the current thesis, an emphasis on reporting all standard metrics was placed, such that, researchers can see how simple questions such as “how well”, “how soon”, and “how much information” can yield varying answers. Furthermore, as evidence from this work suggests, answers to these questions are often interlinked following a common theme. If we are interested in making robust predictions at the earliest possible time, we need to include features that capture the context of learning and any known extrinsic features that constrain the learning of students to be able to make better predictions within the given context. However, later learning can be predicted fairly well, roughly midway through the learning period with prediction accuracies hovering around 80% without including any extrinsic features. However, the questions that remain unanswered and are likely to be left unanswered for the foreseeable future is to what extent prediction accuracies are meaningful and why one needs to care about predictions in the first place. The answer to this question is more nuanced than any single metric (unless our predictions are near 100% accurate all

the time). Thus, inclusion of metrics beyond accuracies are important. Specifically, if the goal is to determine which learners need most support to succeed, then we need to focus on choices that lead to better understand this group of learners. There are several ways to accomplish this. For instance, classification models for the top 25% students and bottom 25% students yield better predictions than a median split. Subsequently, a focus on subclass accuracies, precisions, and recalls for the bottom 25% students will yield a better indicator for predicting who needs most help. Furthermore, manipulating the categorical coding schema where 'positive' corresponds to 'below median' or 'students that need most help' will improve the value of precisions and recalls since these two metrics evaluate the successful classifications of the positive class as selected by the researchers.

Next, it is necessary to compare the value of all metrics (and all results for every model) against a baseline model using a common task framework. Ideally, all competing best algorithms derived by multiple researchers on a single dataset (or relevant standardized publicly available dataset) or task are independently evaluated by third parties against very high-quality baseline models using all performance metrics. However, in reality, this has been very difficult to accomplish within the context of online learning contexts. Most datasets are rarely ever made publicly available or standardized for replicability. Thus, the best possible approach to understand the value of models is to compare the results with itself, bar real data. This provides researchers a way to understand how much better results of a model are (or the lift ratios are) in relation to the baseline. This is ever more valuable in exploratory predictive modeling when researchers have to choose between many different competing models to solve a

prediction task. When researchers are reusing the same data set (or subset) for assessing different competing models to determine which one model to choose, the results lead to overestimates of true predictions. This error, referred to as “human-in-the-loop overfitting” where researchers intentionally report models that have higher prediction accuracies relative to each other rather than relative to the baseline model using the same algorithm. For instance, if two competing models have an adjusted- $R^2$  of 0.91 and 0.89, researchers might choose the former due to the better performance. However, if the baseline performance of the models has an adjusted- $R^2$  of 0.69 and 0.54 respectively, the gain in adjusted- $R^2$  of the second model over its corresponding baseline model is higher ( $0.22 < 0.35$ ). Yet again, comparison of performances of models against baselines act a better indicator of how valuable predictions of different models are as well as how valuable features of interest are in improving predictions of the model of choice. Thus, the current work included baseline models for selecting the models as well as model variant comparisons across all three studies. Of course, there are many approaches to baseline modeling. However, as long as the baseline models are able to demonstrate the best predictions possible by simple heuristics or guess rates, they might act as decent first baseline models.

Next, given the significant differences in features used for making predictions of learning in the literature, the proposed framework used a stepwise inclusion of features that might predict learning in a given context. Given the importance placed on historic performances on the future learning within both explanatory as well as predictive models, the current framework included a step that seeks to exploit the relationship of past performances to future performances. This served two purposes within this

proposed framework. First, this step was used to evaluate if learning performances in a given context alone are sufficient to make predictions of later learning. We have seen that while later learning can be predicted from learning or performances alone, they are only meaningfully robust halfway through the learning phase. Next, this step was used to evaluate if using learner-agnostic and context-agnostic models are useful for making predictions of future learning. For instance, in an offline classroom, instructors can rely on information known about the context, setting, and learner to make pedagogical choices and assess learning. However, within fully online learning contexts, teachers do not have the same opportunity to understand these extrinsic features. Furthermore, predictive models that outright rely on demographic details are known to be prone to biases (Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2018). Furthermore, unlike in explanatory models, predictive modeling approaches cannot “account for” or “control” for baseline performances. Thus, the current framework made use of a model that only relies on learners’ performances/quiz scores alone. This model has shown across all studies that early predictions are poor since the models do not have sufficient understanding of each learner’s performance. Thus, it is important to use and incorporate extrinsic features into the model to improve predictions during the early learning phase.

Finally, the proposed framework utilizes sequential inclusion of two different sets of variables. First, we suggest inclusion of non-malleable features followed by inclusion of malleable features. One of the biggest issues of predictive modeling in online learning contexts (and social sciences in general) is the prevalence of studies that make use of features to make robust predictions over 90% but lacking in guiding the policy-makers,

teachers, and students as to what makes those predictions meaningful. The problem is often attributed to complex modeling approaches that are difficult to interpret. However, even those studies that used simple regression-based prediction models do not offer valuable prescriptions for future actions. Thus, our framework suggests using features that are agreed upon as non-malleable within a given context. For instance, features such as age and gender cannot be manipulated and lead to prescriptions that are restrictive (task X is beneficial for younger adults, task Y is easier for older adults). These restrictive prescriptions are known to be debilitating (negative motivation) to those individuals that are drawn to tasks that they are interested in. Thus, to evaluate the features that are malleable, they are added to the predictive models at the very end to determine if including these features increase predictions. In the current work, we have seen that in Study-1, including WM stimulus type, compensation and supervision in our final model iteration led to significantly high predictive performances. Including these features (alongside the existing features) were able to predict the later performance of learners at Session 1 as much as performances and age were able to predict later learning at Session 8. Thus, including these features might offer a significant advantage while making decisions on how to intervene for those learners that are predicted to be below a predetermined threshold. In Study-2, including demographics and students' study spacing intentions were significant predictors of later learning, whereas including click behaviors hurt the models' performances. This indicated that not all malleable features improve prediction accuracies since the noise associated with certain features lead to overall poor predictions. However, it does not necessarily mean that click-behaviors are not a good indicator for learning. Such a

conclusion would require evaluation of black-swan cases where, click behaviors indeed capture human behavioral traits such as procrastination, study spacing, and task diligence of learners. In Study-3, we have seen results similar to Study-2, where, inclusion of non-malleable features improved predictions and inclusion of some malleable features led to poor or no further improvement in predictions. Furthermore, malleable learner-centric features were able to improve predictions of our models, indicating that click-behaviors that are context-agnostic, are indeed prone to high noise and need very careful consideration when suggesting actionable prescriptions. As a

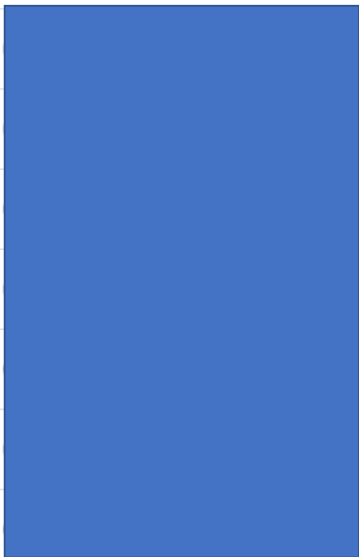
Students (38) ▲	Grade	% On Time	Last Participation	Last Page View	Page Views	Participation
	89%	41%	Mar 13, 2020	Mar 20, 2020	475	21
	83%	47%	Mar 12, 2020	Mar 20, 2020	377	40
	81%	71%	Mar 15, 2020	Mar 22, 2020	312	24
	85%	59%	Mar 13, 2020	Mar 25, 2020	865	37
	93%	82%	Mar 13, 2020	Mar 27, 2020	479	20
	94%	35%	Mar 13, 2020	Mar 20, 2020	525	19
	82%	82%	Mar 12, 2020	Mar 27, 2020	799	23

Figure 5.1 A screenshot of "New Analytics" -- provided on Canvas LMS at UCI. Each row contains information for a single student.

side note, we are actively investigating means to understand click behaviors and their utility in determining learning. For instance, some of our predictive models utilized click behaviors alone to make predictions of learning behaviors. However, due to the highly variable quantities of individuals' clicks, the predictions were below 65% across time points for studies 2 and 3.

Specifically, given that click-behaviors and activity-logs of students on LMS are used for predictive modeling as well as for policymaking by administrators, it is important to discuss the value of click-behaviors derived from LMS. Let us inspect one example of using click-behaviors within LMS at UCI. Figure 5.1 shows a screenshot taken from the Canvas, the LMS used at UC Irvine. LMS provides learner analytics referred to as “New Analytics” to instructors and teaching assistants (TAs) based on students’ participation, performances, and click behaviors. Each row in the figure corresponds to a single student. There are 6 columns each corresponding to a different measure provided for the convenience of the instructors and TAs. These include Grade, % of tasks completed On Time, Last Participation, Last Page View date, Total number of page views, and Participation count (on discussion or forums). Upon closer inspection, Grade of student 2 and student 4 (see corresponding columns) indicated that these two students have similar grade (83% and 85% respectively). However, the total amount of clicks for these students differ by 488 clicks (student 4 has more than twice as many clicks as student 2.) Similarly, students Grade of student 5 and student 6 (see corresponding columns) are very similar (93% and 94% respectively). While, there were no substantial differences between these two students’ clicks, student 5 completed 82% of the tasks on time, whereas student 6 completed a mere 35% of tasks on time. This is one selective example of the differences that students display within online learning contexts. There is no strict one-to-one correspondence of artificially generated metrics such as % completed on time or clicks since they are not directly considered during grading. However, it is difficult to understand the rationale for including those features within LMS for the instructors without knowing the intentions of the

administrators. It is possible that these were included within the LMS to provide insights into the ways the learners interact with the course deployed within the LMS. Furthermore, perhaps these metrics are provided to promote open learning models that were discussed in the previous chapters. However, there are several drawbacks of providing this data to instructors in light of the results from the current study.

Specifically, since click-behaviors, in their current state do not act a good metric for predicting students' grades, the decision to promote "New Analytics" implies that insufficient evidence was used to push this policy. This is critical since, without proper guidance or reasoning for including these measures, instructors might be misled to consider artificially generated measures to be signals of good grades. The other pertinent issues with this information is the lack of control given to teachers in determining what information they can see and cannot see. For instance, if say, a teacher places a special emphasis on timely completion of all assigned tasks and maybe promotes this behavior by assigning some course credit for timely completion, then perhaps it could act as a partial indicator of success in that course. There is no means for the teachers to isolate this information from "New Analytics" dashboard. Furthermore, in a large classroom setting with hundreds of students, it becomes difficult to manually check the differences in patterns of click behaviors. Furthermore, teachers are not provided sufficient training on how to look for relevant signals from these data.

Results from Study-3 showed the importance of measuring students' motivational, affective, and emotional traits that are context-relevant and are continually measured over the learning period are more beneficial measures of students' performances than students' click behaviors. Students' self-efficacy that was measured



at presurvey was a better predictor of students' quiz 1 performance. Furthermore, students goal attainment, costs, task-completion rates were more predictive of weekly quiz performances consistently. Although these measures used items that are specific to the context of the course, it is reasonable to argue that context-specific motivational and affective measures might be more relevant to learning in any learning setting.

Overall, the results of the current work illustrate the benefits of a unified framework to predict learning across many contexts and settings while deriving metrics that are comparable. I believe that the approach I have taken highlight the importance of understanding that the answer to the three fundamental questions related to prediction tasks are related to each other:

*How well can we predict?*

*How soon can we predict?*

*How much information do we need?*

The answers to these questions vary, not just from context to context, but also from model to model. In the current work, I highlighted the shortcomings of the existing literature of predictive modeling within online learning contexts. Furthermore, I have also proposed a solution that demonstrates that by continuing to combine theoretically driven features with machine learning based predictive modeling, using a unified framework we may be able to take steps to overcome these shortcomings. While the endeavor of true personalization is a long way from being realized, I believe that every step taken to promote combining human and machine intelligence as we move forward together will add to the discussions in hopes of achieving the greater goals of *inclusion* and *equity* in learning.

## BIBLIOGRAPHY

- Abrami, P. C., Bernard, R. M., Bures, E. M., Borokhovski, E., & Tamim, R. M. (2011). Interaction in distance education and online learning: Using evidence and theory to improve practice. *Journal of Computing in Higher Education*.  
<https://doi.org/10.1007/s12528-011-9043-x>
- Agudo-Peregrina, Á. F., Iglesias-Pradas, S., Conde-González, M. Á., & Hernández-García, Á. (2014). Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in Human Behavior*.  
<https://doi.org/10.1016/j.chb.2013.05.031>
- Ahn, J., Pellicone, A., & Butler, B. S. (2014). Open badges for education: What are the implications at the intersection of open systems and badging? *Research in Learning Technology*. <https://doi.org/10.3402/rlt.v22.23563>
- Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*. <https://doi.org/10.1016/j.tele.2019.01.007>
- Aletras, N., & Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics, IWCS 2013 - Long Papers*.
- Alpert, W. T., Couch, K. A., & Harmon, O. R. (2016). A randomized assessment of online learning. In *American Economic Review*.  
<https://doi.org/10.1257/aer.p20161057>
- Alqurashi, E. (2016). Self-Efficacy In Online Learning Environments: A Literature Review. *Contemporary Issues in Education Research (CIER)*.  
<https://doi.org/10.19030/cier.v9i1.9549>
- Amanda Stedke. (2017). Differentiation, individualization and personalization: What they mean, and where they're headed | eSchool News.
- Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2014). Engaging with massive online courses. In *Proceedings of the 23rd international conference on World wide web - WWW '14*. <https://doi.org/10.1145/2566486.2568042>
- Anderson, T. (2004). Toward a theory of online learning. In *Theory and Practice of Online Learning*. [https://doi.org/10.1111/j.1467-8535.2005.00445\\_1.x](https://doi.org/10.1111/j.1467-8535.2005.00445_1.x)
- Au, J., Buschkuehl, M., Duncan, G. J., & Jaeggi, S. M. (2016). There is no convincing evidence that working memory training is NOT effective: A reply to Melby-Lervåg and Hulme (2015). *Psychonomic Bulletin and Review*.  
<https://doi.org/10.3758/s13423-015-0967-4>
- Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015a). Improving fluid intelligence with training on working memory: a meta-analysis. *Psychonomic Bulletin and Review*. <https://doi.org/10.3758/s13423-014-0699-x>
- Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015b). Improving fluid intelligence with training on working memory: a meta-analysis. *Psychonomic Bulletin and Review*. <https://doi.org/10.3758/s13423-014-0699-x>
- Baker, R. S. (2016). Stupid Tutoring Systems, Intelligent Humans. *International Journal*

- of *Artificial Intelligence in Education*. <https://doi.org/10.1007/s40593-016-0105-0>
- Bandura, A. (2010). Self-efficacy -Bandura. *The Corsini Encyclopedia of Psychology*. <https://doi.org/10.9780470479216>.
- Bandura, A. (2012). Social cognitive theory. In *Handbook of Theories of Social Psychology: Volume 1*. <https://doi.org/10.4135/9781446249215.n18>
- Bandura, A., & Jones, M. R. (1962). Social learning through imitation. In *Nebraska Symposium on Motivation, 1962*. <https://doi.org/10.1016/j.bpj.2009.11.008>
- Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. *ArXiv Preprint ArXiv:1201.3417*.
- Barbé, J., Mollard, R., & Wolff, M. (2014). Ergonomic approaches to integrate touch screen in future aircraft cockpits. *Journal Europeen Des Systemes Automatises*. <https://doi.org/10.3166/JESA.48.303-318>
- Bates, R., & Khasawneh, S. (2007). Self-efficacy and college students' perceptions and use of online learning systems. *Computers in Human Behavior*. <https://doi.org/10.1016/j.chb.2004.04.004>
- Battle, A., & Wigfield, A. (2003). College women's value orientations toward family, career, and graduate school. *Journal of Vocational Behavior*. [https://doi.org/10.1016/S0001-8791\(02\)00037-4](https://doi.org/10.1016/S0001-8791(02)00037-4)
- Bawa, P. (2016). Retention in Online Courses: Exploring Issues and Solutions—A Literature Review. *SAGE Open*. <https://doi.org/10.1177/2158244015621777>
- Bayer, J., Bydzovská, H., & Géryk, J. (2012). Predicting Drop-Out from Social Behaviour of Students. *Proceedings of the 5th International Conference on Educational Data Mining*.
- Baylor, A. L., & Ritchie, D. (2002). What factors facilitate teacher skill, teacher morale, and perceived student learning in technology-using classrooms? *Computers and Education*. [https://doi.org/10.1016/S0360-1315\(02\)00075-1](https://doi.org/10.1016/S0360-1315(02)00075-1)
- Beaudry, P., Green, D. A., & Sand, B. M. (2016). The Great Reversal in the Demand for Skill and Cognitive Tasks. *Journal of Labor Economics*. <https://doi.org/10.1086/682347>
- Beckmann, J. F., & Birney, D. P. (2012). What happens before and after it happens: Insights regarding antecedences and consequences of adult learning. *Learning and Individual Differences*. <https://doi.org/10.1016/j.lindif.2012.06.005>
- Bennett, S. N. (1978). Recent research on teaching: A dream, a belief, and a model. *British Journal of Educational Psychology*, 48(2), 127–147.
- Bettinger, E., & Loeb, S. (2017). Promises and pitfalls of online education. *Education Next*.
- Bienkowski, M., Feng, M., & Means, B. (2014). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. In *Educational Improvement Through Data Mining and Analytics*.
- Bjork, E. L., & Bjork, R. (2009). Making Things Hard on Yourself, But in a Good Way: Creating Desirable Difficulties to Enhance Learning. In *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*. <https://doi.org/10.1017/CBO9781107415324.004>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*. <https://doi.org/10.1016/b978-0-12-411519-4.00006-9>
- Bloom, B. S. (1976). *Human characteristics and school learning*. McGraw-Hill.

- Bo, J., Borza, V., & Seidler, R. D. (2009). Age-Related Declines in Visuospatial Working Memory Correlate With Deficits in Explicit Motor Sequence Learning. *Journal of Neurophysiology*. <https://doi.org/10.1152/jn.00393.2009>
- Bodily, R., Kay, J., Alevan, V., Jivet, I., Davis, D., Xhakaj, F., & Verbert, K. (2018). Open learner models and learning analytics dashboards: A systematic review. In *ACM International Conference Proceeding Series*. <https://doi.org/10.1145/3170358.3170409>
- Boelens, R., Van Laer, S., De Wever, B., & Elen, J. (2015). *Blended learning in adult education: towards a definition of blended learning. Adult Learners Online! Blended and Online Learning in Adult Education and Training*.
- Bogg, T., & Lasecki, L. (2015). Reliable gains? Evidence for substantially underpowered designs in studies of working memory training transfer to fluid intelligence. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2015.00589>
- Bong, M. (2001). Role of self-efficacy and task-value in predicting college students' course performance and future enrollment intentions. *Contemporary Educational Psychology*. <https://doi.org/10.1006/ceps.2000.1048>
- Bong, M. (2004). Academic Motivation in Self-Efficacy, Task Value, Achievement Goal Orientations, and Attributional Beliefs. *Journal of Educational Research*. <https://doi.org/10.3200/JOER.97.6.287-298>
- Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Interdisciplinary Journal of Information, Knowledge, and Management*. <https://doi.org/10.28945/4184>
- Bowen, H. R., Fincher, C., Bowen, H. R., & Fincher, C. (2019). The Future of American Higher Education. In *Investment in Learning*. <https://doi.org/10.4324/9781351309929-19>
- Bowen, W. (2013). The Potential for Online Learning: Promises and Pitfalls. *Educase Review*.
- Bowen, W. G., Chingos, M. M., Lack, K. A., & Nygren, T. I. (2013). Online learning higher education. *Education Next*.
- Bowen, W. G., Delbanco, A., Gardner, H., Hennessy, J. L., & Koller, D. (2013). *Higher education in the digital age. Higher Education in the Digital Age*. <https://doi.org/10.1515/9781400866137>
- Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*. <https://doi.org/10.1214/ss/1009213726>
- Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *Internet and Higher Education*. <https://doi.org/10.1016/j.iheduc.2015.04.007>
- Brownlee, J. (2016). What is a Confusion Matrix in Machine Learning. *Machinelearningmastery.Com*. <https://doi.org/10.1145/2000824.2000827>
- Bruner, J. S. (1966). *Toward a theory of instruction* (Vol. 59). Harvard University Press.
- Brusilovsky, P., Edwards, S., Kumar, A., Malmi, L., Benotti, L., Buck, D., ... Wollowski, M. (2014). Increasing adoption of smart learning content for computer science education. In *ITiCSE-WGR 2014 - Working Group Reports of the 2014 Innovation and Technology in Computer Science Education Conference*. <https://doi.org/10.1145/2713609.2713611>
- Bull, S., & Kay, J. (2013). Open Learner Models as Drivers for Metacognitive

- Processes. [https://doi.org/10.1007/978-1-4419-5546-3\\_23](https://doi.org/10.1007/978-1-4419-5546-3_23)
- Bull, S., & Kay, J. (2016). SMILI: A Framework for Interfaces to Learning Data in Open Learner Models, Learning Analytics and Related Fields. *International Journal of Artificial Intelligence in Education*. <https://doi.org/10.1007/s40593-015-0090-8>
- Carroll, J. (1963). A model for school learning. *The Teachers College Record*. <https://doi.org/10.1016/j.actatropica.2014.07.009>
- Carroll, J. B. (1989). The Carroll Model: A 25-Year Retrospective and Prospective View. *Educational Researcher*. <https://doi.org/10.3102/0013189X018001026>
- Cattell, R. B., & Child, D. (1975). *Motivation and dynamic structure*. Halsted Press.
- Chan, D. (2008). So why ask me? Are self-report data really that bad? In *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences*. <https://doi.org/10.4324/9780203867266>
- Chang, C. C., Tseng, K. H., Liang, C., & Liao, Y. M. (2013). Constructing and evaluating online goal-setting mechanisms in web-based portfolio assessment system for facilitating self-regulated learning. *Computers and Education*. <https://doi.org/10.1016/j.compedu.2013.07.016>
- Chatti, A. M., Lukarov, V., Thüs, H., Muslim, A., Yousef, F. A. M., Wahid, U., ... Schroeder, U. (2014). Learning Analytics: Challenges and Future Research Directions. *Eleed*.
- Chaturvedi, S., Goldwasser, D., & Daumé III, H. (2014). Predicting Instructor's Intervention in MOOC forums. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*. <https://doi.org/10.3115/v1/P14-1141>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. <https://doi.org/10.1145/2939672.2939785>
- Chiu, C. M., & Wang, E. T. G. (2008). Understanding Web-based learning continuance intention: The role of subjective task value. *Information and Management*. <https://doi.org/10.1016/j.im.2008.02.003>
- Cho, M. H., & Heron, M. L. (2015). Self-regulated learning: the role of motivation, emotion, and use of learning strategies in students' learning experiences in a self-paced online mathematics course. *Distance Education*. <https://doi.org/10.1080/01587919.2015.1019963>
- Cochran, J. D., Campbell, S. M., Baker, H. M., & Leeds, E. M. (2014). The Role of Student Characteristics in Predicting Retention in Online Courses. *Research in Higher Education*. <https://doi.org/10.1007/s11162-013-9305-8>
- Conley, D. T. (2008). Rethinking college readiness. *New Directions for Higher Education*. <https://doi.org/10.1002/he.321>
- Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*. <https://doi.org/10.1016/j.tics.2003.10.005>
- Cooley, W. W., & Leinhardt, G. (1975). The application of a model for investigating classroom processes.
- Corno, L., Cronbach, L. J., Kupermintz, H., Lohman, D. F., Mandinach, E. B., Porteus, A. W., & Talbert, J. E. (2001). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. Routledge.

- Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin and Review*. <https://doi.org/10.3758/s13423-016-1191-6>
- Cronbach, L. J. (2003). Review of Remaking the Concept of Aptitude: Extending the Legacy of Richard E. Snow. *Personnel Psychology*. <https://doi.org/10.1111/j.2042-3306.2010.00181.x>
- Dahlstrom, E., Brooks, D. C., & Bichsel, J. (2014). The Current Ecosystem of Learning Management Systems in Higher Education: Student, Faculty, and IT Perspectives. *EDUCAUSE Center for Analysis and Research*. <https://doi.org/10.13140/RG.2.1.3751.6005>
- Daud, A., Lytras, M. D., Aljohani, N. R., Abbas, F., Abbasi, R. A., & Alowibdi, J. S. (2019). Predicting student performance using advanced learning analytics. In *26th International World Wide Web Conference 2017, WWW 2017 Companion*. <https://doi.org/10.1145/3041021.3054164>
- Day, E. A., Arthur, W., & Gettman, D. (2001). Knowledge structures and the acquisition of a complex skill. *Journal of Applied Psychology*. <https://doi.org/10.1037/0021-9010.86.5.1022>
- Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4), 429–453.
- Deci, E. L., Ryan, R. M., Vallerand, R. J., & Pelletier, L. G. (1991). Motivation and Education: The Self-Determination Perspective. *Educational Psychologist*. <https://doi.org/10.1080/00461520.1991.9653137>
- Deming, D. J., Goldin, C., Katz, L. F., & Yuchtman, N. (2015). Can online learning bend the higher education cost curve? In *American Economic Review*. <https://doi.org/10.1257/aer.p20151024>
- Dennen, V. P., Darabi, A. A., & Smith, L. J. (2007). Instructor-learner interaction in online courses: The relative perceived importance of particular instructor actions on performance and satisfaction. *Distance Education*. <https://doi.org/10.1080/01587910701305319>
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [https://doi.org/10.1007/3-540-45014-9\\_1](https://doi.org/10.1007/3-540-45014-9_1)
- Dillenbourg, P., Schneider, D., & Synteta, P. (2001). Virtual Learning Environments, 3–18. Retrieved from <https://telearn.archives-ouvertes.fr/hal-00190701/>
- Doménech-Betoret, F., Abellán-Roselló, L., & Gómez-Artiga, A. (2017). Self-efficacy, satisfaction, and academic achievement: The mediator role of students' expectancy-value beliefs. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2017.01193>
- Domingos, P. (1999). The role of Occam's Razor in knowledge discovery. *Data Mining and Knowledge Discovery*. <https://doi.org/10.1023/A:1009868929893>
- Dreyfus, S. E., & Dreyfus, H. L. (1980). A five stage model of the mental activities involved in direct skill acquisition. *Research Paper*. <https://doi.org/ADA084551>
- Dutt, A., Ismail, M. A., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2017.2654247>
- Eccles, J. S. (2013). Expectancy Value Motivational Theory. *Education.Com*.
- Eccles, J. S., Adler, T., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In *Achievement*

- and achievement motivation: *Psychological and sociological approaches*.  
<https://doi.org/10.1207/s15327752jpa8502>
- Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as representation of momentum. *Trends in Cognitive Sciences*, 20(1), 15–24.
- Engel, Y., Mannor, S., & Meir, R. (2004). The kernel recursive least-squares algorithm. *IEEE Transactions on Signal Processing*. <https://doi.org/10.1109/TSP.2004.830985>
- Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*. <https://doi.org/10.1111/1467-8721.00160>
- Engle, R. W. (2018). Working Memory and Executive Attention: A Revisit. *Perspectives on Psychological Science*, 13(2), 190–193.  
<https://doi.org/10.1177/1745691617720478>
- Farina, M. (2013). The evolved apprentice. How evolution made humans unique. *Phenomenology and the Cognitive Sciences*. <https://doi.org/10.1007/s11097-012-9276-9>
- Fernandez, A. A., & Shaw, G. P. (2020). Academic Leadership in a Time of Crisis: The Coronavirus and COVID-19. *Journal of Leadership Studies*.  
<https://doi.org/10.1002/jls.21684>
- Figlio, D., Rush, M., & Yin, L. (2013). Is it live or is it internet? experimental estimates of the effects of online instruction on student learning. *Journal of Labor Economics*.  
<https://doi.org/10.1086/669930>
- Flach, P. A. (2003). The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics. In *Proceedings, Twentieth International Conference on Machine Learning*.
- Gabrielle, D. M. (2003). *The effects of technology -mediated instructional strategies on motivation, performance, and self -directed learning*. ProQuest Dissertations and Theses.
- Gagne, R. M. (1977). Analysis of objectives. *LJ Briggs, Instructional Design*, 115–145.
- Gallego, G., & Topaloglu, H. (2019). Online Learning. In *International Series in Operations Research and Management Science*. [https://doi.org/10.1007/978-1-4939-9606-3\\_10](https://doi.org/10.1007/978-1-4939-9606-3_10)
- Gaspard, H., Dicke, A. L., Flunger, B., Schreier, B., Häfner, I., Trautwein, U., & Nagengast, B. (2015). More value through greater differentiation: Gender differences in value beliefs about math. *Journal of Educational Psychology*.  
<https://doi.org/10.1037/edu0000003>
- Gibson, J. J., & Gibson, E. J. (1955). PERCEPTUAL LEARNING: DIFFERENTIATION OR ENRICHMENT? *Psychological Review*. <https://doi.org/10.1037/h0048826>
- Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on “Estimating the reproducibility of psychological science.” *Science*.  
<https://doi.org/10.1126/science.aad7243>
- Glance, D. G., & Barrett, P. H. R. (2014). Attrition patterns amongst participant groups in Massive Open Online Courses. In *Proceedings of ASCILITE 2014 - Annual Conference of the Australian Society for Computers in Tertiary Education*.
- Glaser, R. (1976). Components of a Psychology of Instruction: Toward a Science of Design. *Review of Educational Research*.  
<https://doi.org/10.3102/00346543046001001>
- Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans.

- American Psychologist*. <https://doi.org/10.1037/0003-066X.54.7.493>
- Gore, P. A. (2006). Academic self-efficacy as a predictor of college outcomes: Two incremental validity studies. *Journal of Career Assessment*. <https://doi.org/10.1177/1069072705281367>
- Graesser, A. C., Hu, X., & Sottolare, R. (2018). Intelligent tutoring systems. In *International Handbook of the Learning Sciences*. <https://doi.org/10.4324/9781315617572>
- Graham, C. R. (2006). Blended learning systems: Definition, current trends, and future directions. In *Handbook of blended learning: Global perspectives, local designs*.
- Gray, P. (2009). Play as a foundation for hunter-gatherer social existence. *American Journal of Play*. [https://doi.org/10.1300/J082v41n02\\_07](https://doi.org/10.1300/J082v41n02_07)
- Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. Handbook of educational psychology. In *Handbook of educational psychology*. <https://doi.org/10.1007/s13398-014-0173-7.2>
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*. <https://doi.org/10.1145/3236009>
- Guo, P. J., & Reinecke, K. (2014). Demographic differences in how students navigate through MOOCs. In *Proceedings of the first ACM conference on Learning @ scale conference - L@S '14*. <https://doi.org/10.1145/2556325.2566247>
- Guyon, I., & Elisseeff, A. (2011). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*. <https://doi.org/10.1016/j.aca.2011.07.027>
- Haase, C. M., Heckhausen, J., & Wrosch, C. (2013). Developmental regulation across the life span: Toward a new synthesis. *Developmental Psychology*. <https://doi.org/10.1037/a0029231>
- Haertel, G. D., Walberg, H. J., & Weinstein, T. (1983). Psychological Models of Educational Performance: A Theoretical Synthesis of Constructs. *Review of Educational Research*. <https://doi.org/10.3102/00346543053001075>
- Hamm, J. M., Perry, R. P., Chipperfield, J. G., Murayama, K., & Weiner, B. (2017). Attribution-based motivation treatment efficacy in an online learning environment for students who differ in cognitive elaboration. *Motivation and Emotion*. <https://doi.org/10.1007/s11031-017-9632-8>
- Harnischfeger, A., & Wiley, D. E. (1978). Conceptual issues in models of school learning. *Journal of Curriculum Studies*, 10(3), 215–231.
- Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. <https://doi.org/10.4324/9780203887332>
- Heckhausen, J., Wrosch, C., & Schulz, R. (2010). A Motivational Theory of Life-Span Development. *Psychological Review*. <https://doi.org/10.1037/a0017668>
- Heift, T., & Schulze, M. (2015). Tutorial computer-assisted language learning. *Language Teaching*. <https://doi.org/10.1017/S0261444815000245>
- Hershkovitz, A., Baker, R., Gowda, S. M., & Corbett, A. T. (2013). Predicting Future Learning Better Using Quantitative Analysis of Moment-by-Moment Learning. In *Proceedings of the 6th International Conference on Educational Data Mining*.
- Hertzberg, H. T. E., & Daniels, G. S. (1952). Air Force anthropology in 1950. *American Journal of Physical Anthropology*, 10(2), 201–208.



- Hickey, D., Jovanovic, J., Lonn, S., & Willis, J. E. (2015). 2nd int'l workshop on open badges in education (OBIE 2015): from learning evidence to learning analytics. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15*. <https://doi.org/10.1145/2723576.2723639>
- Hickey, D., Willis, J. E., Jovanovic, J., & Lonn, S. (2015). From Learning Evidence to Learning Analytics. In *2nd International workshop on open badges in education (OBIE 2015)*. <https://doi.org/10.1145/2723576.2723639>
- Hodges, C., & Kim, C. (2010). Email, self-regulation, self-efficacy, and achievement in a college online mathematics course. *Journal of Educational Computing Research*. <https://doi.org/10.2190/EC.43.2.d>
- Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*. <https://doi.org/10.1126/science.aal3856>
- Hong, C. M., Chen, C. M., Chang, M. H., & Chen, S. C. (2007). Intelligent web-based tutoring system with personalized learning path guidance. In *Proceedings - The 7th IEEE International Conference on Advanced Learning Technologies, ICALT 2007*. <https://doi.org/10.1109/ICALT.2007.167>
- Huang, J., Dasgupta, A., Ghosh, A., Manning, J., & Sanders, M. (2014). Superposter behavior in MOOC forums. In *Proceedings of the first ACM conference on Learning @ scale conference - L@S '14*. <https://doi.org/10.1145/2556325.2566249>
- Hubackova, S., & Semradova, I. (2016). Evaluation of Blended Learning. *Procedia - Social and Behavioral Sciences*. <https://doi.org/10.1016/j.sbspro.2016.02.044>
- Hulleman, C. S., Kosovich, J. J., Barron, K. E., & Daniel, D. B. (2017). Making connections: Replicating and extending the utility value intervention in the classroom. *Journal of Educational Psychology*. <https://doi.org/10.1037/edu0000146>
- Ifenthaler, D. (2015). Learning analytics.
- Iguyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*. <https://doi.org/10.1162/153244303322753616>
- Jaeggi, S. M., Buschkuhl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.1103228108>
- Jaeggi, Susanne M., Buschkuhl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. *Memory and Cognition*. <https://doi.org/10.3758/s13421-013-0364-z>
- Jaeggi, Susanne M., Karbach, J., & Strobach, T. (2017). Editorial Special Topic: Enhancing Brain and Cognition Through Cognitive Training. *Journal of Cognitive Enhancement*. <https://doi.org/10.1007/s41465-017-0057-9>
- Jaeggi, Susanne M., Buschkuhl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.0801268105>
- Jan, S. K. (2015). The relationships between academic self-efficacy, computer self-efficacy, prior experience, and satisfaction with online learning. *American Journal of Distance Education*. <https://doi.org/10.1080/08923647.2015.994366>
- Jasny, B. R., & Stone, R. (2017). Prediction and its limits. *Science*. <https://doi.org/10.1126/science.355.6324.468>
- Jordan, K. (2015). Massive open online course completion rates revisited: Assessment,

- length and attrition. *International Review of Research in Open and Distance Learning*. <https://doi.org/10.19173/irrodl.v16i3.2112>
- Jovanovic, J., & Devedzic, V. (2015). Open Badges: Novel Means to Motivate, Scaffold and Recognize Learning. *Technology, Knowledge and Learning*. <https://doi.org/10.1007/s10758-014-9232-6>
- Karbach, J., & Verhaeghen, P. (2014). Making Working Memory Work: A Meta-Analysis of Executive-Control and Working Memory Training in Older Adults. *Psychological Science*. <https://doi.org/10.1177/0956797614548725>
- Katz, B., Jaeggi, S., Buschkuhl, M., Stegman, A., & Shah, P. (2014). Differential effect of motivational features on training improvements in school-based cognitive training. *Frontiers in Human Neuroscience*. <https://doi.org/10.3389/fnhum.2014.00242>
- Khribi, M. K., Jemni, M., & Nasraoui, O. (2009). Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval. *Educational Technology and Society*. <https://doi.org/10.1109/ICALT.2008.198>
- Kitsantas, A., & Zimmerman, B. J. (2006). Enhancing self-regulation of practice: The influence of graphing and self-evaluative standards. *Metacognition and Learning*. <https://doi.org/10.1007/s11409-006-9000-7>
- Kivinen, O., & Piironen, T. (2018). The evolution of Homo Discens: natural selection and human learning. *Journal for the Theory of Social Behaviour*. <https://doi.org/10.1111/jtsb.12157>
- Kizilcec, R. F., & Halawa, S. (2015). Attrition and achievement gaps in online learning. In *L@S 2015 - 2nd ACM Conference on Learning at Scale*. <https://doi.org/10.1145/2724660.2724680>
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *Quarterly Journal of Economics*. <https://doi.org/10.1093/qje/qjx032>
- Klingberg, T. (2010). Training and plasticity of working memory. *Trends in Cognitive Sciences*. <https://doi.org/10.1016/j.tics.2010.05.002>
- Klingsieck, K. B., Fries, S., Horz, C., & Hofer, M. (2012). Procrastination in a distance university setting. *Distance Education*. <https://doi.org/10.1080/01587919.2012.723165>
- Kolowich, S. (2013). Coursera Takes a Nuanced View of MOOC Dropout Rates. *Chronicle of Higher Education*.
- Komarraju, M., Ramsey, A., & Rinella, V. (2013). Cognitive and non-cognitive predictors of college readiness and performance: Role of academic discipline. *Learning and Individual Differences*. <https://doi.org/10.1016/j.lindif.2012.12.007>
- Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of Intelligent Tutoring Systems. *Review of Educational Research*. <https://doi.org/10.3102/0034654315581420>
- Kumar, S. A., & Vijayalakshmi, M. N. (2012). Mining of student academic evaluation records in higher education. In *Proceedings of the 2012 International Conference on Recent Advances in Computing and Software Systems, RACSS 2012*. <https://doi.org/10.1109/RACSS.2012.6212699>
- Kuo, Y. C., Walker, A. E., Schroder, K. E. E., & Belland, B. R. (2014). Interaction, Internet self-efficacy, and self-regulated learning as predictors of student

- satisfaction in online education courses. *Internet and Higher Education*.  
<https://doi.org/10.1016/j.iheduc.2013.10.001>
- Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management*. <https://doi.org/10.1016/j.ijinfomgt.2014.02.002>
- Lack, K. A. (2013). Current status of research on online learning in postsecondary education. *Ithaka S+R*. <https://doi.org/10.18665/sr.22463>
- Larrabee Sønderlund, A., Hughes, E., & Smith, J. (2019). The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology*. <https://doi.org/10.1111/bjet.12720>
- Larrivee, B. (2000). Transforming Teaching Practice: Becoming the critically reflective teacher. *Reflective Practice*. <https://doi.org/10.1080/713693162>
- Lee, S. I., Lee, H., Abbeel, P., & Ng, A. Y. (2006). Efficient L 1 regularized logistic regression. In *Proceedings of the National Conference on Artificial Intelligence*.
- Lin, T., Imamiya, A., & Mao, X. (2008). Using multiple data sources to get closer insights into user cost and task performance. *Interacting with Computers*.  
<https://doi.org/10.1016/j.intcom.2007.12.002>
- Lin, Y. C., Liang, J. C., Yang, C. J., & Tsai, C. C. (2013). Exploring middle-aged and older adults' sources of Internet self-efficacy: A case study. *Computers in Human Behavior*. <https://doi.org/10.1016/j.chb.2013.07.017>
- Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*. <https://doi.org/10.1109/TIT.1982.1056489>
- Macgilchrist, F. (2019). Cruel optimism in edtech: when the digital data practices of educational technology providers inadvertently hinder educational equity. *Learning, Media and Technology*. <https://doi.org/10.1080/17439884.2018.1556217>
- Madigan, R. J., & Bollenbach, A. K. (1986). The effects of induced mood on irrational thoughts and views of the world. *Cognitive Therapy and Research*, 10(5), 547–562.
- Mah, D. K. (2016). Learning Analytics and Digital Badges: Potential Impact on Student Retention in Higher Education. *Technology, Knowledge and Learning*.  
<https://doi.org/10.1007/s10758-016-9286-8>
- Manek, S., Vijay, S., & Kamthania, D. (2016). Educational data mining - a case study. *International Journal of Information and Decision Sciences*, 8(2), 187–201.  
<https://doi.org/10.1504/IJIDS.2016.076517>
- Mangaroska, K., & Giannakos, M. (2019). Learning Analytics for Learning Design: A Systematic Literature Review of Analytics-Driven Design to Enhance Learning. *IEEE Transactions on Learning Technologies*.  
<https://doi.org/10.1109/TLT.2018.2868673>
- Mawjee, K., Woltering, S., Lai, N., Gotlieb, H., Kronitz, R., & Tannock, R. (2017). Working Memory Training in ADHD: Controlling for Engagement, Motivation, and Expectancy of Improvement (Pilot Study). *Journal of Attention Disorders*.  
<https://doi.org/10.1177/1087054714557356>
- Mayer, R. E. (2014). Incorporating motivation into multimedia learning. *Learning and Instruction*. <https://doi.org/10.1016/j.learninstruc.2013.04.003>
- McCuaig, J., & Baldwin, J. (2012). Identifying successful learners from interaction behaviour. In *Proceedings of the 5th International Conference on Educational Data Mining, EDM 2012*.

- McPartlan, P. (2020). *Motivation for and within online college courses. Dissertation Abstracts International Section A: Humanities and Social Sciences.*
- Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2009). Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies. Retrieved from <http://repository.alt.ac.uk/629/>
- Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of Math Anxiety and Its Influence on Young Adolescents' Course Enrollment Intentions and Performance in Mathematics. *Journal of Educational Psychology*. <https://doi.org/10.1037/0022-0663.82.1.60>
- Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology*. <https://doi.org/10.1037/a0028228>
- Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (1983). *Machine learning: An artificial intelligence approach. Machine Learning: An Artificial Intelligence Approach.* <https://doi.org/10.1109/MWSYM.2007.380062>
- Miltiadou, M., & Savenye, W. C. (2003). Applying social cognitive constructs of motivation to enhance student success in online distance education. *AACE Journal*.
- Mohamed, A., Husain, W., & Rashid, A. (2015). The Third Information Systems International Conference A Review on Predicting Student 's Performance using Data Mining Techniques. *Procedia - Procedia Computer Science*. <https://doi.org/10.1016/j.procs.2015.12.157>
- Moore, B. R. (2004). The evolution of learning. *Biological Reviews of the Cambridge Philosophical Society*. <https://doi.org/10.1017/S1464793103006225>
- Moreno, R., & Mayer, R. (2007). Interactive Multimodal Learning Environments. *Educational Psychology Review*. <https://doi.org/10.1007/s10648-007-9047-2>
- Nadelson, L. S., Semmelroth, C., Martinez, G., Featherstone, M., Fuhrman, C. A., & Sell, A. (2013). Why Did They Come Here? –The Influences and Expectations of First-Year Students' College Experience. *Higher Education Studies*. <https://doi.org/10.5539/hes.v3n1p50>
- Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*. [https://doi.org/10.1016/0364-0213\(90\)90024-Q](https://doi.org/10.1016/0364-0213(90)90024-Q)
- Nickel, C. E., & Overbaugh, R. C. (2012). Cooperative and collaborative strategies in blended and online learning environments. In *Educational Communities of Inquiry: Theoretical Framework, Research and Practice*. <https://doi.org/10.4018/978-1-4666-2110-7.ch012>
- Nicol, D., & MacFarlane-Dick, D. (2006). Formative assessment and selfregulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*. <https://doi.org/10.1080/03075070600572090>
- Norcini, J. (2010). The power of feedback. *Medical Education*. <https://doi.org/10.1111/j.1365-2923.2009.03542.x>
- Nussbaumer, A., Hillemann, E.-C., Gütl, C., & Albert, D. (2015). A Competence-based Service for Supporting Self-Regulated Learning in Virtual Environments. *Journal of Learning Analytics*. <https://doi.org/10.18608/jla.2015.21.6>
- Olesen, P. J., Westerberg, H., & Klingberg, T. (2004). Increased prefrontal and parietal activity after training of working memory. *Nature Neuroscience*, 7(1), 75–79. <https://doi.org/10.1038/nn1165>
- Open Science Collobaration. (2015). Estimating the reproducibility of psychological

- science: Open Science Collobaration. *Science*.  
<https://doi.org/10.1126/science.aac4716>
- Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., & Mirriahi, N. (2019). Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*. <https://doi.org/10.1111/bjet.12592>
- 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 2018*.
- Parker, P. C., Perry, R. P., Chipperfield, J. G., Hamm, J. M., & Pekrun, R. (2018). An attribution-based motivation treatment for low control students who are bored in online learning environments. *Motivation Science*.  
<https://doi.org/10.1037/mot0000081>
- Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2009). Learning styles concepts and evidence. *Psychological Science in the Public Interest, Supplement*.  
<https://doi.org/10.1111/j.1539-6053.2009.01038.x>
- Patterson, B., & McFadden, C. (2009). Attrition in online and campus degree programs. *Online Journal of Distance Learning Administration*.
- Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*.  
<https://doi.org/10.1016/j.eswa.2013.08.042>
- Pergher, V., Shalchy, M. A., Pahor, A., Van Hulle, M. M., Jaeggi, S. M., & Seitz, A. R. (2019). Divergent Research Methods Limit Understanding of Working Memory Training. *Journal of Cognitive Enhancement*, 1–21.
- Perry, R. P., & Hamm, J. M. (2016). An Attribution Perspective on Competence and Motivation BT - Hanbook of Competence and Motivation Theory and Applications. *Hanbook of Competence and Motivation Theory and Applications*.
- Picciano, A. G. (2009). Blending with purpose: The multimodal model. *Journal of Asynchronous Learning Network*. <https://doi.org/10.4013/base.2011.84.02>
- Porter, A. C., & Polikoff, M. S. (2012). Measuring Academic Readiness for College. *Educational Policy*. <https://doi.org/10.1177/0895904811400410>
- Premlatha, K. R., Dharani, B., & Geetha, T. V. (2016). Dynamic learner profiling and automatic learner classification for adaptive e-learning environment. *Interactive Learning Environments*. <https://doi.org/10.1080/10494820.2014.948459>
- Qualtrics. (2016). Qualtrics: The World's Leading Research & Insights Platform.  
<https://doi.org/10.3109/03639045.2013.790903>
- Rakes, G. C., & Dunn, K. E. (2015). Teaching Online: Discovering Teacher Concerns. *Journal of Research on Technology in Education*.  
<https://doi.org/10.1080/15391523.2015.1063346>
- Rakes, G. C., Dunn, K. E., & Rakes, T. A. (2013). Attribution as a predictor of procrastination in online graduate students. *Journal of Interactive Online Learning*.
- Ramesh, A., Goldwasser, D., Huang, B., Daume, H., & Getoor, L. (2014). Learning latent engagement patterns of students in online courses. In *Proceedings of the National Conference on Artificial Intelligence*.
- Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries. *Proceedings of the First Instructional Conference on Machine Learning*.
- Reid, D. K., & Stone, C. A. (1991). Why Is Cognitive Instruction Effective? Underlying

- Learning Mechanisms. *Remedial and Special Education*.  
<https://doi.org/10.1177/074193259101200303>
- Reinig, M. (2010). The theory and practice of online learning. *Language, Learning and Technology*.
- Reiss, S. (2012). Intrinsic and Extrinsic Motivation. *Teaching of Psychology*.  
<https://doi.org/10.1177/0098628312437704>
- Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*. <https://doi.org/10.1037/a0026838>
- Rivard, R. (2013). Measuring the MOOC dropout rate. *Inside Higher Ed*.
- Rizzardini, R. H., Chan, M. M., & Guetl, C. (2016). An Attrition Model for MOOCs: Evaluating the Learning Strategies of Gamification. In *Formative Assessment, Learning Data Analytics and Gamification: In ICT Education*.  
<https://doi.org/10.1016/B978-0-12-803637-2.00014-2>
- Roca, J. C., & Gagné, M. (2008). Understanding e-learning continuance intention in the workplace: A self-determination theory perspective. *Computers in Human Behavior*.  
<https://doi.org/10.1016/j.chb.2007.06.001>
- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*. <https://doi.org/10.1145/2684822.2685324>
- Rogowsky, B. A., Calhoun, B. M., & Tallal, P. (2015). Matching learning style to instructional method: Effects on comprehension. *Journal of Educational Psychology*. <https://doi.org/10.1037/a0037478>
- Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*. <https://doi.org/10.1016/j.eswa.2006.04.005>
- Romero, Cristobal, & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*.  
<https://doi.org/10.1002/widm.1075>
- Romero, Cristobal, & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. <https://doi.org/10.1002/widm.1355>
- Saavedra, J. (2020). Educational challenges and opportunities of the Coronavirus (COVID-19) pandemic. Retrieved December 4, 2020, from <https://blogs.worldbank.org/education/educational-challenges-and-opportunities-covid-19-pandemic>
- Salas, D. J., Baldiris, S., Fabregat, R., & Graf, S. (2016). Supporting the acquisition of scientific skills by the use of learning analytics. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [https://doi.org/10.1007/978-3-319-47440-3\\_32](https://doi.org/10.1007/978-3-319-47440-3_32)
- Schapire, R. E. (2003). The Boosting Approach to Machine Learning: An Overview. [https://doi.org/10.1007/978-0-387-21579-2\\_9](https://doi.org/10.1007/978-0-387-21579-2_9)
- Scheffel, M., Drachsler, H., Stoyanov, S., & Specht, M. (2014). Quality indicators for learning analytics. *Journal of Educational Technology & Society*, 17(4), 117–132.
- Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence.
- Schulz, R., & Heckhausen, J. (1996). A Life Span Model of Successful Aging. *American*

- Psychologist*. <https://doi.org/10.1037/0003-066X.51.7.702>
- Schunk, D. H., & Zimmerman, B. J. (2012). *Motivation and self-regulated learning: Theory, research, and applications*. *Motivation and Self-Regulated Learning: Theory, Research, and Applications*. <https://doi.org/10.4324/9780203831076>
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*. <https://doi.org/10.1037/0003-066X.54.2.93>
- Seaman, J., Allen, I. E., & Seaman, J. (2018). Grade increase: Tracking distance education in the United States. *Babson Survey Research Group*.
- Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P., & Pritchard, D. E. (2014). Who does what in a massive open online course? *Communications of the ACM*. <https://doi.org/10.1145/2500876>
- Sembing, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. (2011). PREDICTION OF STUDENT ACADEMIC PERFORMANCE BY AN APPLICATION OF DATA MINING TECHNIQUES. In *MANAGEMENT AND ARTIFICIAL INTELLIGENCE*.
- Shanks, D. R., Holyoak, K. J., & Medin, D. L. (1996). *The psychology of learning and motivation: advances in research and theory*. *Causal learning*. Academic Press.
- Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*. <https://doi.org/10.1037/a0027473>
- Shum, S. B., & Ferguson, R. (2012). Social learning analytics. *Educational Technology and Society*. <https://doi.org/10.1145/2330601.2330616>
- Škrinjarčić, B. (2014). William G. Bowen: Higher Education in the Digital Age. *Croatian Economic Survey*. <https://doi.org/10.15179/ces.16.1.7>
- So, H. J., & Brush, T. A. (2008). Student perceptions of collaborative learning, social presence and satisfaction in a blended learning environment: Relationships and critical factors. *Computers and Education*. <https://doi.org/10.1016/j.compedu.2007.05.009>
- Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training revisited: A multi-level meta-analysis of n-back training studies. *Psychonomic Bulletin and Review*. <https://doi.org/10.3758/s13423-016-1217-0>
- Tan-Wilson, A., & Stamp, N. (2015). College students' views of work–life balance in STEM research careers: Addressing negative preconceptions. *CBE Life Sciences Education*. <https://doi.org/10.1187/cbe.14-11-0210>
- Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. In *Data Classification: Algorithms and Applications*. <https://doi.org/10.1201/b17320>
- Tempelaar, D. T., Rienties, B., & Giesbers, B. (2011). Student Learning Preferences in a Blended Learning Environment: Investigating the Relationship Between Tool Use and Learning Approaches. In *Building Learning Experiences in a Changing World*. [https://doi.org/10.1007/978-94-007-0802-0\\_12](https://doi.org/10.1007/978-94-007-0802-0_12)
- Thomas, P. R., & McKay, J. B. (2010). Cognitive styles and instructional design in university learning. *Learning and Individual Differences*. <https://doi.org/10.1016/j.lindif.2010.01.002>
- Vallerand, R. J., & Blissonnette, R. (1992). Intrinsic, Extrinsic, and Amotivational Styles as Predictors of Behavior: A Prospective Study. *Journal of Personality*. <https://doi.org/10.1111/j.1467-6494.1992.tb00922.x>
- Van Bruggen, J. (2005). Theory and practice of online learning. *British Journal of Educational Technology*. [https://doi.org/10.1111/j.1467-8535.2005.00445\\_1.x](https://doi.org/10.1111/j.1467-8535.2005.00445_1.x)

- Varun, D., & Chadha, A. (2011). An Empirical Study of the Applications of Data Mining Techniques in Higher Education. *International Journal of Advanced Computer Science and Applications*. <https://doi.org/10.14569/ijacsa.2011.020314>
- Veltri, G. A. (2017). Big Data is not only about data: The two cultures of modelling. *Big Data and Society*. <https://doi.org/10.1177/2053951717703997>
- Venezia, A., & Jaeger, L. (2013). Transitions from high school to college. *Future of Children*. <https://doi.org/10.1353/foc.2013.0004>
- Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Van Assche, F., Parra, G., & Klerkx, J. (2014). Learning dashboards: An overview and future research opportunities. *Personal and Ubiquitous Computing*. <https://doi.org/10.1007/s00779-013-0751-2>
- Vesin, B., Mangaroska, K., & Giannakos, M. (2018). Learning in smart environments: user-centered design and analytics of an adaptive learning system. *Smart Learning Environments*. <https://doi.org/10.1186/s40561-018-0071-0>
- Visser, J. (2012). Reflections on a definition: Revisiting the meaning of learning. In *Second International Handbook of Lifelong Learning*. [https://doi.org/10.1007/978-94-007-2360-3\\_12](https://doi.org/10.1007/978-94-007-2360-3_12)
- Wager, T. D., & Smith, E. E. (2003). Neuroimaging studies of working memory: A meta-analysis. *Cognitive, Affective and Behavioral Neuroscience*. <https://doi.org/10.3758/CABN.3.4.255>
- Walberg, H.J. (1981). A psychological theory of educational productivity. *Psychology and Education*, Ed. F.H. Farley and N. Gordon.
- Walberg, Herbert J. (1984). Improving the Productivity of America's Schools. *Educational Leadership*.
- Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a Knowledge Base for School Learning. *Review of Educational Research*. <https://doi.org/10.3102/00346543063003249>
- Wickelgren, W. A. (1981). HUMAN LEARNING AND MEMORY. *Ann. Rev. Psychol* (Vol. 32). Retrieved from [www.annualreviews.org](http://www.annualreviews.org)
- Wilkowski, J., Deutsch, A., & Russell, D. M. (2014). Student skill and goal achievement in the mapping with google MOOC. In *Proceedings of the first ACM conference on Learning @ scale conference - L@S '14*. <https://doi.org/10.1145/2556325.2566240>
- Winne, P. H., Adesope, S., Code, J., Gress, C., Jordanov, M., Kumar, V., ... Shi, B. (2006). The learning kit project: Advancing research on learning as learners learn in everyday settings. In *Proceedings - Sixth International Conference on Advanced Learning Technologies, ICALT 2006*. <https://doi.org/10.1109/icalt.2006.1652607>
- Witten, I H, Frank, E., & Hall, M. A. (2005). *Data Mining Practical Machine Learning Tools And Techniques*. *Data Mining*. [https://doi.org/10.1007/0120884070\\_9780120884070](https://doi.org/10.1007/0120884070_9780120884070)
- Witten, Ian H., Frank, E., & Hall, M. A. (2011). Introduction to Weka. In *Data Mining: Practical Machine Learning Tools and Techniques* (pp. 403–406). <https://doi.org/10.1016/b978-0-12-374856-0.00010-9>
- Witten, Ian H, Frank, E., & Hall, M. a. (2011). *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook)*. Complementary literature None.
- Xu, H., Caramanis, C., & Mannor, S. (2009). Robust regression and Lasso. In *Advances in Neural Information Processing Systems 21 - Proceedings of the 2008 Conference*.
- Yang, G. (2018). Understanding Continuous Use Intention of MOOCs -A Perspective



- from Subjective Task Value BT - 2018 4th International Conference on Social Science and Higher Education (ICSSHE 2018). Atlantis Press.  
<https://doi.org/https://doi.org/10.2991/icsshe-18.2018.178>
- Zimmerman, B. J. (1989). A Social Cognitive View of Self-Regulated Academic Learning. *Journal of Educational Psychology*. <https://doi.org/10.1037/0022-0663.81.3.329>
- Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In *Handbook of Self-Regulation*. <https://doi.org/10.1016/B978-012109890-2/50031-7>
- Zins, J. E., Weissberg, R. P., Wang, M. C., & Walberg, H. J. (2004). *Building Academic Success on Social and Emotional Learning: What Does the Research Say?* Teachers College Press. <https://doi.org/10.1525/mp.2012.30.1.85>

# Appendix A

Table A.1

*Table of survey measures used to measure course-level learner-centric measures.*

Name	p	p	Construct	Item	Response Values	Item note
Stem - In Github Codebo ok	r	o				
	e	s				
	t					
se1	x	x	Self-Efficacy	I'm certain I can master the skills taught in this course	slider: 1=Not true at all, 5 = Very true	
se2	x	x	Self-Efficacy	I'm certain I can figure out how to learn even the most difficult material in this course	slider: 1=Not true at all, 5 = Very true	
se3	x	x	Self-Efficacy	I can do almost all the work in this course if I don't give up	slider: 1=Not true at all, 5 = Very true	
se4	x	x	Self-Efficacy	Even if the work in this course is hard, I can learn it	slider: 1=Not true at all, 5 = Very true	
se5	x	x	Self-Efficacy	I can do even the hardest work in this course if I try	slider: 1=Not true at all, 5 = Very true	
wgrade	x	x	Grade Expectations	What grade do you want to get in this course?	13=A+, 12=A, 11=A-, 10=B+, 9=B, 8=B-, 7=C+, 6=C, 5=C-, 4=D+, 3=D, 2=D-, 1=F	*0-100% slider at post-survey
egrade	x	x	Grade Expectations	What grade do you expect to get in this course?	13=A+, 12=A, 11=A-, 10=B+, 9=B, 8=B-, 7=C+, 6=C, 5=C-, 4=D+, 3=D, 2=D-, 1=F	*0-100% slider at post-survey
badgrade	x	x	Grade Expectations	What is the worst grade you would still be satisfied with getting in this course?	13=A+, 12=A, 11=A-, 10=B+, 9=B, 8=B-, 7=C+, 6=C, 5=C-, 4=D+, 3=D, 2=D-, 1=F	*0-100% slider at post-survey
wgradef	x		Grade Expectations	Think about your grade on the final - What grade do you now want to get?	slider: 0-100%	
egradef	x		Grade Expectations	Think about your grade on the final - What grade do you now expect to get?	slider: 0-100%	

Name Stem - In Github Codebook	p r e s t	p o s t	Construct	Item	Response Values	Item note
badgra def		x	Grade Expectations	Think about your grade on the final - What's the worst grade you will still be satisfied with?	slider: 0-100%	
orsh1	x	x	Online Self Regulation	how often do you...work in a place where you can read and work on assignments without distractions?	slider: 1=Never, 5=All the time	
orsh2	x	x	Online Self Regulation	how often do you...ignore distractions around you when you study?	slider: 1=Never, 5=All the time	
orsh3	x		Online Self Regulation	how often do you...spend 20 hours a week on this course?	slider: 1=Never, 5=All the time	
orsh4	x	x	Online Self Regulation	how often do you...keep a record of what your assignments are and when they are due?	slider: 1=Never, 5=All the time	
orsh5	x	x	Online Self Regulation	how often do you...plan your work in advance so that you can turn in your assignments on time?	slider: 1=Never, 5=All the time	
orsh6	x	x	Online Self Regulation	how often do you...make sure people around you will help you study and not try to distract you?	slider: 1=Never, 5=All the time	
orsh7	x	x	Online Self Regulation	how often do you...use email and other online tools to ask your classmates and instructors questions?	slider: 1=Never, 5=All the time	
olsrl1	x	x	Online Self Regulation	stay on task when studying on your computer	slider: 1=Never, 7=All the time	
olsrl2	x	x	Online Self Regulation	get distracted when studying on your computer	slider: 1=Never, 7=All the time	*negatively-coded
olsrl3	x	x	Online Self Regulation	find yourself getting distracted by social media while studying on your computer	slider: 1=Never, 7=All the time	*negatively-coded
olsrl4	x	x	Online Self Regulation	find yourself getting distracted by internet browsing	slider: 1=Never, 7=All the time	*negatively-coded
olsrl5	x	x	Online Self Regulation	find yourself getting distracted by your phone while studying on your computer	slider: 1=Never, 7=All the time	*negatively-coded
onlexp1	x		Online Expectations	When taking an online course, I expect to perform...	slider: 1=Not at all well, 7 = Very well	
onlexp2	x		Online Expectations	How good would you be at learning something new in an online course?	slider: 1=Not at all good, 7 = Very good	
er1	x	x	Effort Regulation	I often feel so lazy or bored when I study for this class that I quit before I finish what I planned to	slider: 1=Strongly disagree, 5=Strongly	*negatively-coded

Name Stem - In Github Codebook	p	p	Construct	Item	Response Values	Item note
				do	Agree	
er2	x	x	Effort Regulation	I work hard even if I do not like what I am doing	slider: 1=Strongly disagree, 5=Strongly Agree	
er3	x	x	Effort Regulation	When coursework is difficult, I give up or only study the easy parts	slider: 1=Strongly disagree, 5=Strongly Agree	*negatively-coded
er4	x	x	Effort Regulation	Even when course materials are dull and uninteresting, I manage to keep working until I finish	slider: 1=Strongly disagree, 5=Strongly Agree	
util1	x	x	Utility Value	How beneficial for your daily life is understanding the biology and chemistry of cooking?	slider: 1=Not beneficial at all, 7 = Very beneficial	has highest loading among all utility items
util2	x	x	Utility Value	How useful in everyday life and leisure time is knowledge of biology and chemistry of cooking?	slider: 1=Not at all useful, 7 = Very useful	
util3	x	x	Utility Value	How applicable in everyday life is knowledge of biology and chemistry of cooking?	slider: 1=Not at all applicable, 7 = Very applicable	
util4	x	x	Utility Value	How much will you be able to impress others with your knowledge of the biology and chemistry of cooking?	slider: 1 = Not at all, 7 = A lot	
util5	x	x	Utility Value	How important is it to you to get a good grade in this course for your academic career?	slider: 1=Not at all important, 7 = Very important	
int1	x	x	Interest Value	How often do you wonder about the science behind cooking?	slider: 1 = Never, 7 = Very often	
int2	x	x	Interest Value	How curious are you to learn about the science behind cooking?	slider: 1=Not at all curious, 7 = Very curious	
int3	x	x	Interest Value	How interested are you in the science behind food and cooking?	slider: 1=Not at all interested, 7 = Very interested	has highest loading among all interest items
int4	x	x	Interest Value	How much fun will learning about the biology and chemistry of cooking be?	slider: 1=Not at all fun, 7 = Very fun	

Name Stem - In Github Codebo ok	p r e t	p o s t	Construct	Item	Response Values	Item note
att1	x	x	Attainment Value	How important to you, personally, is it to be a person who understands the science behind cooking?	slider: 1=Not at all important, 7 = Very important	
att2	x	x	Attainment Value	How important is it that others see you as knowledgeable about the science behind food and cooking?	slider: 1=Not at all important, 7 = Very important	
att3	x	x	Attainment Value	How important to your identity is it to be knowledgeable about the science behind food and cooking?	slider: 1=Not at all important, 7 = Very important	has highest loading among all attainment items
cost1	x	x	Cost Value (Emotional)	How stressful will this class be?	slider: 1=Not at all stressful, 7 = Very stressful	
cost2	x	x	Cost Value (Emotional)	How frustrating will this class be?	slider: 1=Not at all frustrating, 7 = Very frustrating	has highest loading among all cost (emotional) items
cost3	x	x	Cost Value (Emotional)	How emotionally draining will this class be?	slider: 1=Not at all draining, 7 = Very draining	
cost4	x	x	Cost Value (Loss of Valued Alternatives)	How much do you have to sacrifice to do well in this course?	slider: 1=Nothing, 7 = An incredible amount	
cost5	x	x	Cost Value (Loss of Valued Alternatives)	How many other valued activities does this class require you to give up?	slider: 1=None, 7 = An incredible amount	
cost6	x	x	Cost Value (Loss of Valued Alternatives)	How many opportunities will you be missing out on if you commit fully to this class?	slider: 1=None, 7 = An incredible amount	has highest loading among all cost (lova) items
cost7	x	x	Cost Value (Outside Effort)	How much will your other commitments get in the way of you putting forth effort in class?	slider: 1=Not at all, 7 = Completely	
cost8	x	x	Cost Value (Outside Effort)	How much time will you have for this class after taking care of more important activities?	slider: 1=Not nearly enough, 7 = Enough	*negatively-coded
cost9	x	x	Cost Value (Outside Effort)	How much effort will you have left for this class after taking care of more important activities?	slider: 1=Not nearly enough, 7 = Enough	*negatively-coded, has highest loading among all cost (outside effort) items
studyda	x	x	Course Plan	How many days of each week will you work on	1 to 7	

Name Stem - In Github Codebo ok	p r e t	p o s t	Construct	Item	Response Values	Item note
ys				this course?		
studyplan_ chg	x		Course Plan	Think about the study plan that you had at the beginning of the course. Did you end up changing your study plan?	0=No, I stuck to my study plan, 1=I never had a study plan, 2=Yes, I changed my study plan a bit, 3=Yes, I changed my study plan a lot	
courses	x	x	Other Courses	How many other courses are you taking this summer?	1=0, 2=1, 3=2, 4=3+	*at post-survey, changed to "compared to other courses you took in Summer Session 1..."
courses _imp	x	x	Other Courses	Compared to other courses you are currently taking, how important is this course?	4=Most important, 3=Second-most important, 2=Third-most important, 1=Fourth-most important	*available choices depend on previous question
oact1	x		Other Activities List	What other important activities do you plan on doing in July while completing this course? (e.g., working for pay, caring for family members, taking another course, playing sports, completing home projects, etc.)	open-ended	
oact2	x		Other Activities List	What other important activities do you plan on doing in July while completing this course? (e.g., working for pay, caring for family members, taking another course, playing sports, completing home projects, etc.)	open-ended	
oact3	x		Other Activities List	What other important activities do you plan on doing in July while completing this course? (e.g., working for pay, caring for family members, taking another course, playing sports, completing home projects, etc.)	open-ended	

Name Stem - In Github Codebo ok	p r e t	p o s t	Construct	Item	Response Values	Item note
oact4	x		Other Activities List	What other important activities do you plan on doing in July while completing this course? (e.g., working for pay, caring for family members, taking another course, playing sports, completing home projects, etc.)	open-ended	
oact5	x		Other Activities List	What other important activities do you plan on doing in July while completing this course? (e.g., working for pay, caring for family members, taking another course, playing sports, completing home projects, etc.)	open-ended	
oact6	x		Other Activities List	What other important activities do you plan on doing in July while completing this course? (e.g., working for pay, caring for family members, taking another course, playing sports, completing home projects, etc.)	open-ended	
oact7	x		Other Activities List	What other important activities do you plan on doing in July while completing this course? (e.g., working for pay, caring for family members, taking another course, playing sports, completing home projects, etc.)	open-ended	
oact8	x		Other Activities List	What other important activities do you plan on doing in July while completing this course? (e.g., working for pay, caring for family members, taking another course, playing sports, completing home projects, etc.)	open-ended	
oact9	x		Other Activities List	What other important activities do you plan on doing in July while completing this course? (e.g., working for pay, caring for family members, taking another course, playing sports, completing home projects, etc.)	open-ended	
oact10	x		Other Activities List	What other important activities do you plan on doing in July while completing this course? (e.g., working for pay, caring for family members, taking another course, playing sports, completing home	open-ended	

Name Stem - In Github Codebo ok	p r e t	p o s t	Construct	Item	Response Values	Item note
				projects, etc.)		
oact11	x		Other Activities List	During Summer Session 1, were there any other important activities you that you didn't plan to do but ended up spending a lot of time on? (e.g., working for pay, caring for family members, taking another course, playing sports, completing home projects, etc.) - .	open-ended	
oact12	x		Other Activities List	During Summer Session 1, were there any other important activities you that you didn't plan to do but ended up spending a lot of time on? (e.g., working for pay, caring for family members, taking another course, playing sports, completing home projects, etc.) - .	open-ended	
oact13	x		Other Activities List	During Summer Session 1, were there any other important activities you that you didn't plan to do but ended up spending a lot of time on? (e.g., working for pay, caring for family members, taking another course, playing sports, completing home projects, etc.) - .	open-ended	
oact14	x		Other Activities List	During Summer Session 1, were there any other important activities you that you didn't plan to do but ended up spending a lot of time on? (e.g., working for pay, caring for family members, taking another course, playing sports, completing home projects, etc.) - .	open-ended	
oact15	x		Other Activities List	During Summer Session 1, were there any other important activities you that you didn't plan to do but ended up spending a lot of time on? (e.g., working for pay, caring for family members, taking another course, playing sports, completing home projects, etc.) - .	open-ended	
oactcomp 1	x		Other Activities Completion	At the beginning of the course, you said you planned on doing the activities below. Did you end up doing them?	0=No, 1=Yes	



Name Stem - In Github Codebo ok	p r e t	p o s t	Construct	Item	Response Values	Item note
oactcomp 2	x		Other Activities Completion	At the beginning of the course, you said you planned on doing the activities below. Did you end up doing them?	0=No, 1=Yes	
oactcomp 3	x		Other Activities Completion	At the beginning of the course, you said you planned on doing the activities below. Did you end up doing them?	0=No, 1=Yes	
oactcomp 4	x		Other Activities Completion	At the beginning of the course, you said you planned on doing the activities below. Did you end up doing them?	0=No, 1=Yes	
oactcomp 5	x		Other Activities Completion	At the beginning of the course, you said you planned on doing the activities below. Did you end up doing them?	0=No, 1=Yes	
oactcomp 6	x		Other Activities Completion	At the beginning of the course, you said you planned on doing the activities below. Did you end up doing them?	0=No, 1=Yes	
oactcomp 7	x		Other Activities Completion	At the beginning of the course, you said you planned on doing the activities below. Did you end up doing them?	0=No, 1=Yes	
oactcomp 8	x		Other Activities Completion	At the beginning of the course, you said you planned on doing the activities below. Did you end up doing them?	0=No, 1=Yes	
oactcomp 9	x		Other Activities Completion	At the beginning of the course, you said you planned on doing the activities below. Did you end up doing them?	0=No, 1=Yes	
oactcomp 10	x		Other Activities Completion	At the beginning of the course, you said you planned on doing the activities below. Did you end up doing them?	0=No, 1=Yes	
courser ank	x	x	Other Activities Rank	Please drag and drop your responsibilities during this course in order from most important to least important	1 to 16	
oact1hr s	x	x	Other Activities Time	On average, how many hours per week will you spend on each of these activities in the month of July?	0 to 40	*only appears if and something is listed in this space on previous question

Name Stem - In Github Codebook	p	p	Construct	Item	Response Values	Item note
oact2hrs	x	x	Other Activities Time	On average, how many hours per week will you spend on each of these activities in the month of July?	0 to 40	*only appears if and something is listed in this space on previous question
oact3hrs	x	x	Other Activities Time	On average, how many hours per week will you spend on each of these activities in the month of July?	0 to 40	*only appears if and something is listed in this space on previous question
oact4hrs	x	x	Other Activities Time	On average, how many hours per week will you spend on each of these activities in the month of July?	0 to 40	*only appears if and something is listed in this space on previous question
oact5hrs	x	x	Other Activities Time	On average, how many hours per week will you spend on each of these activities in the month of July?	0 to 40	*only appears if and something is listed in this space on previous question
oact6hrs	x	x	Other Activities Time	On average, how many hours per week will you spend on each of these activities in the month of July?	0 to 40	*only appears if and something is listed in this space on previous question
oact7hrs	x	x	Other Activities Time	On average, how many hours per week will you spend on each of these activities in the month of July?	0 to 40	*only appears if and something is listed in this space on previous question
oact8hrs	x	x	Other Activities Time	On average, how many hours per week will you spend on each of these activities in the month of July?	0 to 40	*only appears if and something is listed in this space on previous question
oact9hrs	x	x	Other Activities Time	On average, how many hours per week will you spend on each of these activities in the month of July?	0 to 40	*only appears if and something is listed in this space on previous question
oact10hrs	x	x	Other Activities Time	On average, how many hours per week will you spend on each of these activities in the month of	0 to 40	*only appears if and something is listed

Name Stem - In Github Codebo ok	p r e t	p o s t	Construct	Item	Response Values	Item note
				July?		in this space on previous question
oact11hrs	x		Other Activities Time	On average, how many hours per week will you spend on each of these activities in the month of July?	0 to 40	*only appears if and something is listed in this space on previous question
oact12hrs	x		Other Activities Time	On average, how many hours per week will you spend on each of these activities in the month of July?	0 to 40	*only appears if and something is listed in this space on previous question
oact13hrs	x		Other Activities Time	On average, how many hours per week will you spend on each of these activities in the month of July?	0 to 40	*only appears if and something is listed in this space on previous question
oact14hrs	x		Other Activities Time	On average, how many hours per week will you spend on each of these activities in the month of July?	0 to 40	*only appears if and something is listed in this space on previous question
oact15hrs	x		Other Activities Time	On average, how many hours per week will you spend on each of these activities in the month of July?	0 to 40	*only appears if and something is listed in this space on previous question
sp1	x		Consciousness	I see myself as someone who ... does a thorough job	1 = Strongly disagree, 3 = Neither agree nor disagree, 5 = Strongly agree	
sp2	x		Consciousness	I see myself as someone who ... can be somewhat careless	1 = Strongly disagree, 3 = Neither agree nor disagree, 5 = Strongly agree	*negatively-coded
sp3	x		Consciousness	I see myself as someone who ... is a reliable worker	1 = Strongly disagree, 3 = Neither agree nor disagree, 5 = Strongly agree	

Name Stem - In Github Codebo ok	p r e s e n t	p o s i t i v e	Construct	Item	Response Values	Item note
sp4	x		Consciousness	I see myself as someone who ... tends to be disorganized	1 = Strongly disagree, 3 = Neither agree nor disagree, 5 = Strongly agree	*negatively-coded
sp5	x		Consciousness	I see myself as someone who ... tends to be lazy	1 = Strongly disagree, 3 = Neither agree nor disagree, 5 = Strongly agree	*negatively-coded
sp6	x		Consciousness	I see myself as someone who ... perseveres until the task is finished	1 = Strongly disagree, 3 = Neither agree nor disagree, 5 = Strongly agree	