Predicting Student Success in a Self-Paced Mathematics MOOC

by

James Allan Cunningham

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved March 2017 by the
Graduate Supervisory Committee:

Gary Bitter, Chair
Rebecca Barber
Ian Douglas

ARIZONA STATE UNIVERSITY

May 2017

ABSTRACT

While predicting completion in Massive Open Online Courses (MOOCs) has been an active area of research in recent years, predicting completion in self-paced MOOCS, the fastest growing segment of open online courses, has largely been ignored. Using learning analytics and educational data mining techniques, this study examined data generated by over 4,600 individuals working in a self-paced, open enrollment college algebra MOOC over a period of eight months.

Although just 4% of these students completed the course, models were developed that could predict correctly nearly 80% of the time which students would complete the course and which would not, based on each student's first day of work in the online course. Logistic regression was used as the primary tool to predict completion and focused on variables associated with self-regulated learning (SRL) and demographic variables available from survey information gathered as students begin edX courses (the MOOC platform employed).

The strongest SRL predictor was the amount of time students spent in the course on their first day. The number of math skills obtained the first day and the pace at which these skills were gained were also predictors, although pace was negatively correlated with completion. Prediction models using only SRL data obtained on the first day in the course correctly predicted course completion 70% of the time, whereas models based on first-day SRL and demographic data made correct predictions 79% of the time.

To my wife, Michele, you are the best of the best.

ACKNOWLEDGMENTS

Dr. Adrian Sannier pulled me aside at the ASU-GSV Summit in San Diego along with the chair of my committee, Dr. Bitter, and insisted that I analyze the data from his college algebra MOOC for my dissertation. I already had other plans and was nearly ready to present my research proposal to my committee. I hesitated, but you insisted, Dr. Sannier, and I thank you for believing in me and urging me to go in this direction. This was a fascinating dataset to work with.

Dr. Gary Bitter has been an unending source of encouragement to pursue and publish work relating to "big data" and education in the service of students around the world. You encouraged me to take Dr. Sannier up on his offer. I am glad I did.

To Dr. Rebecca T. Barber: You are one of the smartest educational data scientists I know. Thank you for your guidance, technical knowledge and support. You have been amazing.

To Dr. Ian Douglas: Thank you for your encouragement to pursue qualitative research as well as quantitative research in MOOCs. Planning for the next phase has already begun.

Thank you to Kim Watson, my editor, for all the heavy lifting you performed to polish the final product.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

Technological innovation has been the driving force for increasing the amount of information that can be communicated over distance (Poe, 2011). The invention of the printing press in late medieval Europe allowed for information in a portable format to be transported and distributed across the continent. The sudden availability and distribution of texts such as the Bible and the works of Plato and Aristotle in the 14th through 16th centuries created a political revolution and the destruction of an economic order that had dominated Europe for a thousand years (Eisenstein, 1983). The invention of the telegraph and the telephone in the 19th century brought another wave of change in the delivery of information when the spoken word could be transmitted almost instantly over long distances. When the concept of mass media was introduced in the form of radio and television, the ability to transmit information over long distances went beyond communication between two individuals to instantaneous communication to large populations over long distances. Most recently, in the past two decades, information communication has once again been revolutionized through the advent of the Internet and the World Wide Web (Gross & Harmon, 2016).

Because one of the core elements of education is the communication of information, each of the above technological developments in information transmission has had an impact on education (Thelin, 2009). However, not all have had the same degree of impact. Some information communication technologies, such as the printing press, have had a more direct impact on education than others, such as the telegraph. Yet, the newest innovation in the communication of information over great distances, the invention of the Internet, is having a profound impact on education. How far reaching this

impact will be is just now beginning to be understood (Mason, 2000; Hollands & Tirthalia, 2015).

The first classes offered over the Internet by institutions of higher learning began to be available in the mid-1990s through the Open University in the United Kingdom, and Walden University and the University of Phoenix in the United States (Hollands & Tirthalia, 2015). Today, twenty years later in the U.S., one in seven university students is taking all her classes online (Poulin & Straut, 2016). In recent years in the United States, private not-for-profit universities and public universities have taken the lead in the growth of online education. At one point, for-profit universities dominated online higher education in the U.S.; however recently, for-profit enrollments have declined (9%). On the other hand, the most recent government surveys for the period between 2012 and 2014 reported private not-for-profit university enrollments in online education growing at a robust rate of 33% and a 12% growth in enrollments of online students by public institutions of higher education (Poulin & Straut, 2016).

Statement of the Problem

One of the fastest growing segments of online education in recent years has been Massive Open Online Courses (MOOCs) (Hollands & Tirthalia, 2015). These courses are offered by universities for free or at a very low cost to students anywhere in the world (Belleflamme & Jacqmin, 2014). The "openness" of these online courses means that anyone with access to a computer and an Internet connection can participate. One of the reasons for the massive number of students in these classes *is* their openness to all. It is not unusual for thousands of students to enroll in a course. However, because of the low bar for entry, often at no cost to the student and requiring no academic prerequisites,

attrition in these courses is dramatic. Researchers have found the completion rate in many MOOCs to be less than 10% (Amnueypornsakul, Bhat, & Chinprutthiwong, 2014) and, since the advent of large MOOC platforms such as Coursera and edX in 2012-13, there has been much interest by researchers in the causes for students either persisting or dropping out of MOOC-type courses (Allione & Stein, 2016; Zheng, Han, Rosson, & Carroll, 2016; Xiong et al., 2015; Skrypnyk, De Vries, & Hennis, 2015). Although various approaches to predicting attrition, retention, and completion have been taken with respect to several MOOC courses, these issues have not been addressed in two of the fastest growing segments of the MOOC market—self-paced and credit-bearing MOOCs (Shah, 2016). There are several reasons why prediction may be more critical for these courses and why research into attrition and completion might yield insights over and above that for prediction for session-based non-credit or certificate-type courses:

1. MOOC courses offering college credit have a well-defined goal that involves course completion. Whether or not offering credit contributes to or increases the achievement of this goal, completion itself must be examined.
2. The characteristics of students who pursue credit through MOOCs must be determined.
3. Self-paced courses do not have the cohort structure of many of the original MOOCs. One result of this is less active forums (Shah, 2016). How this affects course completion must also be studied.
4. Success in a self-paced MOOC is highly dependent on students being able to self-regulate their behavior. Traces of self-regulation in student online behavior may

offer valuable insights for strengthening prediction models, which can then offer insights into how students self-regulate when working in a self-paced MOOC.

Research Questions

The three primary research questions of this dissertation are: (1) What demographic characteristics and online behaviors are exhibited by students who complete or do not complete a self-paced mathematics MOOC? (2) How early can we predict that a student will either complete or not complete a self-paced MOOC? (3) Is there evidence in the daily activity logs of a self-paced mathematics MOOCs that can show evidence of self-regulation on the part of users? Students working online produce millions of lines of data every day. By mining the information created by students as they work, researchers can better understand the behavior patterns associated with student persistence in online courses, and when a student is at risk of dropping out.

A Brief History of MOOCs

Tuition-based online offerings by institutions of higher education have been available in the United States for two decades. In recent years, however, the trend of offering higher educational resources for free over the Internet has accelerated (Hollands & Tirthalia, 2015). One of the most ambitious of these free programs emerged in the early 2000s as the Massachusetts Institute of Technology's (MIT's) OpenCourseWare initiative. In 2001, a new type of public copyright license was established, known as the Creative Commons license (Creative Commons, 2016). This license allows individuals and organizations to give permission for the free distribution of works that would normally fall under copyright restrictions. Under these licenses in 2004, MIT uploaded to the Internet videos, handouts, and resources used in its undergraduate and graduate

courses. Following MIT's lead, several other universities have since started their own open-courseware initiatives. Currently, 80 institutions of higher education offer free courseware over the Internet from more than 25,000 courses (Danver, ed., 2016).

Another significant development during this decade was the 2006 launch of Khan Academy. Using a Yahoo Doodle notepad to create short math tutorials that could be uploaded to YouTube, Salman Khan began tutoring his niece who was located several states away. Soon, friends and family were accessing the videos and making their own requests (Pinkus, 2015). These online videos became so popular that Sal Khan quit his job as a hedge-fund manager and formed the non-profit educational company, Khan Academy. Now Khan Academy offers free instruction to 26 million registered students in 190 countries in subjects from physics and computer science to art history and American civics (Scorza, 2016).

While the primary focus of Khan Academy was not the higher education market, its widespread success along with the open-courseware initiatives of major universities such as MIT, Yale, and Carnegie Mellon paved the way for the newest wave in online education: the MOOC.

The Original MOOC

The term MOOC was coined in 2008 by two Canadian professors, Dave Cormier and Bryan Alexander (Parr, 2013). Both Cormier and Alexander had spent a decade experimenting with how educators could use the Internet to enhance instruction. They were subsequently enlisted by George Siemens of the University of Manitoba and Steven Downes, senior research officer of Canada's National Research Council, to help create a new course. These two professors designed a class called *Connectivism and Connective*

Knowledge. Twenty-five fee-paying students from the University of Manitoba enrolled, but the course was also opened to anyone who wanted to join over the Internet. Over 2,000 students enrolled, thereby accessing the course for free (Yuzer & Kurubackak, eds., 2014).

This first MOOC grew out of a theory of learning developed by Siemens and Downes called *connectivism.* Connectivism views learning as analogous to a computer network composed of nodes and links. This theory sees the learning process as taking place when connections develop between individuals and between individuals and non-human "appliances" such as databases. Emphasis is placed on the non-linear development of knowledge and how learning and knowledge develop in an organic fashion when knowledge flow is unimpeded and continuously updated within the network (Siemens, 2005). Two websites that exemplify Siemens' view of how connectivism works in the age of the Internet are the online encyclopedia, Wikipedia, and the question-and-answer programming site, Stack Overflow (D. Bruff, 2016). Both Wikipedia and Stack Overflow rely on cooperative communities to contribute information and to keep the sites up-to-date. Conversations are at the heart of both projects. In Wikipedia, conversations about the content of any one topic take place in the wiki background as knowledge is curated by an expert community for presentation in the visible part of the encyclopedia. In Stack Overflow, conversations about the best answer to a question posed by a programmer occur in the open and members vote for the answers they think are best. MOOCs built around the connectivist theory have come to be known as cMOOCs. Four activities are key to cMOOCs: *aggregation*, the accumulation of learning materials that are continuously updated by participants as the MOOC progresses; *remixing*, the act by

learners of making connections between different learning materials and then sharing those insights with other participants through blogging, social bookmarking, or tweeting; and *repurposing* and *feeding forward*, the processes of creating internal connections between the materials and the insights of others and then afterwards forming new connections (Yeager, Hurley-Dasgupta, & Bliss, 2013). Although cMOOCs were the first MOOCs to appear on the online scene and have been defended as having been designed and grounded in educational theory in order to be authentic and to generate new learning, at present they constitute only a small part of the overall MOOC landscape (Caulfield, 2013).

The Beginning of xMOOCs

In 2011, a different type of MOOC emerged. Unlike the cMOOCs championed by Siemens and Downes, this type of MOOC is more linear, instructor-driven, platform-driven, and similar in many ways to the format of the traditional university classroom, while also adapted in certain respects for the Internet age (Sokolik & Bárrcena, 2015). Although xMOOCs (as these instructor-driven MOOCs have come to be known) seemed to burst onto the scene in 2011, much like the cMOOCs, they actually reflected years of experimentation and incremental development by educators who wished to combine the power of the Internet with the goals of learning (Ng & Widom, 2014).

Stanford University was the pioneer in the field of xMOOCs. Stanford's first foray into the MOOC space grew out of the open-courseware movement in the early part of the 21st century. In 2007, Andrew Ng, a Stanford professor and leading researcher in machine learning, collaborated with the Stanford Center for Professional Development to videotape and post online ten courses complete with lecture notes and self-graded

homework (Ng & Widom, 2014). What was different about these courses was their completeness. Up to this point, open-courseware consisted of educator *resources* posted online. For example, a professor might post her syllabus and outlines of lectures, handouts, and videos of student projects. While this made materials available to other educators, they were not meant to replace the actual courses. When Ng constructed his complete courses in cooperation with the Stanford Center for Professional Development and made them available as open-courseware, he labeled the set of courses the Stanford Engineering Everywhere (SEE) project. Ng wanted students anywhere in the world to be able to access not only course resources, but complete course contents with merely a computer and an available Internet connection. As part of the SEE project, Ng and his colleague, Jennifer Widom, experimented with several innovations. One of these innovations was their version of the Khan Academy-style tablet recordings of instructional videos, which they sought to incorporate into their courses (Ng & Widom, 2014).

While Andrew Ng was working on his "teach the world" project, another colleague at Stanford University, Daphne Koller, was working on a different learning problem: how to incorporate a "flipped classroom" model into her own courses by uploading videos of her lectures to YouTube with the intent of making her class time with students more productive (Severance, 2012). The "flipped classroom" is a pedagogical model in which students view taped lectures at home, so class time can be devoted to exercises, projects, or discussions (EDUCAUSE learning initiative, 2012). Eventually, Ng and Koller joined forces to achieve both goals at once: enhance their face-to-face classes while making the content of their courses available to the world. In 2011,

Stanford launched the first three xMOOCs: *Databases,* taught by Jennifer Widom, *Machine Learning,* taught by Andrew Ng, and *Artificial Intelligence* taught by Sebastian Thrun and Peter Norvig (Ng & Widom, 2014). The response to these three free Stanford courses exceeded everyone's expectations. Over 160,000 students signed up for Thrun and Norvig's *Artificial Intelligence* course alone. All three courses in the initial xMOOC offering had over 100,000 enrollees (Severance, 2012). Professor Thrun's students were located in 190 countries and included soldiers on active duty in Afghanistan and single mothers in the United States (Chafkin, 2013). The next year, Thrun and Michael Sokolsky launched the startup Udacity, which is based on a computer platform that was designed to teach Thrun's AI course, and Ng, Koller, and Widom used the computer platform that was used to teach *Machine Learning* and *Databases* to create the startup Coursera (Ng & Widom, 2014). With an explosion of enrollees at Coursera exceeding 1.7 million in 2012, Ng boasted to the New York Times, "We're growing faster than Facebook!" (Pappano, 2012).

Within a few months of the launch of Coursera, an xMOOC competitor was born on the east coast of the United States at Harvard and MIT. This MOOC platform, dubbed edX, was created to produce xMOOCs based on course content from these universities.

MOOCs are continuing to grow at a rapid pace. In 2016, 11.3% of institutions of higher education offer MOOCs (Allen & Seaman, 2016). Class Central, a website devoted to organizing and disseminating information about MOOCs, reported that more than 1,200 free courses were beginning in the month of November 2016, and 127 of those courses were brand new (Shah, 2016).

The Computer Science Connection

In order to understand the emergence of MOOCs, it is important to note the close connection between MOOCs and computer science. Both the founders of cMOOCs and xMOOCs were deeply interested in computer science and technology and all the founders of the xMOOCs were programmers with an interest in artificial intelligence (Severance, 2012). Although the course content of the original cMOOCs and the xMOOCs centered around computers, there are key differences between these two genres of open online courses. George Siemens and Stephen Downes, the founders of the cMOOC, are particularly interested in the theory and philosophy surrounding human-computer interaction (HCI) and human-to-human connections. Ng, Widom, and Thrun, in contrast, were programmers as well as professors who taught various aspects of computer science at Stanford University. As an elite university adjacent to Silicon Valley, California, the connection between xMOOCs and high technology was evident from the beginning.

This difference in background and objectives—on the one hand, contextualizing computer science in the modern world through theory and philosophy, and on the other hand, the practical application of computer science—has affected the trajectory of cMOOCs and xMOOCs. For example, the original cMOOC, *Connectivism and Connective Knowledge,* has gone through several iterations and has matured into a growing community of connected users who continue to benefit from the interactions created through the course (MoocGuide, 2016; Caulfield, 2013). xMOOCs, on other hand, pioneered the development of the computer code necessary to simultaneously deliver educational content to hundreds of thousands of students and has resulted in some of the largest and most successful MOOC platforms, including Coursera, edX, and

Udacity (Fowler, 2013). The popularity of the first three xMOOCs highlights the fact that there is a hunger worldwide for high-quality instruction in the computer sciences, and that computers are particularly well adapted to delivering instruction regarding computers. Although MOOCs currently cover almost every educational discipline, computer science continues to be a cornerstone and driver of MOOC content (Shah, 2016). MOOCs are also dependent on the computer science community for the development of new computer code to continue the process of innovation in MOOC content delivery (Waks, 2016).

This deep computer science connection, especially with respect to xMOOCs, has become a major source of criticism regarding xMOOCs and of MOOCs in general (Papa, 2014). Critics argue that the xMOOC movement is not driven by educators but by computer programmers, is disconnected from contemporary educational theory, and is rooted in obsolete theories of education and psychology (Papa, 2014; Armellinini & Rodriguez, 2016; Kelly, 2014; Fischer, 2014). Some of this criticism is coming from the proponents of cMOOCs. In a 2013 interview with Chris Parr of the *Times Higher Education* website, Siemens, Downes, and Cormier criticized xMOOCs as "static and passive education," comparable to television or online textbooks, depending on pedagogy that was "several decades behind," and devoid of an understanding of the history of online education (Parr, 2013). Each of these points has become part of an ongoing and active debate about what kind of "education" MOOCs actually deliver. This dissertation focuses on yet another criticism of both xMOOCs and cMOOCs: their high dropout and low completion rates.

The Problem of Low Rates of Completion in MOOCs

While educators were surprised and pleased by the number and variety of students who signed up for the first MOOCs, they were equally appalled at the steep drop-off rates of students over the life of these courses. When Sebastian Thrun launched Udacity, typical completion rates for the xMOOC-type courses were 7–10% (Chafkin, 2013). Even more discouraging was how stubbornly these numbers persisted over time even as Thrun struggled to mitigate attrition through innovations in curriculum and delivery as he developed new courses and revamped the original one. Thrun's disillusionment regarding the low course completion rate was one of the primary drivers that caused him to abandon higher education as the content of Udacity's platform and instead tweak Udacity's focus to skill-based professional development courses contracted through large companies such as AT&T and Google (Chafkin, 2013). Sebastian Thrun was not the only one to note the steep attrition rates in MOOCs. In the first years following the launches of Coursera and edX, low course completion rates became a major point of criticism of MOOCs and led some to proclaim that MOOCs were a grand failure (Konnikova, 2014).

In the intervening four years since the establishment of the major xMOOC platforms, attrition and completion rates for MOOCs have become the focus of intense research (Ferguson, Coughlan, & Herodotou, 2016; Breslow, 2016; Veletsianos & Shepherdson, 2016). There have been several drivers of this research. The first two and most obvious were the twin goals of MOOC proponents to raise completion rates and to shed light on why students drop out of MOOCs. But this research has also made a valuable contribution in having opened the discussion regarding the kind of completion rates to be expected in MOOCs. Many researchers believe this debate over expectations

to have been very fruitful because it has begun to shape a dialogue on how MOOCs fit into the higher education landscape (Clark, 2016; Ho et al., 2014; Koller, Ng, Do & Chen, 2013). Some examples of the kinds of questions that are being asked in the debate over MOOC attrition include:

1. In what ways are MOOCs similar to and different from their corresponding face-to-face courses taking place in the classroom at colleges and universities?
2. Is learning taking place in MOOCs; and if it is taking place, what kind of learning is it, and how can it be measured?
3. Is a MOOC more like a book in a library that students check out to obtain specific information and then return to the shelf, or are MOOCs more like college classes from which benefits arise from interactions with experts in a particular field, discussions with fellow students, and the credentialing that comes with formal education?

Due to the diversity of MOOC offerings and MOOC learners, none of these questions have simple answers.

New Trends

Growth of Self-Paced MOOCs

In November 2016, Dhawal Shah of Class Central reported on the recent developments in MOOC trends of 2016 in an article titled, "MOOC Trends of 2016: MOOCs No Longer Massive" (2016a). In the article, Shah pointed out that while MOOC participation continues to grow, recently passing the 35 million mark, individual course sizes are down. One of the major drivers of both the growth in participant numbers in MOOCs and the smaller class sizes is the increased number of self-paced courses and the

number of courses with rolling enrollment offered on a monthly or bi-weekly basis. The trend toward rolling enrollment has more than doubled the number of courses being offered in any given month (Figure 1).

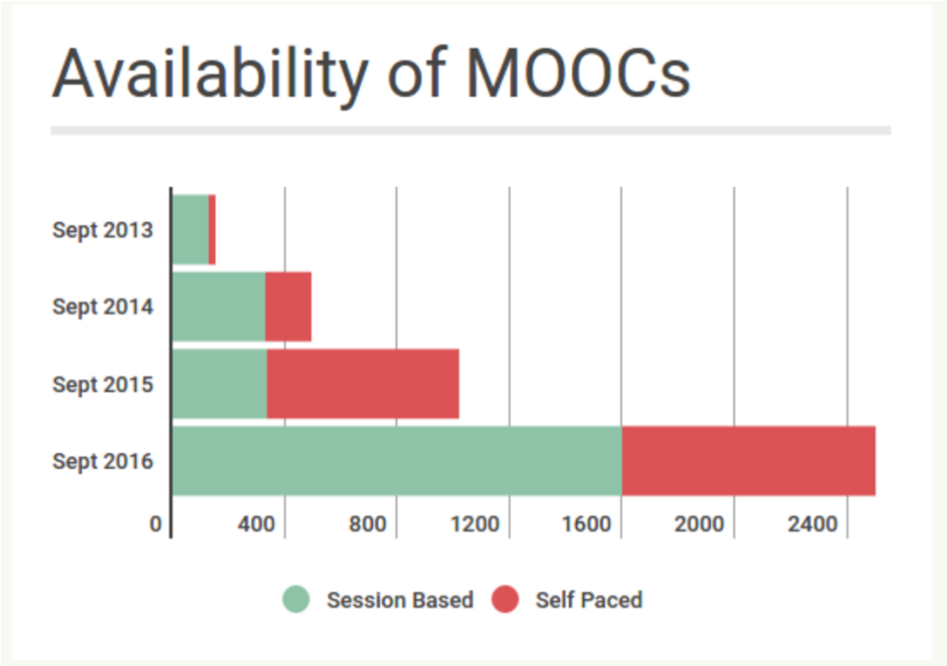


Figure 1. Growth of self-paced courses from Sept 2013 to Sept 2016 (Shah, 2016b).

MOOCs for College Credit

Another trend emerging in 2016 is the offering of MOOCs for college credit from accredited universities or accreditation organizations (Shah, 2016a). Six universities from five countries collaborated with the MOOC platform edX to produce MOOCs with transferable college credit. One of the leaders in this trend has been Arizona State University. In April 2015, Arizona State University and edX announced the launch of the Global Freshman Academy (Lewin, 2015). The Academy offers a full year of university courses that can be taken for transferable college credit. One of these courses offered for transferable credit, college algebra, is also a self-paced course.

CHAPTER 2

Literature Review

In the three years since The New York Times declared 2012 the "Year of the MOOC," the issue of MOOC completion and attrition rates has become a very active research area (Gašević, Kovanović, Joksimović, & Siemens, 2014; Ferguson, et al., 2015; Breslow, 2016; Veletsianos & Shepherdson, 2016; Ferguson, Coughlan, & Herodotou, 2016). Overall, MOOC completion rates are very low. An average of 5–10% of students who enroll in MOOCs will go on to complete them (Veletsianos & Shepherdson, 2016; Skrypnyk, De Vries, & Hennis, 2015; Glance & Barrett, 2014; Koller, Ng, Do, & Chen, 2013). In addition, these high attrition and low completion rates have been stubbornly resistant to innovations in MOOC design and interventions directed at students (Bacon et al., 2015; DeBoer, Ho, Stump, & Breslow, 2014; Glance & Barrett, 2014).

Because MOOCs are similar in many ways to conventional university courses, i.e., containing lectures, quizzes, exams, and homework assignments, the stark difference in dropout rates is disturbing. Beyond the simplistic argument that there is a low bar for entry (anyone can sign up and, often, at no cost), researchers have tried to determine who the students in MOOCs are and what do they seek to gain from their participation (Goldwasser, Mankoff, Manturuk, Schmid, & Whitfield, 2016). One way researchers have sought to explain the difference in dropout rates is in terms of MOOC participant intent. Many participants in MOOCs, these researchers argue, never intended to earn a certificate or complete the course in the first place ("Massive study on MOOCs," 2015). Instead, the MOOC participants can be viewed as belonging to different classes of students with differing motivations. For instance, a MOOC may contain passive

participants, active participants, and community contributors (Koller, Ng, Do, & Chen, 2013). Other researchers have argued that course completion is the wrong measure of MOOC success in the first place. Participants may achieve their goals with respect to their participation in MOOCs and these goals may not involve completion (Clark, 2016; Ho et al., 2014). In some ways, the more interesting question has become—why do students stay in a course rather than why do they drop out.

Psychological Factors Affecting Dropout

In order to investigate this question regarding the intent of participants in MOOCs, several studies have examined the internal factors of participants that might determine whether a student will complete a course or drop out early. These internal factors include motivation, affect, goal-striving, grit, self-efficacy, and sentiment, among others (Breslow, 2016; Ferguson, Coughlan & Herodotou, 2016; Khalil, 2014). One of the major approaches for investigating these intrinsic factors has been the use of surveys filled out by participants before, during, or after the course is completed (Cupitt & Golshan, 2014; Green, Oswald, & Pomerzantz, 2015; Oakley, Poole & Nestor, 2016). Among the internal factors most studied is the motivation of participants in taking the course (Ferguson, Coughlan, & Herodotaou, 2016; Breslow, 2016). Adamopoulos found that student attitudes toward the professor of a course had a major effect on the motivation of students to remain in the course. The author also identified course difficulty and length as major motivating factors (2013). Xiong et al. examined intrinsic motivation versus extrinsic motivation in MOOCs and found both to be important for student retention, whereas social motivation was not as important (2015). Other researchers have found that the intent expressed by participants at the beginning of their courses was a

good indicator of whether they were motivated to complete the course (Pursel, Zhang, Jablokow, Choi, & Velegol, 2016).

Other internal factors related to motivation that have been connected to the successful completion of MOOCs include goal-striving, sentiment, and affect. Kizilcec & Halawa found goal-striving to be a characteristic that was more commonly found in successful completers of MOOCs than those who did not complete (2015). On the other hand, negative grit scores, especially for at-risk students, were found to be correlated with poor performance in this medium (Cupitt & Golshan, 2014). In addition to goal-striving, how a student feels about participating in MOOCs has also been explored as a possible clue to whether MOOC users will persist or drop out. In some studies, differences in affect were not shown to be statistically significant in predicting who would persist in a MOOC (Heutte, Kaplan, Fenouillet, Caron, & Rosselle, 2014). However, sentiment, measured on the basis of forum comments, was found to be predictive of student dropout in some studies (Chaplot, Rhim, & Kim, 2015; Tucker & Divinsky, 2014). However, in other studies, sentiment analysis was not found to be predictive of who would persist and who would not (Dmoshinskaia, 2016).

Demographic Factors Affecting Dropout

In addition to internal psychological factors, demographic variables have been measured with respect to determining their impact on MOOC completion and attrition rates. Since participants in MOOCs can be located anywhere in the world, several researchers have been interested in how physical location may affect persistence in these courses. In a MOOC experiment by the French Ministry of Higher Education, researchers looked at differences between European and African participants (Heutte, Kaplan,

Fenouillet, Caron, & Rosselle, 2014). While African participants displayed a higher degree of enthusiasm at the beginning of the course, they were less likely to persist and complete it. Ho et al. reported on the first year of Harvard and MIT MOOCs and found that although the United States had the largest number of enrollees (almost a third of that of the top 25 countries), Spain, Greece, and the Czech Republic were the countries with the largest percentage of registered users who went on to receive certification (2014). Hone & El Said reported a 32% completion rate among students from the University of Cairo who were encouraged to take a MOOC for their own academic development—a rate of completion well above the overall average completion rates of 5–10% for MOOCs (2016).

Other demographic factors that have been examined in relation to MOOC completion, retention, and attrition rates have been gender, language, and socio-economic and educational background. Dillahunt, Wang, & Teasley compared how students from developed countries fared in MOOCs versus those from developing regions of the world (2014). While a higher percentage of participants from developed regions earned certificates and completed the courses, participants from developing regions were more likely to earn a certificate with distinction. Another demographic factor that has been studied is gender. MOOCs, especially STEM MOOCs, have been dominated by men (Ho et al., 2014). Jiang, Schenke, Eccles, Xu, and Warschauer found that not only is participation in STEM MOOCs dominated by males, completers of STEM MOOCs are disproportionately male. However, this statistic varies widely by country and culture. For example, in Indonesia, a STEM MOOC completer is, on average, slightly more likely to be female, whereas in Japan, almost none of the completers are female (2016).

Barriers to Completion of MOOCs

Another approach researchers have taken to consider why participants in MOOCs do not persist is by looking at barriers to completion. In 2013, Belanger & Thorton surveyed participants in Duke University's first MOOC, called *Bioelectricity: A Quantitative Approach*, to determine why they had not finished the course. The most common answers were: lack of time, insufficient math background, and an inability to transfer their learning from the conceptual to application. The time factor tends to be one of the reasons cited most often by students who drop out of MOOCs (Xiong et al., 2015; Khalil, 2014; Thille et al., 2014). This may have to do with expectations. To master the material from rigorous courses like *Bioelectricity* takes time. Even a motivated student may find it very difficult to meet the demands of a rigorous course in addition to fulfilling other life responsibilities. In their study, *Learner's Strategies in MOOCs*, Veletsianos, Reich, & Pasquini report hearing repeatedly from students the necessity of "stealing time" from other critical life activities in order to complete a MOOC that was a priority for them (2016).

Prediction of MOOC Completion

In addition to identifying reasons why participants may or may not complete a MOOC, research into MOOC completion has focused on prediction. Predictable patterns of persistence or attrition offer insight into the drivers of MOOC completion as well as providing touchpoints for actionable interventions that can increase completion rates. The data used for the prediction of MOOC persistence has come from five sources: demographic data on participants, participant surveys, clickstream data of various types, participation in forums and social media, and work done within the courses themselves

(Goldwasser, Mankoff, Manturuk, Schmid, & Whitfield, 2016; Sharkey & Sanders, 2014; Zheng, Han, Rosson, & Carroll, 2016; Kizilcec, Piech, & Schneider, 2013).

Video viewing. Several types of clickstream data have been used as variables to develop predictive models. One of the most common variables in MOOC predictive models is clicks related to video viewing (Breslow, 2016). Actions related to a video, such as hitting the pause button, can be stored as a server-side event in a file written in JavaScript Object Notation (JSON) (Balakrishnan, 2013). Methods for measuring video watching to predict persistence vary widely from measuring video watching as a simple binary (watched or did not watch the video) (Stein & Allione, 2014) to capturing elaborate video watching patterns such as re-watching, skipping, fast watching, and slow watching (Sinha, Jermann, Li, & Dillenbourg, 2014). Greater amounts of video watching in MOOCs have been conclusively shown to be predictive of course completion and has even been linked with greater satisfaction with the course (Kizilcec, Piech, & Schneider, 2013). One interesting finding showed that it is not necessary for participants to have watched videos from beginning to end to demonstrate this predictive effect, especially near the end of the course (Balakrishan, 2013).

Quiz attempts. Another clickstream variable found to be predictive of course completion in MOOCs is quiz attempts. Like videos, quiz data has been measured in different ways. Studies have looked at quiz attempts on both practice and graded quizzes (De Barba, Kennedy, & Ainley, 2016) and purely at graded quiz attempts (Amnueypornsakul, Bhat, & Chinprutthiwong, 2014). Another study looked at quiz attempts in conjunction with other behaviors such as referring to other materials (Sharkey & Sanders, 2014). Stein & Allione (2014) found that participants in MOOCs who took

the first quiz were 30% less likely to drop out of the course, and De Barba, Kennedy, & Ainley found the number of quiz attempts to be more predictive of completion than video hits (2016).

Forum activity. A third variable that has proven very predictive of MOOC persistence is forum activity. Forum activity can be looked at as a predictive variable in many ways. For example, it can be deduced purely from clickstream data as forum page views (Kloft, Stiehler, Zheng, & Pinkwart, 2014). A more in-depth approach to forum posts is counting how many posts a student contributes and then analyzing the text through natural language processing to identify characteristics such as sentiment (Chaplot, Rhim, & Kim, 2015). Other forum behavior such as up-voting, down-voting, or starting new threads can also be measured (Sinha, Li, Jermann, & Dillenbourg, 2014). All these measures have been found to be predictive of MOOC persistence and completion. In addition, Jiang, Williams, Schenke, Warschauer, & O'Dowd found that forum behavior could be predictive of those who would receive a certificate with distinction.

Other clickstream traces. Several other behaviors that can be detected from clickstream behavior have been linked by researchers to MOOC persistence and completion. Some of the major ones include the number of active days a student spends in a course (Lim, 2016; Kloft, Stiehler, Zheng, & Pinkwart, 2014; Laurillard, 2014; DeBoer, Ho, Stump, & Breslow, 2014). Another is the student's pace in moving through the material (Thille et al., 2014) or the number and length of breaks or stop-outs a student takes from a course (Halawa, Greene, & Mitchell, 2014). Although these measures taken together do not create perfect models, they have the potential to give early indications as to who will be a persistent completer and who may drop out of the course along the way.

Prediction Methods

Because of the amount of data associated with online courses (often millions of lines of computer code), machine learning techniques are favored among researchers for building predictive models for MOOC completion and attrition. These analysis methods include methods of classification (Kizilcec, Piech, & Schneider, 2013; Sinha, Jermann, Li, & Dillenbourg, 2014) and clustering (Kizilcec, Piech, & Schneider, 2013). Other machine learning algorithms used are Hidden Markov Models (HMM) (Breslow, 2016; Kizilcec & Halawa, 2015) and neural networks (Chaplot, Rhim, & Kim, 2015). More conventional educational research techniques such as Structural Equation Modeling (SEM) (Aparicio, Bacao, & Oliveira, 2016; Xiong et al., 2015) and logistic regression (Semenova, 2016; Thille et al., 2014) have been used to build predictive models for MOOCs. Other techniques such as survival analysis have also been used (Allione & Stein, 2016; DeBoer, Ho, Stump, & Breslow, 2014; Greene, Oswald, & Pomerantz, 2015).

Many excellent predictive models have been created in relation to courses in online education (for example see Barber & Sharkey, 2012), and even though MOOCs are a recent innovation in online education, predicting completion in MOOCs has been an active area of research. Predictive models based on case studies of individual MOOCs are common, as are studies that aggregate data from multiple MOOCs. The primary sources of data for this research consist of demographic information from students, clickstream data generated as students work online and navigate through a course, and survey data collected from willing participants before, during, and after a course is completed. Major sources of clickstream data come from video views, completion of quizzes and

homework, and forum activity. The use of machine learning techniques dominate in the creation of predictive models due to the volume of data produced by students in MOOCs. Attempts to find effective predictors of MOOC completion have generally been successful although there is plenty of room for improvement. Almost all of the research on MOOC completion is based on session-type MOOCs with single start and completion dates. There is almost no completion-prediction research associated with self-paced MOOCs or rolling-enrollment MOOCs, even though this is currently the fastest growing segment of the MOOC market (Shah, 2016). In addition, there is almost no prediction research associated with credit-bearing MOOCs. This omission is less surprising since, at this point, credit-bearing MOOCs make up a very small part of the total number of MOOCs offered.

CHAPTER 3

Methodology

The source of the data for this study is a self-paced MOOC from the Global Freshman Academy (GFA). The GFA is offered by Arizona State University (ASU) through the MOOC edX. The Global Freshman Academy offers several online courses that are typically taken by freshman at ASU. These courses are offered in a "try before you buy" format where students can enroll and take the course for free (or a small fee for identity verification through edX) and then if students successfully complete the course they can decide if they want to purchase college credit. Courses range from first year English composition to astronomy. More information on the Global Freshman Academy can be found at https://gfa.asu.edu/.

Knowledge Space Theory

The course studied here is the college algebra course. This mathematics MOOC uses the Assessment and LEarning in Knowledge Spaces (ALEKS) intelligent tutoring system (ITS) to present the mathematical content of the college algebra course to students. ALEKS uses an artificial intelligence (AI) engine that was developed based on Knowledge Space Theory. This is a mathematical cognitive theory developed by mathematical psychologist Jean-Claude Falmagne (McGraw Hill Education, 2016). The central concept of Knowledge Space Theory is the "knowledge state," which is defined as a set containing all the problems (in this case, college algebra mathematics problems) that an individual is capable of solving (Falmagne, Koppen, Villano, Doignon, & Johannesen, 1990). A collection of knowledge states constitutes a "knowledge structure" and certain classes of knowledge structures are called "knowledge spaces." The purpose of

knowledge spaces is to map commonalities between math components that need to be mastered and to use these commonalities to create accurate assessments of a student's math knowledge and to design custom learning pathways for each student based on that student's current math knowledge. Each pathway is designed to result in the student's successful mastery of all the components in the domain (McGraw Hill Education, 2016).

Theoretical Framework

Whereas the Knowledge Space Theory was the theoretical framework for the creation of the ALEKS mathematics program, the theoretical framework for this research is self-regulated learning (SRL) theory. SRL, like knowledge space theory is concerned with the learning process within the individual student. SRL explores how students use consciousness of cognition, behavior, and motivation to control these aspects within themselves in order to actively pursue an academic task (Alexander, Dinsmore, Parkinson & Winters). When learners are self-regulated, they engage in volitional behaviors that further their learning goals. These behaviors can include goal setting, developing task strategies, and help seeking (Barnard-Brak, Lan, & Osland Paton, 2010). SRL is therefore seen as a lens through which achievement differences in students can be explained and predicted (Marzouk et al., 2016). Bussey (2011) observed that self-regulation takes on greater importance as technology gains greater prominence in education. Because of technology, students are becoming more responsible for the pace, timing, and place of learning as learning takes place more and more frequently outside the classroom, especially due to the availability of learning over the Internet (Douglas & Alemanne, 2007).

The performance and persistence of students in a self-paced mathematics course, such as ALEKS presented in a MOOC format, relies heavily on student self-regulation to be successful. Although this self-regulation is not directly observable through activity log data, the results of self-regulation can be viewed in behaviors such as completing the initial knowledge check (the pretest in ALEKS), the amount of time students work in the intelligent tutoring system, and how steadily they progress through the material. These *traces of self-regulation* constitute the building blocks of the models used in this study.

The Data

The data for this study were gathered from college algebra students in a self-paced MOOC offered through a large public university during an eight-month period from May 1 to December 31, 2016. This MOOC is marketed through the edX platform and is one of the few MOOCs currently available that is offered for college credit. There are two sets of data associated with each student. The first dataset, provided by edX, contains demographic data featuring the following variables:

- Student ID
- Gender
- Birth year
- Country location
- Education level

Each of these variables was tested as part of the model to gauge their respective predictive values regarding student persistence, completion, and attrition.

The second dataset was drawn from ALEKS in the form of JSON files converted into flat files as Excel spreadsheets. In these files, a line of data reflects a single day in which a

student has worked in the course along with several data points collected or calculated from the daily activity logs, including:

- How much time the student spent working in the course
- How close the student is to mastering all 419 mathematical skills
- The date of last login
- The date of last assessment
- How many skills have been mastered
- How many skills have been learned

This raw data generated by students working online is longitudinal and fairly granular. The unit of time is a twenty-four-hour period, so it does not reflect a moment-by-moment record of student activity. For example, if a student works in ALEKS more than once in a single day, hours in ALEKS will be aggregated for that day and will not be recorded as separate sessions. Each new skill learned by a student in a twenty-four-hour period is recorded and a comparison can reveal if a topic has moved from learned to mastered during that day.

Splitting Data to Make Predictions

In the field of machine learning, the analysis undertaken here would be regarded as a "classification problem" (Witten & Frank, 2000). In this case, college algebra students were in the sample were separated into two classes: "completers" and "non-completers." From there, the goal was to predict in the early stages of the course which students will fall into each class. To develop models that can correctly predict these two outcomes, a dataset for which the ultimate outcome is known must be used. The retrospective dataset for the self-paced college algebra fulfills this requirement because

for each of these students who began on May 1, 2016, and worked in the course until December 31, 2016, who completed the course and who did not is known.

Here, to prevent "overfitting" the model, a technique commonly employed in machine learning known as "splitting the dataset" was employed (Witten & Frank, 2000). To split the data, an algorithm is used to divide the dataset into two evenly matched groups with respect to the distribution of the outcome characteristic of completing or not completing. With this technique, a subset sample is created from the sample for the purpose of making the model more highly generalizable. If predictions are customized too closely to the current dataset, there is a chance that when fresh data is applied to the model—such as new students in the college algebra course—the model will not predict as well as it did using the original dataset for which it was designed. This is due to quirks that are unique to the original dataset and is called "overfitting" the data.

Splitting the dataset into training and testing sets mitigates overfitting by mimicking the process of introducing the model to fresh data and observing how it performs. The "training set" is used as the basis on which to build the model and contains a larger portion of the data (usually between 70–80%). Once the model has been completed using the training set, it is used on the "testing set" containing the remainder of the data to determine whether the model predicts as well as it did on the training set. If its performance is similar, there is a high probability that the model will predict similarly when introduced to new data for which the outcome is unknown. In this study, for the logistic regression dataset, I introduced a 70/30 data split to produce the training and testing sets.

Techniques

To create and analyze variables, linear and logistic regression were used as primary techniques after data cleaning and performed all data cleaning and data analysis in R version 3.3.1 (2016-06-21) within the open source integrated development environment (IDE) RStudio version 0.99.902. Extensive data cleaning was required to remove from the dataset students who never completed the initial knowledge check and to remove lines of data that were redundant or superfluous. For example, a line of data is created for each day a student is in the course, but new lines continue to be created every day until the chosen cut-off date, whether or not the student is active in the course on any of those subsequent days. Determining when students cease to be active in the course and removing the associated extra lines of data are essential steps in preparing data for analysis. A heuristic of thirty days of inactivity was used to determine that a student had become "inactive" in this course. The last active day before a student becomes inactive was considered the last day in the course. Repeated lines of data after this last active day were the lines that were removed. In addition, all the skills students master are stored in lengthy character vectors. These vectors must be converted to binaries (does this student have this skill or not) and numerical values (how many skills has this student mastered on this day of the course).

After the data is cleaned, the variables must be summarized and calculated to be added to the predictive models. Table 1 lists the variables used in the models:

Table 1
Independent Variables, data types, and values

Variable	Type	Possible Values
Age range	Categorical	18-22; 23-29; 30-39; 40-49; 50-59; 60-69; Over 70; Under 18; Unknown
Gender	Categorical	Male, Female, Other, Unknown
Country/Region	Categorical	One of 142 countries or aggregated as one of 20 world regions defined by UN DESA, Unknown
Education	Categorical	Associate, Bachelor, Doctorate, Elementary, High School, Junior High School, Master, None, Other, Unknown
Verified	Dichotomous	Verified, Audit, Unknown
Placement Test Score	Continuous	0 - 419
Total days in course	Continuous	1 - 240
Total active days	Continuous	1 - 240
Ratio of active to total days in course	Continuous	0 - 1
Hours in course	Continuous	1 – 5,760
Mean topics per hour mastered	Continuous	0 - 419
Mean topics per day mastered	Continuous	0 - 419
Mean topics per week mastered	Continuous	0 - 419
Time in initial knowledge check	Continuous	0 - 50
Days in initial knowledge check	Continuous	0 - 50
Greatest number of active days in a row	Continuous	1 - 240
Number of breaks from the course	Continuous	1 - 120
Greatest number of non-active days	Continuous	1 - 240

- **Age Range**. Age range is a variable of interest because self-regulation may be affected by age due to exposure to education or work experience and may therefore be predictive of course completion.
- **Gender and country**. Researchers have found that self-regulation can vary in online learning by gender and culture, especially in STEM subjects (Bussey, 2011; McInerney & Schunk, 2011).
- **Education.** Because school is a primary environment in which self-regulation is learned, level of education is expected to be predictive of completion or attrition in the MOOC.
- **Verified.** Choosing edX verification allows students to opt to convert their completion in the course to university credit; therefore, choosing verification constitutes evidence of goal setting—a key component of self-regulated learning—and is expected to be predictive of completion.
- **Placement.** Placement within the course pretest is a key indicator of background knowledge.
- **Active days and ratio of active days to total days**. Active days are operationalized as any day a student spends working in the course whether or not they make any progress. Days working on the pretest before it is completed are considered to be time spent doing a pretest, and are not counted as active days. Number of active days has been shown to be predictive in other MOOC models and may be evidence of self-regulation (Lim, 2016; Kloft, Stiehler, Zheng, & Pinkwart, 2014; Laurillard, 2014; DeBoer, Ho, Stump, & Breslow, 2014). Total

days is operationalized as all active or not active days between and including the first and last active days in the course.

- **Hours in course**. Amount of time is measured as whole hours and tenths of hours. Amount of time is operationalized by a student logging into ALEKS through edX and working in the course. Time spent working in the course is regarded as evidence of self-regulation and is expected to be predictive of completion.
- **Pace**. Pace is measured as the number of topics learned or mastered per hour worked in the course. Average pace as well as acceleration or deceleration of pace are considered as predictive variables of completion and drop-out. Pace has been used by other researchers to predict completion in MOOCs (Thille et al., 2014). In a self-paced MOOC like this one, pace may be even more predictive.
- **Breaks between active days**. The number and length of breaks between active days is also considered as a predictive variable. Other MOOC researchers have considered these breaks in their predictive models of MOOC completion (Halawa, Greene, & Mitchell, 2014).
- **Number of consecutive days working in the course**. The number of consecutive days a student works within the course has not been one of the major variables considered by other MOOC researchers, but in a self-paced MOOC, this variable could be predictive and is included here.

Using linear regression models, the above variables were tested to predict the following:

1. The total number of math skills a student will learn in the course.
2. Whether or not this student will complete the course.

Because attrition is so high in MOOCs, based on earlier research results, it can expected that signals related to dropout or completion to be weak at the beginning of the course and grow stronger as data accumulates (Kloft, Stiehler, Zheng, & Pinkwart, 2014).

Logistic Regression

Because the outcome variables are dichotomous, logistic regression can be used as a technique for predicting these binary-type variables. For example, "Will this student show up next week and work on the algebra course?" is a yes or no question. Logistic regression creates probabilities for the answers yes or no. In order to have data to both create and then test the model, the data was split into training and testing sets. The packages in R can do this automatically, e.g., the "caTools" package. To create a training data set, 70% of the data was used because there was plenty of student data and it was possible to make the testing data set larger to increase the level of confidence in the model. Next, the variables were tested for multicollinearity using the "cor" function in R. In cases of variables with high correlations, the variable with the highest correlation with the outcome variable was retained and the others were discarded.

Next, a logistic model in R was created using the remaining variables. The logistic function can be written as follows: $f(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$

where β_0 = the intercept

$\beta_n x_n$= the regression coefficient multiplied by the value of the predictor.

Because of the large number of dropouts in MOOCs, it is important to minimize false negatives for students staying in the course (Chaplot, Rhim, & Kim, 2015). To choose an effective threshold and to minimize false negatives in our model, a confusion matrix was

created to compare actual with predicted outcomes and to examine the sensitivity and specificity of the model. To assist in adjusting the threshold, an ROC curve was graphed in R to determine an ideal threshold value that minimizes both false positives and false negatives. The area under the ROC curve (AUC) was also examined to evaluate the model quality.

Limitations of the Dataset

This dataset had several limitations. First, daily activity logs were the main evidence in this study of self-regulated learning. Although consistent work in a course and persistence may seem to be evidence of self-regulation, actual contact with students through surveys or other instruments is essential to confirm that self-regulation is taking place and to show that there is a firm connection between self-regulation and the behaviors used as predictors in the models. Secondly, although a large number of students were evaluated in this case study ($N \cong 4{,}600$), the study addressed only a single self-paced MOOC course offered by a single university in a single-domain, mathematics.

CHAPTER 4

Results

This chapter is divided into six major sections, in which:

1. Descriptive statistics of the sample are detailed.
2. Differences between completers and non-completers in the sample are described.
3. Variables that served as traces of self-regulation in the daily activity log data.
4. The results of the linear regression models are reviewed.
5. There is a review of the results of the logistic regression models that predict completion and non-completion based on data from the entire course.
6. There is a review of the results of the logistic regression models that use only data from the first active day students were enrolled in the course both with and without the contribution of demographic variables.

Description of the Sample

In the sample, the work of students who participated in a self-paced, open-enrollment college algebra MOOC between May 1, 2016 and December 31, 2017 was examined. Anyone who began the MOOC before May 1st but showed activity extending into the examined time period was excluded as well as anyone who began during the time period but continued to work past December 31st. Like most MOOCs, this course experienced severe attrition (Figures 1–2). Because of the structure of the course, several distinct periods in which students tend to drop out were examined (see Figure 3). In his paper titled, "MOOCs and the funnel of participation," Douglas Clow likened the attrition in MOOCs to a marketing sales funnel (2013).

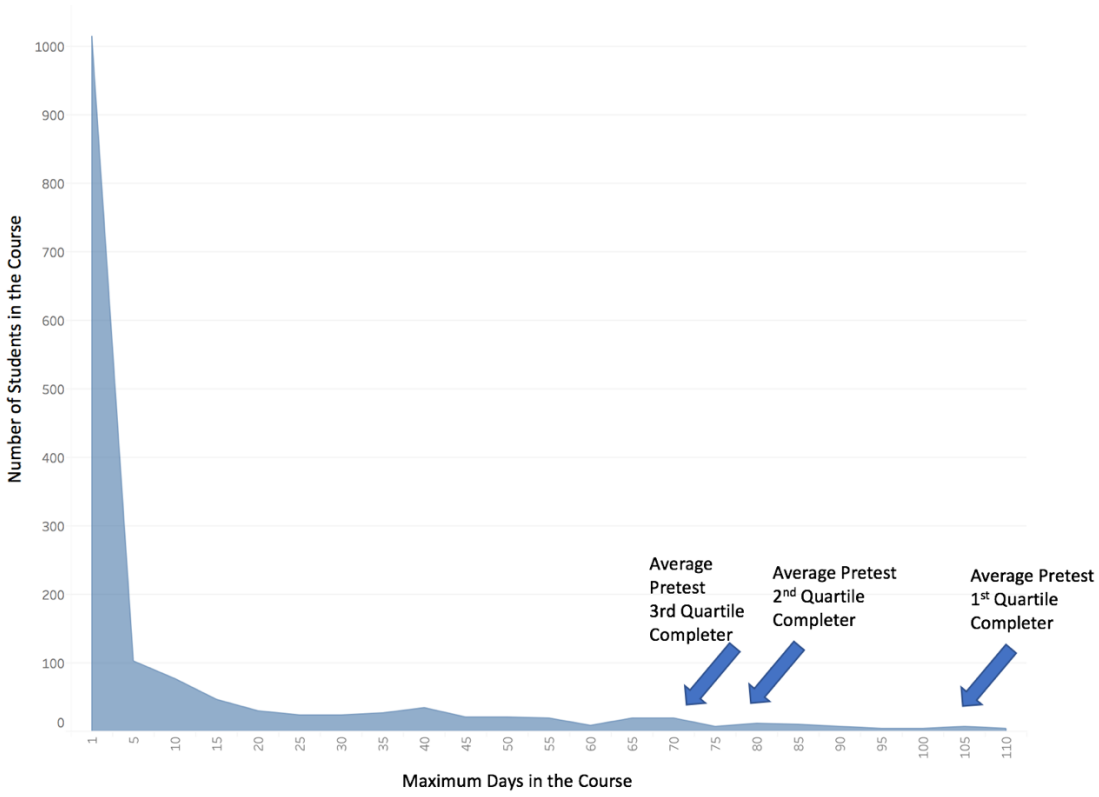


Figure 2. Visualization of attrition in the college algebra course for students who have completed the pretest, based on the total number of active and inactive days in the course. As shown in the graph, approximately 90% of the students were no longer active by their 5th day. Average completers of the course persisted for between 70 and 105 days.

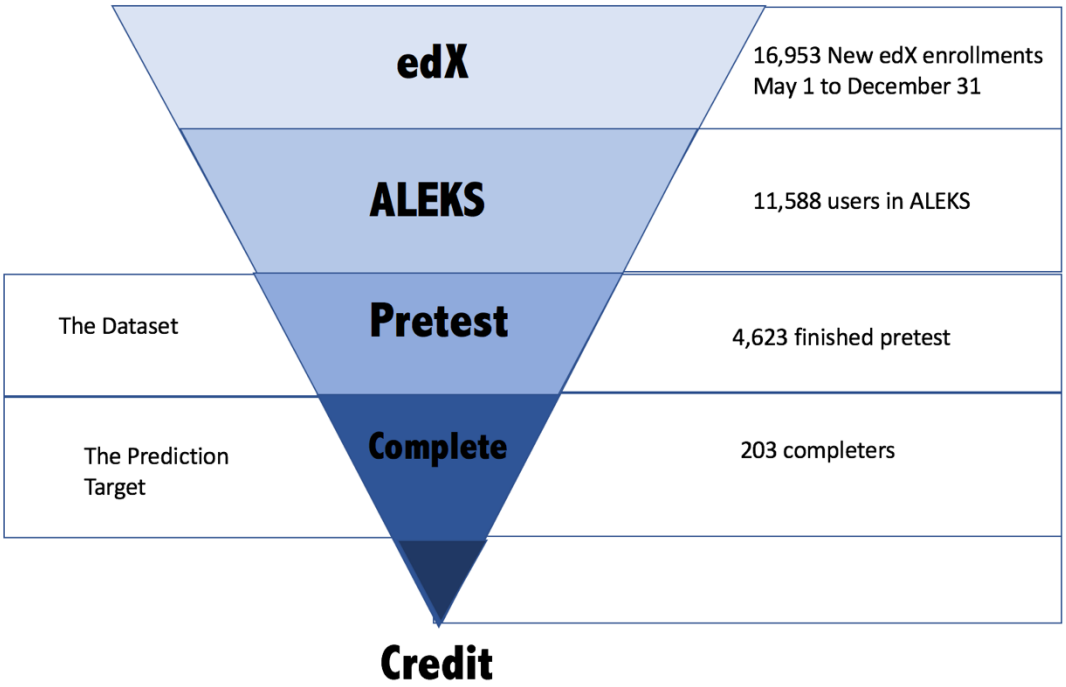


Figure 3. Funnel of participation in the open-enrollment self-paced college algebra course examined in this study. The dataset used in this study included anyone who completed the pretest between May 1, 2016 and December 31, 2016.

The sample used in this study was restricted not only by the designated time period, but also to only those students who had completed the pretest, which was a prerequisite in ALEKS for proceeding with the course. The pretest is used to assess what the student already knows and is used by the intelligent tutoring system to identify skills that the student should work on during the fall semester within that student's zone of proximal development (ZPD) (Vygotskiĭ, 1978).

Sample Demographics

Descriptive statistics for the sample studied are presented in Table 2.

Table 2
Demographics (All data N = 4623, Complete-cases with no missing demographic data N = 3264)

Characteristic	n	% Total Cases	% Complete Cases
Gender			
Male	2515	54	61
Female	1487	32	38
Other	48	1	1
Missing	573	12	NA
Age			
0 – 18	521	11	14
19 – 25	1221	26	33
26 – 35	1249	27	31
36 – 50	676	15	15
51 – 65	236	5	5
Over 65	78	2	1
Missing	642	14	NA
Country			
United States	2043	44	53
India	208	4	6
Canada	132	3	3
Great Britain	123	3	3
Australia	70	2	2
Other	1248	27	33
Missing	799	17	NA
Education			
None	12	.3	.3
Elementary	45	1	1
Junior high school	333	7	9
High school	1603	35	43
Associate degree	335	7	9
Bachelor's degree	908	20	21
Master's degree	522	11	12
Other	189	4	4
Missing	676	15	NA
ID Verified			
Audit	4360	94	95
Verified	188	4	5
Missing	75	2	NA

Demographic data of students enrolled in this course was obtained from a short voluntary survey that is presented to users upon enrolling in a course through edX. The category "ID Verified" is an option available to students to have their identity verified through edX, which is required if they intend to take the final exam and receive college credit from any university offering a course through edX. Over half of the respondents were male (54%), with the percentage of female respondents notably smaller (32%). As with all the survey categories, a substantial portion of respondents chose not to answer the question regarding their gender (12%). The mean age reported by course participants was 30 years old and there were 9 respondents who reported their age as being younger than 7 years. As this is unlikely for a college algebra course, the minimum age more likely ranged between 9–11 years—the next lowest reported age group. While the age data of those who reported to be less than 7 years old was viewed as probably erroneous when calculating the minimum age of the sample, the records were retained because the negative impact of this error was determined to be minimal. The maximum age reported was 89 years.

Participants in this course were from 142 countries in 20 world regions, with 47% coming from North America. The region with the next highest percentage of students was South Asia with 281 participants, or 6% of the sample studied. When participants were asked for their "highest level of education completed," the most frequent answer was high school (35%). However, 38% reported holding some sort of college degree (Associates, Bachelor's and Master's combined). Only a small number of students (4%) requested identity verification, which is an important indication of the intention to complete the course.

Missing Data

All the demographic variables have a substantial amount of missing data, ranging from 12% for gender to 17% for country. None of the data containing evidence of self-regulated learning—the activity log data—was missing. The nature of the missing data was explored both visually and through logistic regression modeling to determine whether the missing data should be characterized as: (1) *missingness completely at random,* (2) *missingness at random,* (3) *missingness that depends on unobserved predictors,* or (4) *missingness that depends on the missing value itself* (Gelman & Hill, 2016). Figure 4 shows a visual model of the distribution for all the sample missing data. A visualization like this one is helpful in looking for patterns of missing data. For example, it is obvious from the visualization that country of origin is not correlated with missingness in education, age or gender.

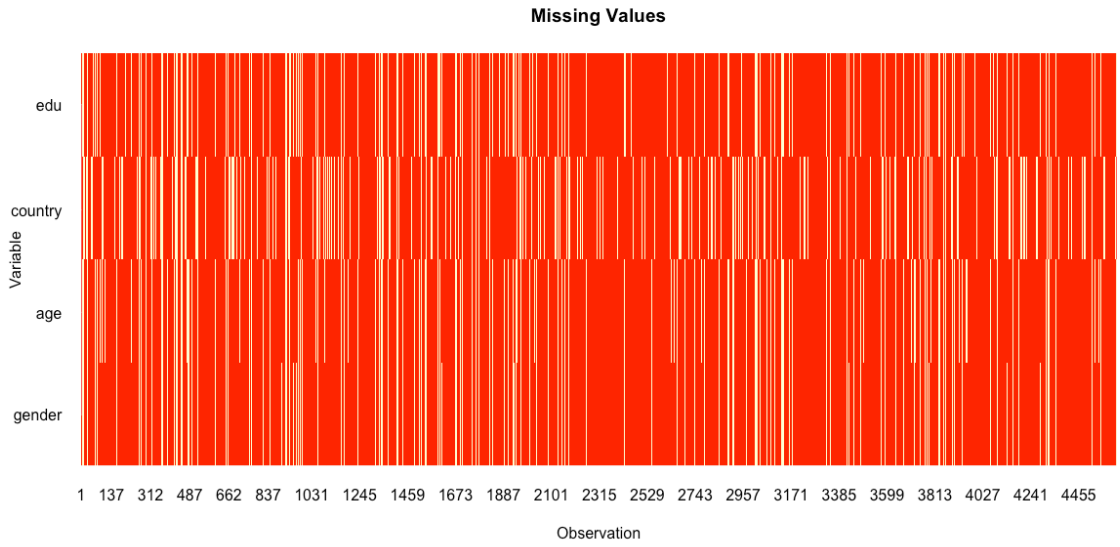


Figure 4. Visualization of the distribution of missing observations in relation to each other. Red represents data that is not missing. White represents missing data.

Table 3 shows the results of the logistic regression models regarding the missing data. When regressed on the dependent variable of completion, missing versus non-missing data of gender, age, country, and education was not able to predict completion and none of the demographic variables was statistically significant in the regression models.

Table 3. *Results of four logistic regression models using completion as the dependent variable and the missingness of each of the demographic characteristics as the independent variables.*

Variable	*B*	SE_B	$\boldsymbol{\beta}$	Sig.	AIC
Missing Gender	.00	.22	-.04	.97	1669.9
Missing Age	.00	.21	-.04	.97	1669.9
Missing Country	-.04	.19	-.21	.84	1669.9
Missing Edu	-.07	.21	-.34	.73	1669.8

Based on the data visualization evidence and that of the logistic regression models, the missing data was classified as *missingness at random* (MAR) since *missingness completely at random* (MCAR) is almost impossible to determine (Gelman & Hill, 2016). There seemed to be no observable patterns of missingness in the data, which served as justification for excluding the missing data when performing the complete cases logistic regression analysis on day 1 data that included demographic variables in the prediction. The *N's* for cases with no missing data are presented in Table 2.

Comparisons of Demographic Data for Completers and Non-completers

When comparing distributions of the demographic characteristics for completers and non-completers, several differences were evident. Although those choosing to be ID verified represent a minority of both completers and non-completers, it is evident in the distributions of ID verified and audit that a greater proportion of completers opted for ID verification than non-completers (Figure 5).

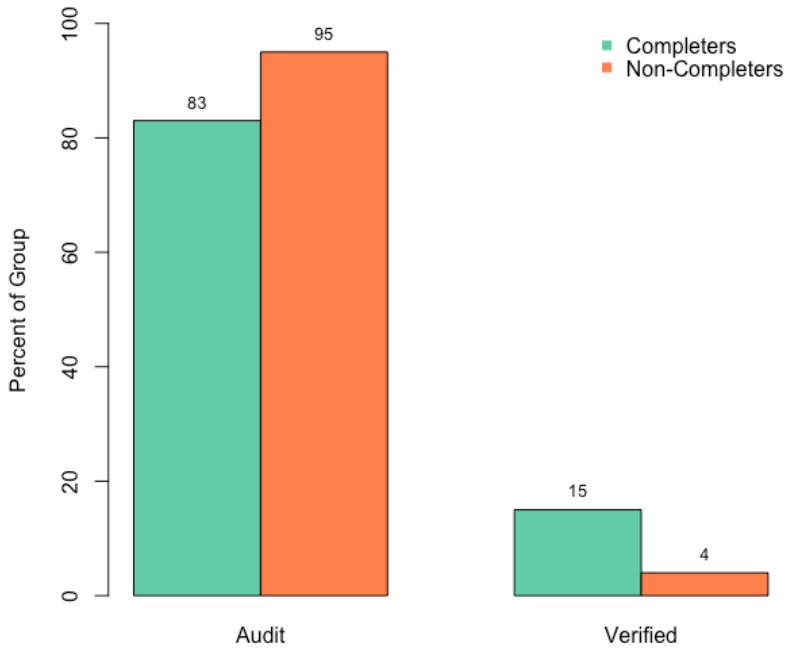


Figure 5. Comparison of ID verification for completers and non-completers.

Unlike ID verification, the distributions of gender for completers and non-completers differed substantially (Figure 6). Although males comprised about half of the course participants overall, among completers, they accounted for almost 70%. In addition, indicating gender on the survey appears to be highly correlated with completion and may be an indication of engagement.

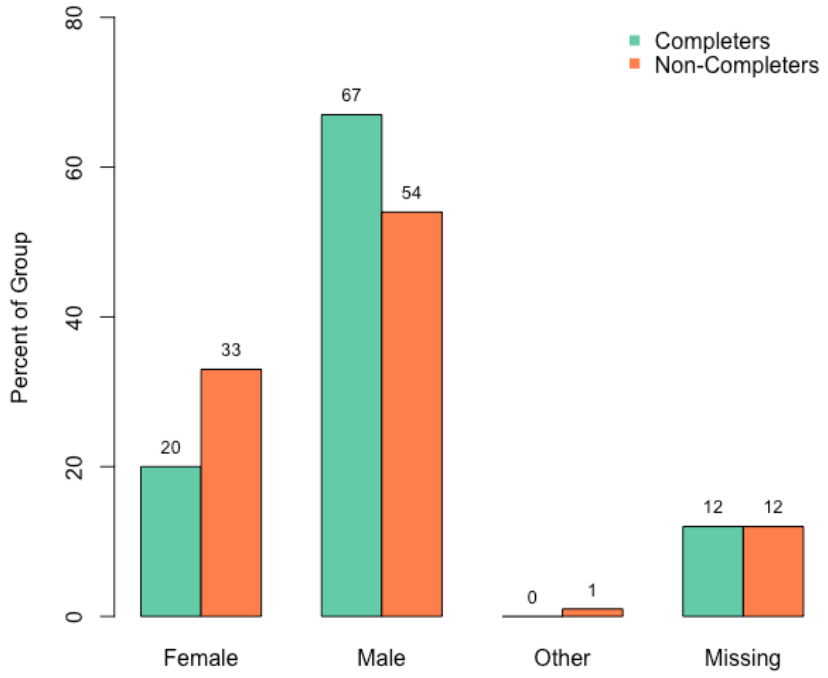


Figure 6. Comparison of distributions of gender for completers and non-completers.

The mean age of completers in this sample was 37 years whereas the mean age of non-completers was 30 years. Visualizing the distribution of age shows that a large majority the non-completers are younger than 36 years (76% of those who responded to the survey), whereas a majority of the completers were 36 years of age or older (59% of those who responded) (Figure 7).

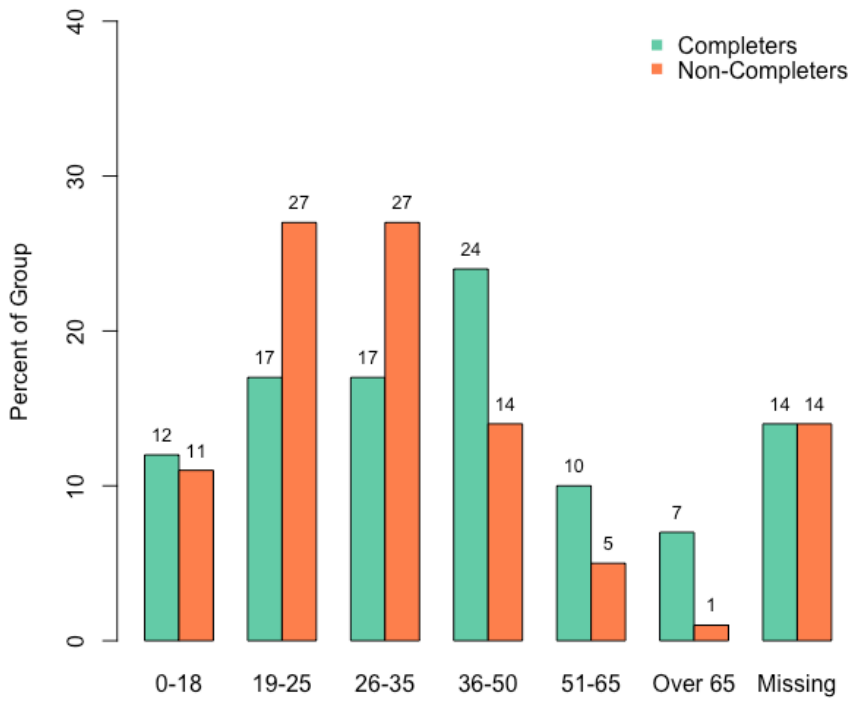


Figure 7. Comparison of age groups between completers and non-completers.

While there are differences in the age distributions of completers and non-completers, the educational background of both groups is very similar. Both have a large proportion of students with only a high school diploma and another large group who possess a college degree. In both groups, nearly the same proportions have only a high school diploma or some sort of degree (Figure 8).

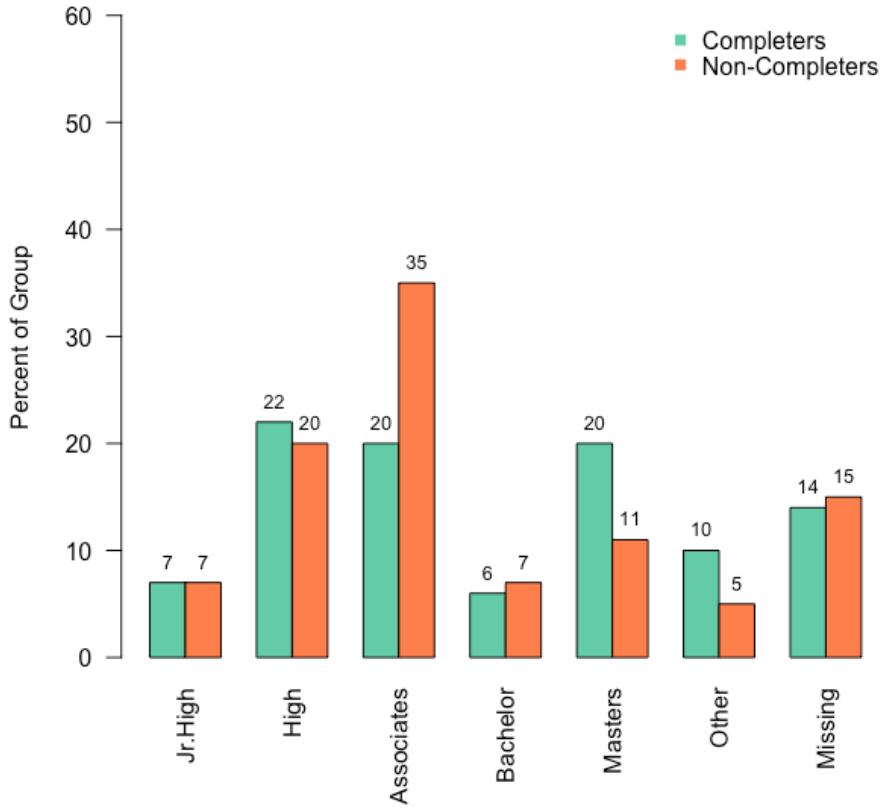


Figure 8. Visualization of the distributions of the highest level of education achieved by completers and non-completers.

Much like highest level of education achieved, the geographic distributions were very similar between the groups, with a slightly higher proportion of students in the non-completer group coming from North America (47%) versus those in the completer group (39%) (Figure 9). Around half of the students in both groups were from regions of the world other than the United States, Canada, or Mexico.

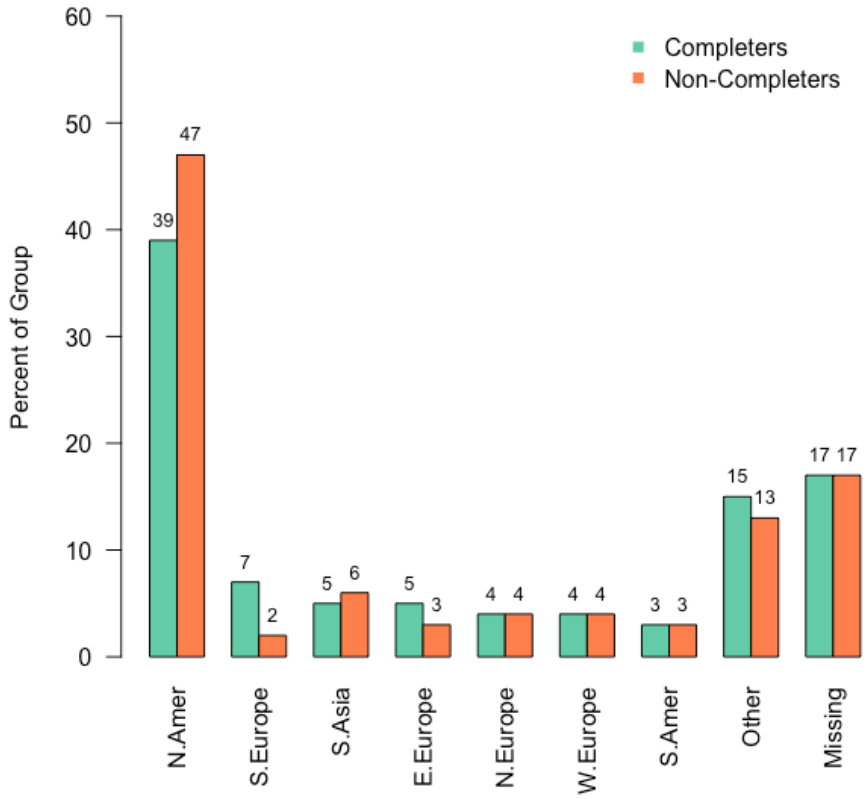


Figure 9. Distribution of students by region comparing between completers and non-completers.

Comparing Completers with Non-completers and Traces of Self-regulation

One of the objectives of this study was to examine the role that self-regulation may play with respect to achievement in this open-enrollment self-paced college algebra MOOC. Although individual self-regulation strategies and/or motivations cannot be directly observed through activity logs, there is evidence of self-regulation in the course behaviors that are present in the behaviors represented in the data. For example, in the daily activity logs, it can be observed that a student spent two hours working in the

course. Why this student did so is not obvious, but it is reasonable to assume that this data may constitute evidence of self-regulation.

SRL Measured as Maximum Time in One Day and Average Skills Gained. One way to examine the effects of self-regulation on achievement was to examine the correlations between the average time spent in the course on any day a student worked in the course and the average number of mathematics skills that were gained per active day in the course. These traces can best be seen by visualizing the data. These visualizations compare in graphic fashion the average number of skills gained each day with the average number of hours worked each by completers and non-completers. Regression lines represent predictions for each group based on these variables. It was also important to determine if these correlations varied depending on the extent of background knowledge a student brought to the course—especially those who scored in the first, second, and third quartiles on the pretest (individuals who scored in the fourth quartile on the pretest are of less interest because they have demonstrated that they have already mastered most of the course content). Figures 10–12 show the results of these comparisons.

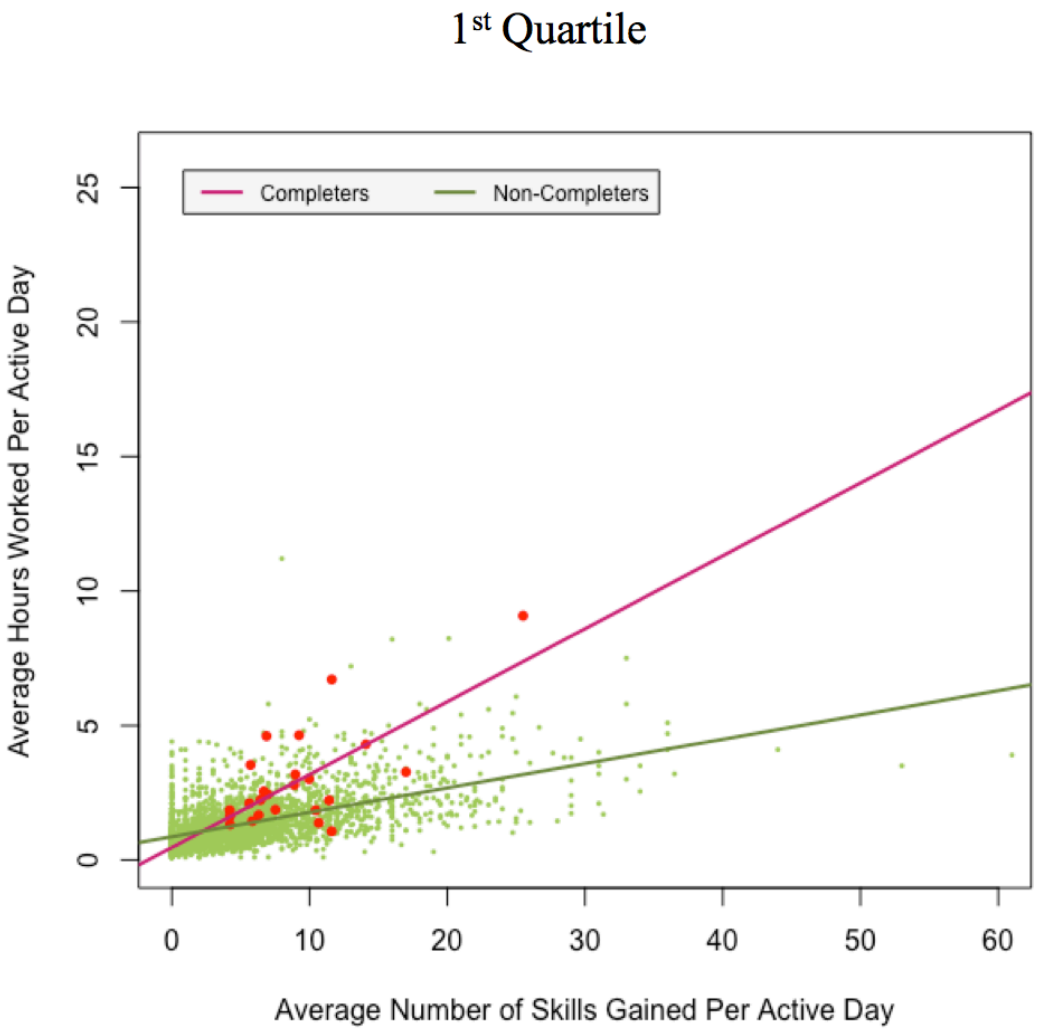


Figure 10. Comparison of SRL traces for students who placed in the 1st quartile on the pretest.

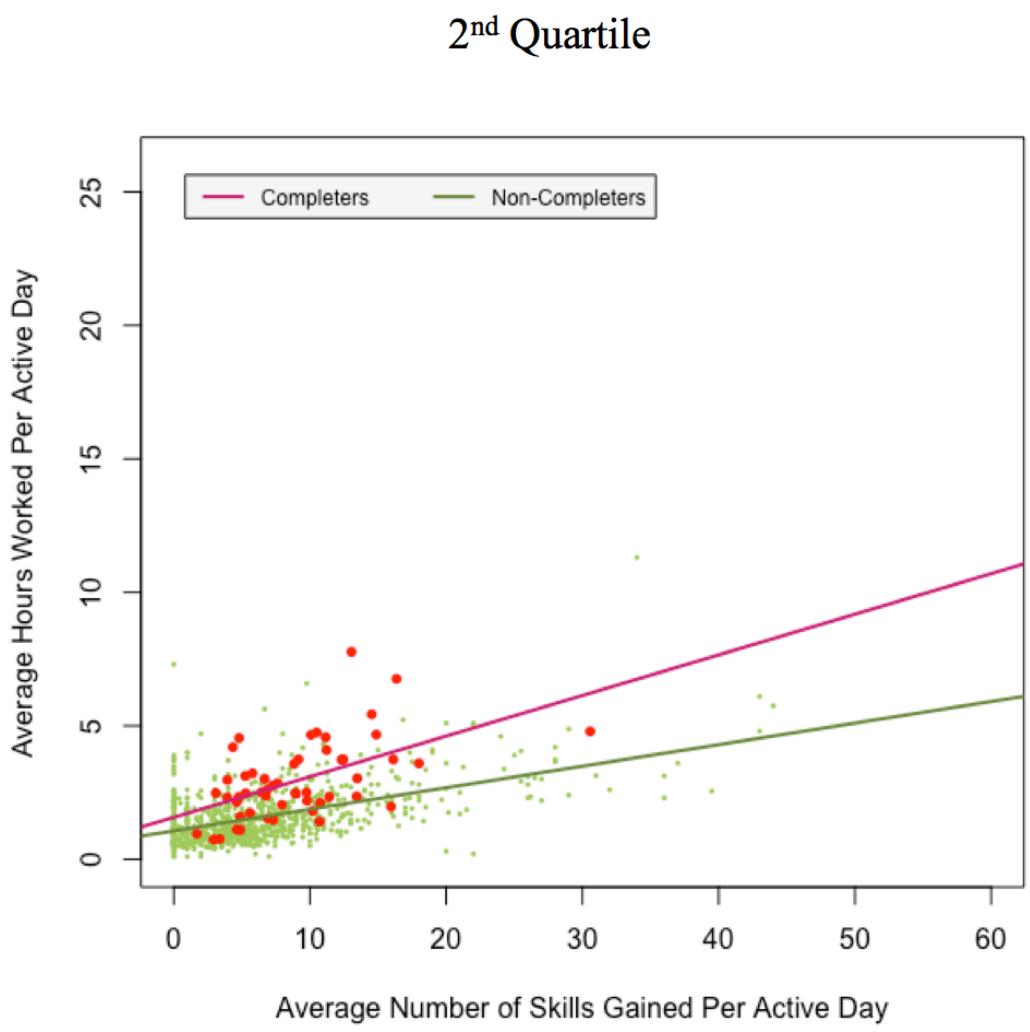


Figure 11. Comparison of SRL traces for students who placed in the 2nd quartile on the pretest.

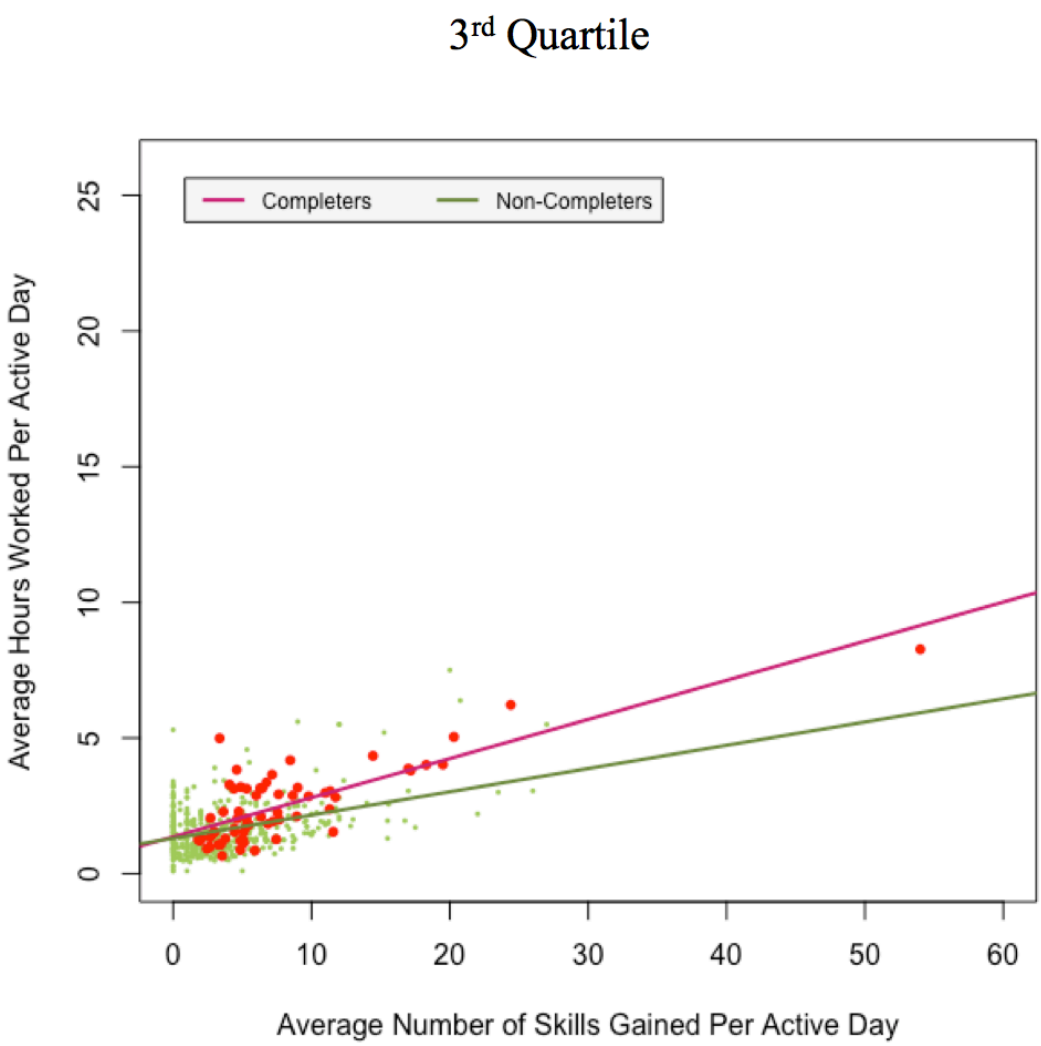


Figure 12. Comparison of SRL traces for students who placed in the 3rd quartile on the pretest.

By comparing these visualizations, it can be observed that completers in all quartiles show higher average hours worked and skills gained with the greatest difference in slope being reflected by students in the first quartile.

SRL Measured as Average Number of Hours Spent during Active Days.

Another indicator of self-regulation is the difference in the number of hours that

completers and non-completers devoted to the course averaged over the total number of their active days in the course. Table 4 documents the differences between completers and non-completers in time spent in the course in terms of the average number of hours spent on active days and total number of active days in the course. Students who scored in the first quartile on the pretest spent 1.6 hours more on average working in the course than non-completers who scored in the first quartile—reflecting the largest difference in all the quartiles of the average time spent working in the course on active days.

Table 4: *Relationship between completion average hours worked and average number of active days in the course.*

	Av Hrs/ A-day	Min Hrs/ A-day	Max Hrs/ A-day	Av A-days	Min A-days	Max A-days
All Complete	2.74	0.60	9.08	25.66	1	134
All Non-Complete	1.42	0.10	11.3	5.63	1	143
Q1 Complete	2.95	1.07	9.08	45.00	12	89
Q1 Non-Complete	1.35	0.10	11.2	5.81	1	143
Q2 Complete	2.95	0.74	7.77	36.85	7	134
Q2 Non-complete	1.54	0.10	11.3	6.11	1	81
Q3 Complete	2.51	0.65	8.27	27.56	3	69
Q3 Non-complete	1.60	0.10	7.5	4.25	1	36

Figure 13 illustrates how much more time on average completers spent working in the course during active days in contrast to non-completers. Overall, the majority of completers (64%) spent more than two hours on days when they were working in the course, whereas most non-completers (66%) spent 1.5 hours or less.

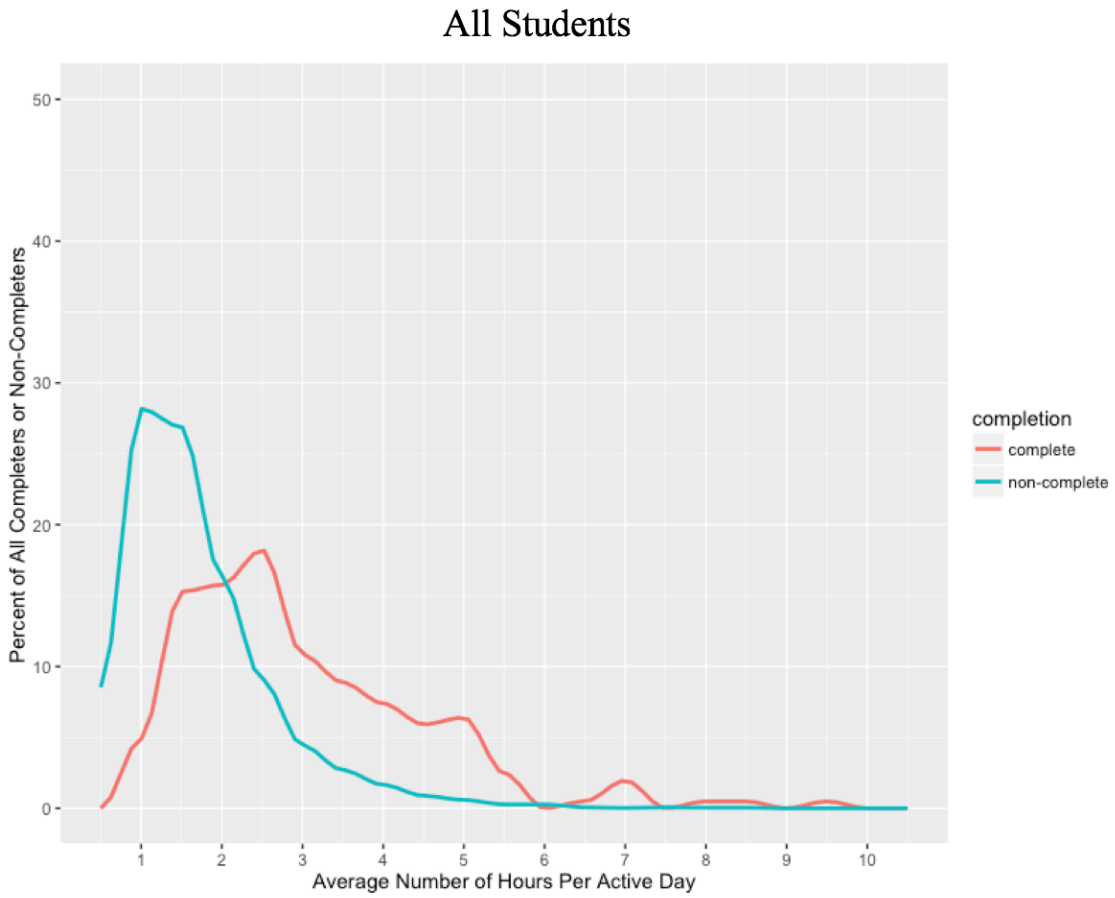


Figure 13. SRL traces shown as average hours spent per active day (all students).

When broken down by quartile, as in Figures 14–16, the average time spent by non-completers exhibits a definite peak and drop-off. In quartiles 1 and 2, the largest group of non-completers spent roughly an hour working in the course, whereas in quartile 3, the largest group of non-completers spent around 1.5 hours. Unlike the non-completer

plots, the completer plots are jagged, thereby reflecting clusters of students in each quartile working about the same amount of time in the course. This may be further evidence of self-regulation on the part of completers. These clusters could reflect groups of students who committed more time to working in the course because they naturally learn at a slower pace or because they had set goals for themselves to finish the course in a shorter period of time.

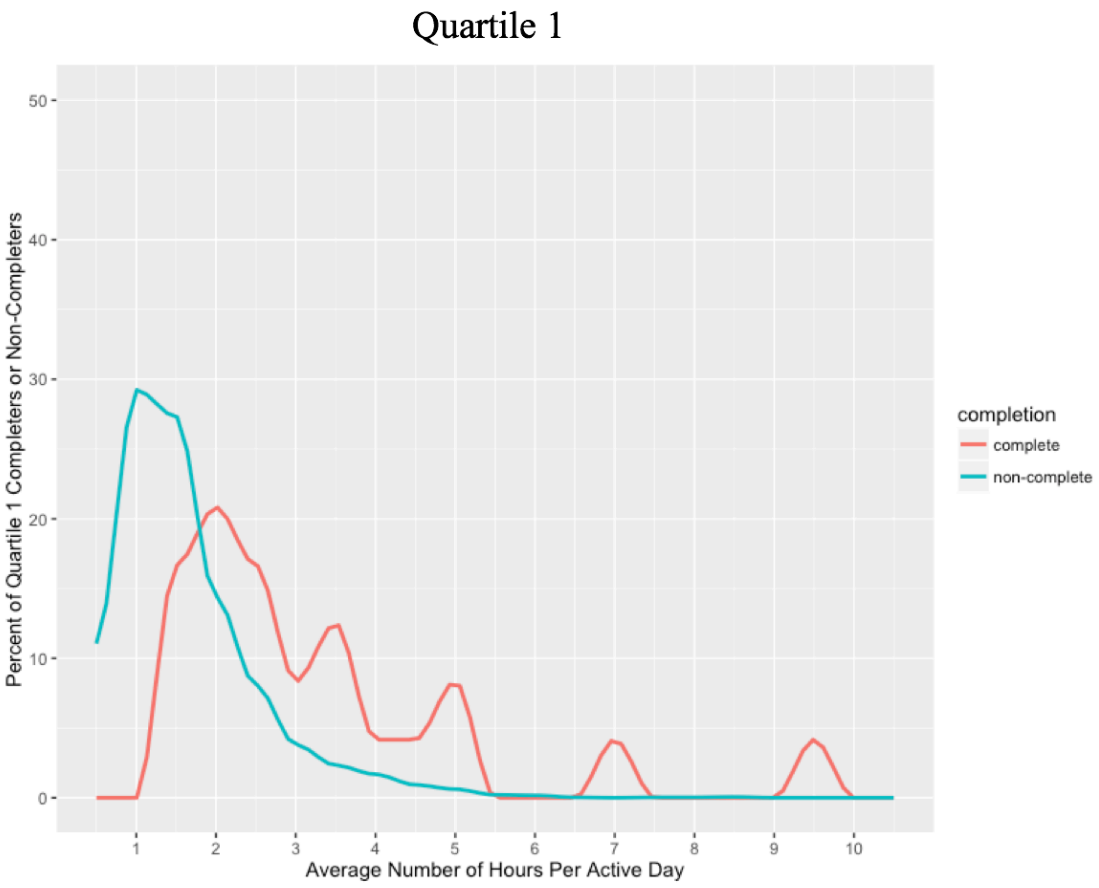


Figure 14. SRL traces shown as average hours spent per active day (1st quartile).

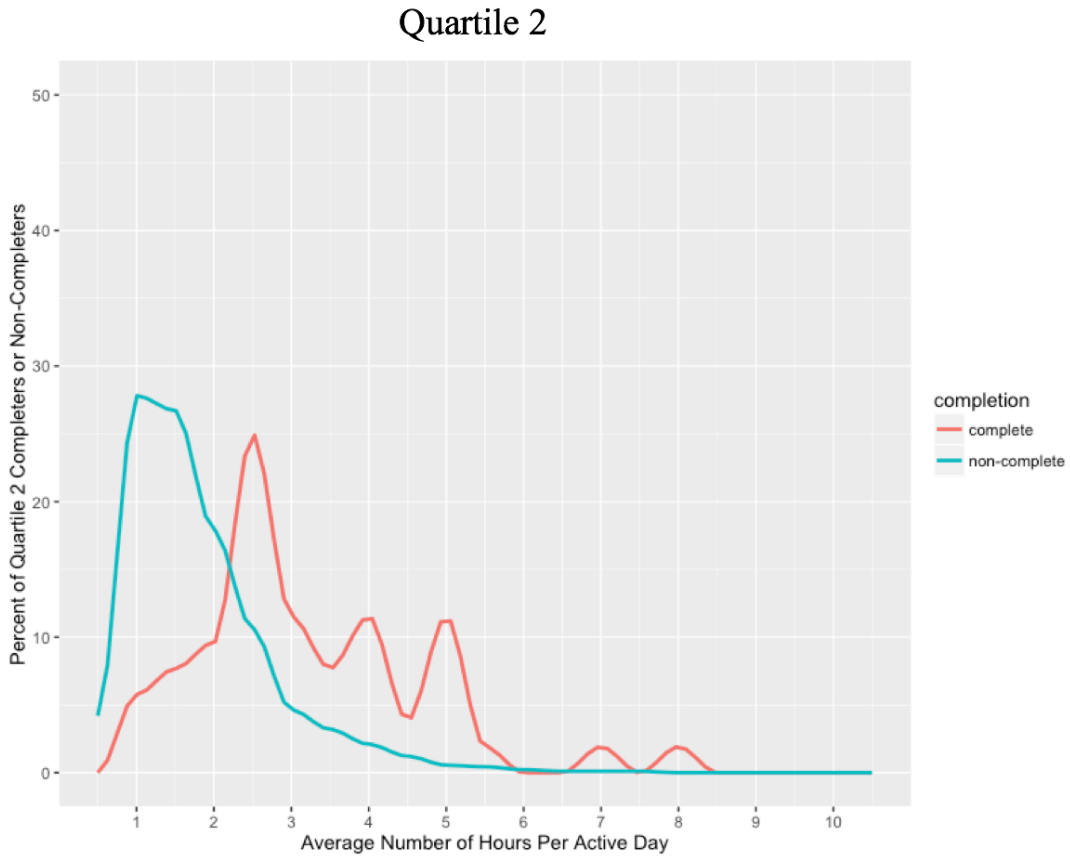


Figure 15. SRL traces shown as average hours spent per active day (2nd quartile).

As reflected in Table 3, Figure 15 shows that completers in quartile 2 put in the most time on average of all the students in the course, and that the majority of quartile 2 completers (52%) put in 2.5 or more hours into the course each active day whereas most of the non-completers (59%) worked in the course for 1.5 hours or less.

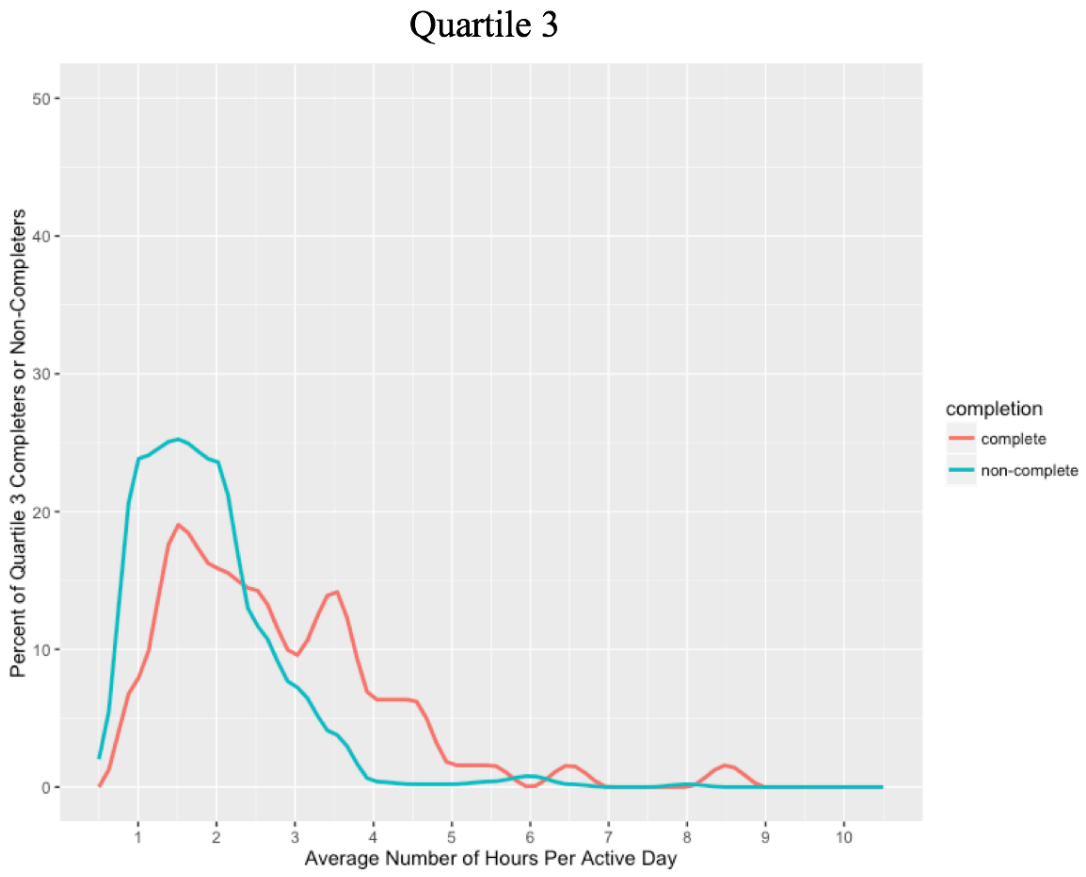


Figure 16. SRL traces shown as average hours spent per active day (3rd quartile).

SRL Measured as Early Indicator of Ultimate Achievement in Course.

A third way self-regulation was measured was by examining how well activity log data that seemed to indicate self-regulation on the first day of the course correlated with that student's achievement in the course. In this case, student achievement was compared with the number of mathematical skills gained after the pretest. Figures 17–19 illustrate the results of these relationships in visual depictions of linear regression predictions with Locally Weighted Smoothing (LOESS) to enhance the visualization of the relationship between the variables and to show trends (Cleveland, 1979).

Figure 17 shows the total number of skills gained after the pretest regressed against the number of hours worked on the first day. This figure indicates the strong relationship between the time spent by the student on the first day and that student's future performance in the course. The more time a student put in on the first day, the stronger was the probability that the maximum number of skills that student would achieve in the course would be high.

Figure 18 shows that there is a moderate relationship between the number of skills gained on the first day and the maximum number of skills a student will achieve by their last active day in the course. The shading on the graph shows that the variance for skills achieved on the first day is larger than the variance in the hours put in on that first day. This may mean that the hours put in on the first day are a signal of motivation—an SRL variable strongly correlated with achievement (Marzouk et al., 2016).

The wide bands of standard error on Figure 19 show the lack of correlation between velocity on the first day and the number of skills ultimately mastered by students after the pretest. When combined with other variables in the linear regressions, velocity was negatively correlated with achievement. This shows that working faster may be a sign of impatience or other indication of the lack of self-regulation.

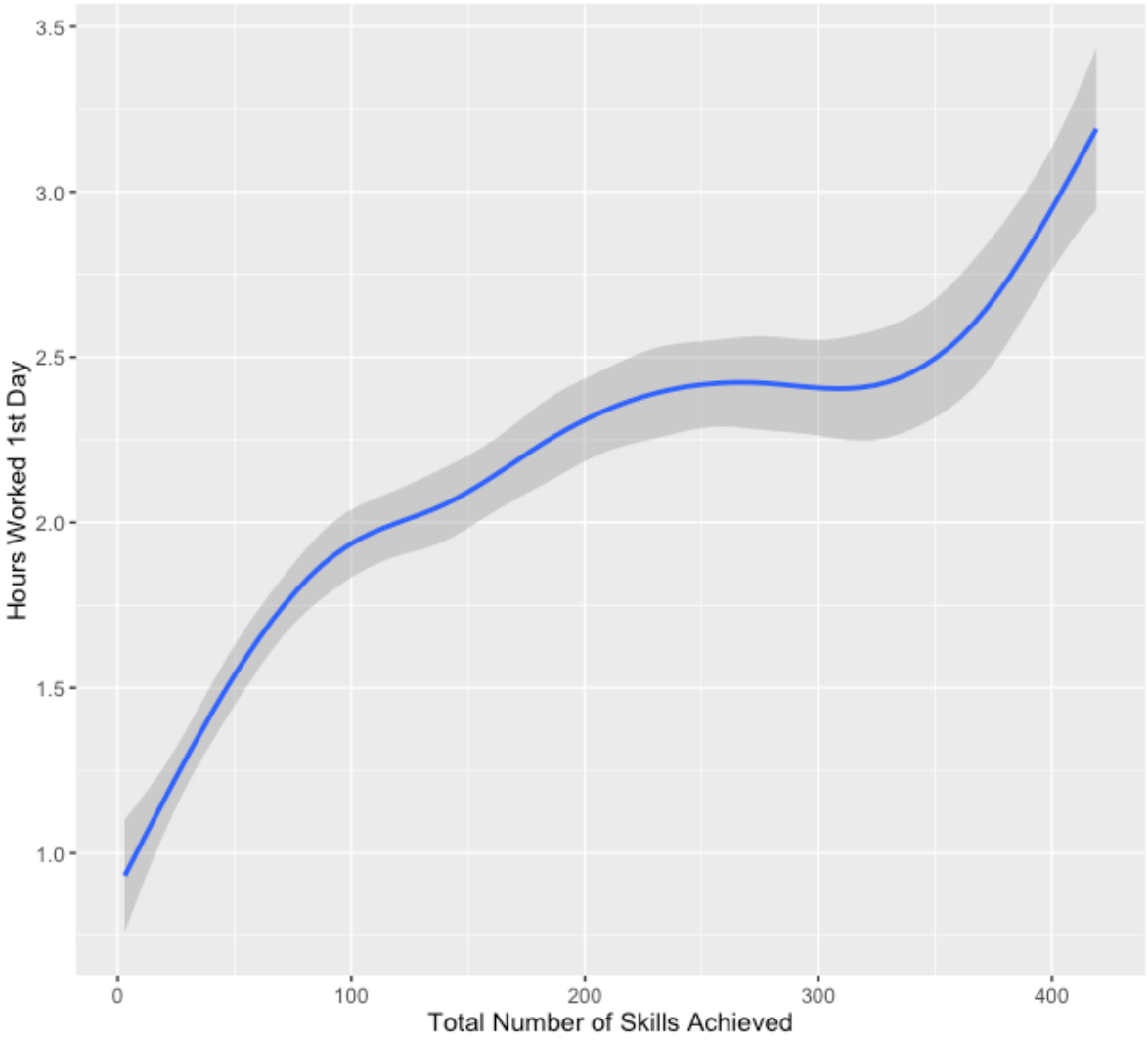


Figure 17. Total number of skills gained after the pretest (max = 419) regressed on the number of hours worked in the course on the first day with predictions fitted with LOESS smoothing. Shading reflects the standard error.

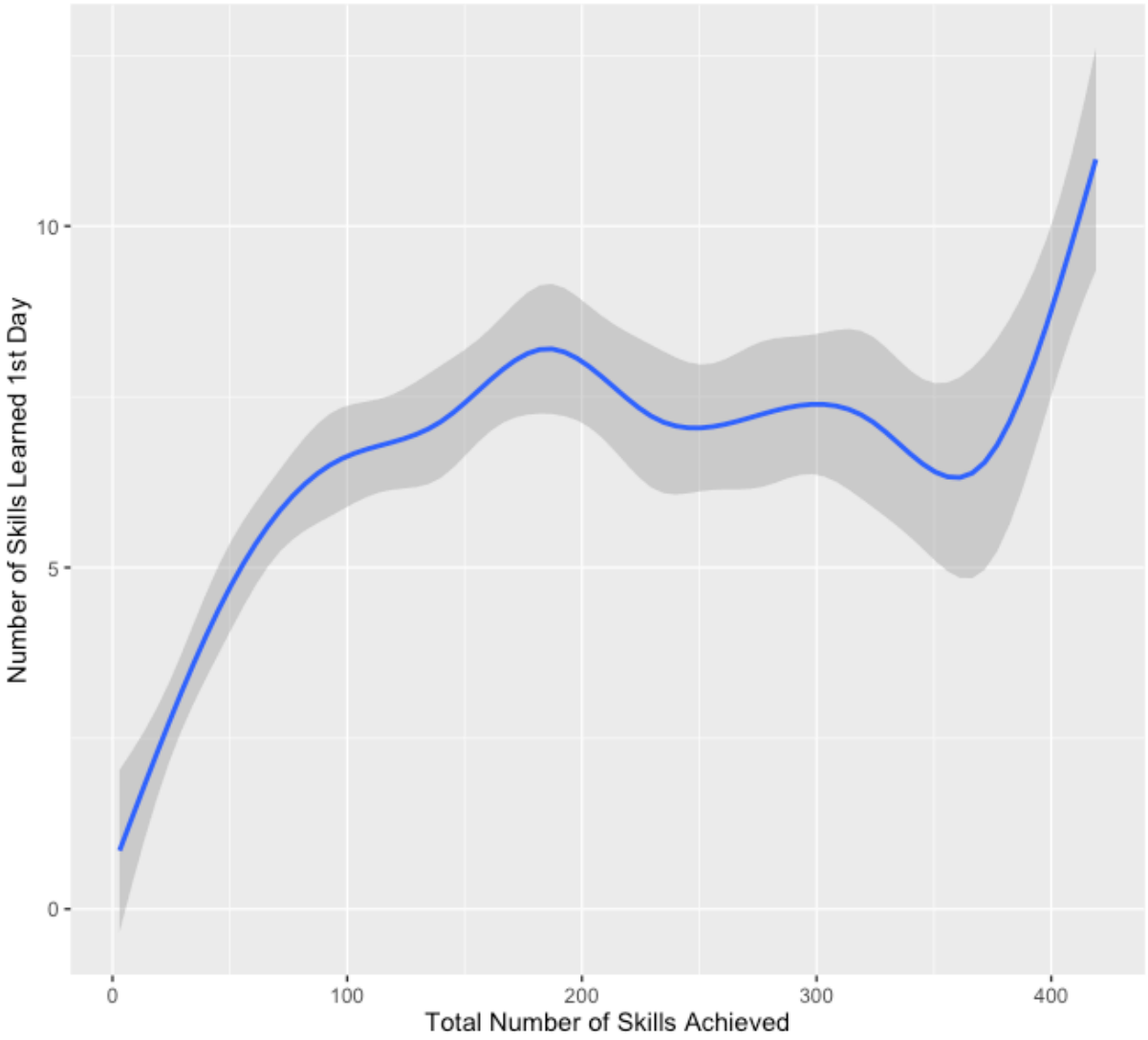


Figure 18. Total number of skills gained after the pretest (max = 419) regressed on the number of skills in the course achieved with predictions fitted with LOESS smoothing. Shading reflects the standard error.

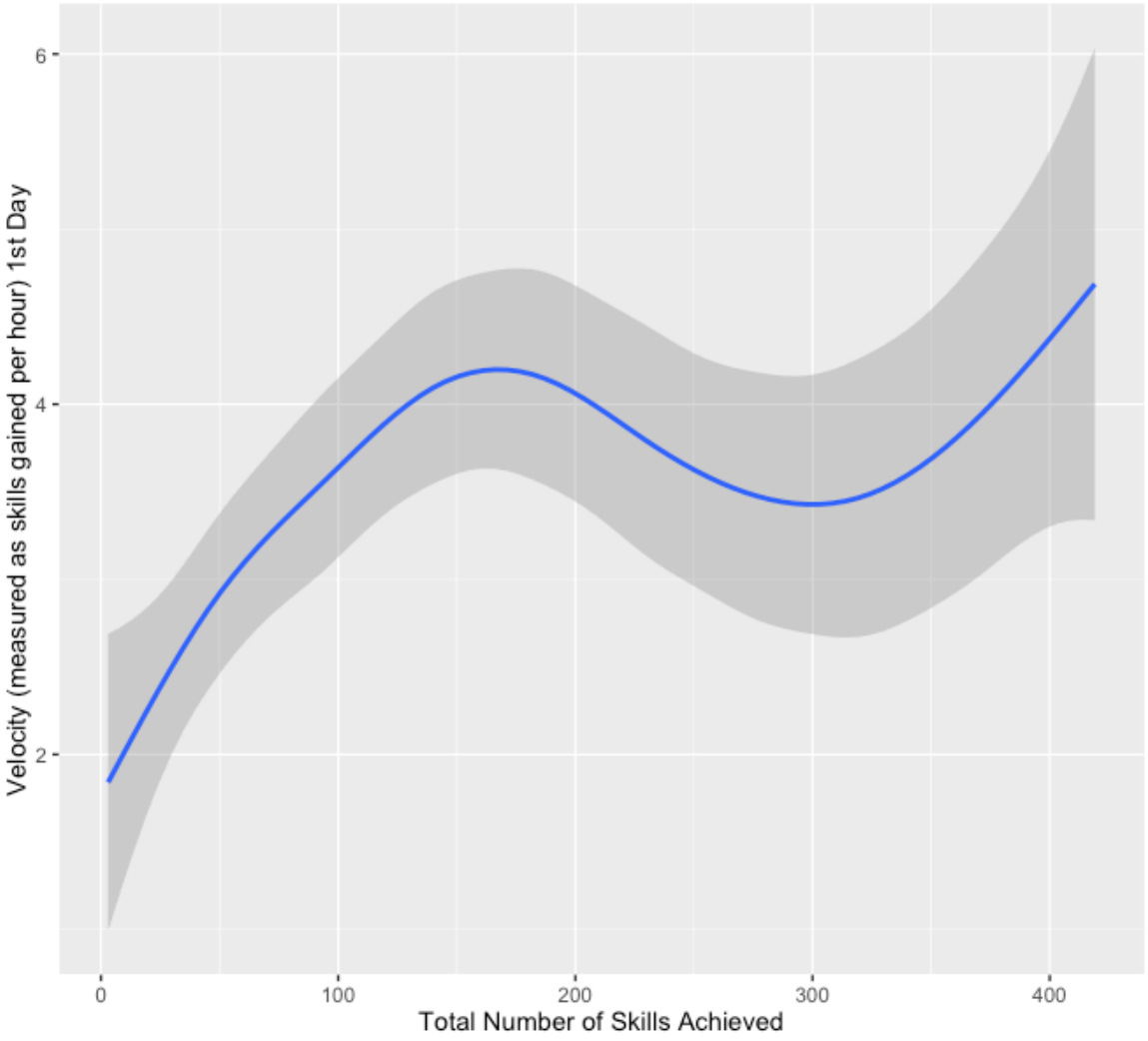


Figure 19. Total number of skills gained after the pretest (max = 419) regressed on the velocity on the first day with predictions fitted with LOESS smoothing. Shading reflects the standard error.

Linear Regression Models

Based on the available data, achievement can be measured in this college algebra course in two ways. It can be viewed as a binary outcome, completion versus non-completion, or in terms of how many of the 419 course math skills the student mastered.

Completion is the outcome of primary interest in this study. However, although the number of skills is actually an ordinal variable, this outcome can be treated as a continuous variable for the purposes of performing regression and making approximate correlations.

Using only the daily activity log data, eighteen independent variables were derived directly from the logs or calculated based on data that had been taken directly from it. Table 5 lists the characteristics of these variables.

Table 5. *Characteristics of the eighteen variables derived from daily activity logs for all students*

	Mean	SD	Skewness	Kurtosis
Pretest (419 max)	101.90	93.84	1.09	0.22
Topics learned after pretest	36.99	57.21	2.55	7.41
Percent course completion	33.17	26.43	0.84	-0.23
Total days in course	33.15	45.12	1.83	2.98
Active days in course	6.45	9.86	4.15	27.95
Hours spent in course	10.58	20.51	4.62	29.11
Average hours per day in the course	0.79	0.93	2.46	8.96
Average hours per active day	1.48	1.03	2.15	8.76
Total number of topics tested and learned	139	110.73	0.84	-0.23
Total number of topic learned	17.73	18.30	1.34	1.05
Ratio of active days to total days	0.50	0.38	0.26	-1.55
Velocity (topics gained per hour)	3.81	4.91	8.83	147.98
Velocity (topics gained per day)	23.56	38.80	2.68	8.32
Velocity (topics gained per active day)	5.28	5.86	2.26	8.70
Average velocity per active day	25.32	61.04	3.30	11.22
Max hours spent in a single day	2.64	2.22	2.23	8.23
Number of breaks away from course	2.63	3.85	3.00	12.41
Longest break away from course (days)	18.36	31.26	2.73	8.63

The large standard deviations in many of the variables in Table 5 reflect the great degrees of variability in this dataset. For example, “Topics learned after the pretest,”

"Total days in the course," and "Active days in the course" have standard deviations greater than their means. The high skewness and kurtosis values indicate that most of these values are not normally distributed and exhibit heavy tails. Using standard statistical procedures could be difficult when working with this kind of dataset. Fortunately, because there is a large amount of data, it is still possible to detect significant signals within the dataset when performing predictions. Large amounts of data is one of the big advantages to be gained from working with courses in a MOOC format (Ferguson & Clow, 2015).

Correlations Between Variables in Complete Dataset and Among 1st Day Variables

Next, with a focus on how well these variables might predict achievement in the course, correlations between these variables were examined and the outcome variable of how many skills the students mastered after taking the pretest. These relationships were tested at two levels—the whole dataset and only the data that could be gained from a student's first active day in the course.

The whole dataset correlations. Table 6 lists these correlations. The top five strongest correlations between the dependent and independent variables were the total hours spent in the course (.55), the total number of skills gained in the course (including skills determined in the pretest to have already been mastered) (.49), the average number of skills mastered divided by total number of days in the course (.49), the number of active days in the course (.44), and the maximum number of hours spent working in the course in a single day (.38). The potential for multicollinearity between these variables was a definite concern, so correlations between the independent variables were also examined.

Table 6. *Correlations between independent variables and the dependent variable* math skills gained after the pretest.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1 Math Skills Post-Pretest																	
2 Placement	.33																
3 Total skills	.49	.03															
4 Total days	.14	.02	.41														
5 Active days	.44	.04	.79	.53													
6 Hours	.55	.08	.83	.37	.83												
7 Av hours per day	.18	.14	.11	-.38	-.04	.18											
8 Av hours aday	.26	.21	.35	-.05	.11	.40	.70										
9 Max skills 1 day	.16	-.02	.70	.31	.48	.39	-.02	.23									
10 Aday ratio	.02	.02	-.14	-.60	-.17	-.08	.66	.13	-.25								
11 Velocity (hour)	-.02	-.06	.14	.05	.00	-.04	-.10	-.08	.38	-.10							
12 Max hours	.38	.17	.64	.25	.45	.67	.36	.76	.44	-.11	-.03						
13 Av skills (day)	.49	.03	.95	.40	.75	.81	.13	.36	.66	-.12	.13	.62					
14 Av skills (aday)	.11	.00	.41	.01	.05	.16	.32	.53	.64	-.03	.47	.44	.40				
15 Av velocity (day)	.04	.37	-.22	-.29	-.22	-.17	.42	.13	-.28	.50	-.09	-.13	-.19	-.08			
16 Breaks away	.24	.00	.62	.64	.83	.57	-.25	-.02	.49	-.38	.04	.32	.58	.02	-.27		
17 Longest break	.01	.01	.11	.86	.14	.09	-.39	-.07	.11	-.60	.05	.09	.13	.01	-.24	.22	
18 Pretest quartile	.32	.95	.01	.00	.03	.06	.12	.18	-.06	.02	-.07	.15	.00	-.03	.36	.00	.00

First day correlations. Correlations were examined and compared with the independent variables derived only from activity log data available from the students' first active day. The first active day is operationalized as the first day in the course in which the student completed the pretest. For some students, this first active day contained only pretest information. For others, after finishing the pretest, they went on to start learning some of the new math skills in the course. The difference was determined entirely by the choices made by the students themselves.

The number of available variables was substantially more limited (6 independent variables versus 17 independent variables for the complete dataset). This reduced number of variables was partially due to the unavailability of certain variables from the whole dataset, such as the maximum hours worked on a single day. Other variables were the same on the first day, such as total active days. Table 7 lists the correlations between the first day variables.

Table 7. *Correlations between the dependent variable and independent variables from day 1 data*

	1	2	3	4	5	6
1 Max skills post-pretest						
2 Placement	.32					
3 Total days	.00	-.02				
4 Hours	.16	.22	-.01			
5 Skills learned	.09	-.04	-.03	.48		
6 Velocity (hour)	.05	.33	-.02	-.20	.09	
7 Pretest quartile number	.32	.95	-.01	.18	-.06	.31

One of the strongest first-day correlations that was linked to achievement was the relationship between the number of hours the student worked on the first day and the maximum number of skills the student learned throughout the course. The total number of days in the course was negatively correlated with several other variables.

Determining which independent variables to use in models. By examining the correlations between the dependent variable and the independent variables and the correlations between the independent variables themselves, it was determined that a single variable should be chosen from four categories for the linear and logistic regression models. Three of these variables reflect self-regulation on the part of the students, including (1) a time variable that reflects how much time students were investing in the course, (2) a skill variable that reflects how many skills students were gaining as they worked in the course, and (3) a velocity variable that reflects how fast students were gaining skills in the course. The fourth variable was the student's placement on the pretest, which functioned as a covariate to control for background knowledge. It was also desirable to choose variables that were available both in the complete dataset and on the first active day in the course in order to be able to draw comparisons between models. As a result, the following SRL variables were selected:

1. *Total number of hours* spent in the course (time variable).
2. *Total number of skills learned* between formative assessments (skills variable). In ALEKS, students perform mathematical exercises and are then tested on the skills developed in those exercises. These tests involve formative assessments in a mastery-type format connected to the AI within ALEKS. Skills learned constitutes skills gained between these formative assessments.

3. *Total number of skills they learned divided by the total number of hours spent in the course* (velocity variable).

The above three variables, along with the *pretest score* (ranging from 0–419), comprise the four variables used in all of the linear and logistic regression models.

Multiple Linear Regression Models

Multiple linear regression models on the training set were created, comprising 70% of the complete dataset, a testing set with 30% of the complete dataset, a training set with 70% of the day 1 dataset, and a testing set with 30% of the day 1 dataset. Table 8 presents the *Ns* of these datasets.

Table 8. *Ns for each dataset. The Day 1 Complete Cases dataset was used only in the final logistic regression models that included demographic variables as predictors. Complete cases represent the students who had all the demographic variables present in their data.*

	All	Completers	Non-Completers
All Data (100%)	4623	203	4420
All Data (Train, 70%)	3236	142	3094
All Data (Test, 30%)	1387	61	1326
Day 1 Data (100%)	4623	203	4420
Day 1 (Train, 70%)	3236	142	3094
Day 1 (Test, 30%)	1387	61	13226
Day 1 Complete Cases (100%)	3264	143	3121
Day1 CC Train (70%)	2285	100	2185
Day 1 CC Test (30%)	979	43	936

The multiple linear regression models were created to determine the strength of the relationship between the dependent variable, the total number of mathematics skills acquired in the course, and the combination of self-regulation variables—hours spent in the course, mathematical skills learned, velocity as defined above, and placement on the

pretest as a control for background knowledge. Tables 9–12 present the results of these linear regression models.

Table 9. *Multiple linear regression of the complete training dataset.*

Variable	*B*	SE_B	$\boldsymbol{\beta}$	*t-value*	*Sig.*
Hours	1.78	.02	.33	90.05	< 2e-16***
Skills Learned	1.43	.02	.23	62.65	< 2e-16***
Velocity	-0.67	.07	-.03	-9.11	< 2e-16***
Pretest	0.99	.00	.84	262.73	< 2e-16***

Note. * $p < .05$ ** $p < .01$ *** $p < .001$. $R^2 = 0.97$

Table 10. *Multiple linear regression of the complete testing dataset.*

Variable	*B*	SE_B	$\boldsymbol{\beta}$	*t-value*	*Sig.*
Hours	1.80	.03	.34	53.76	< 2e-16***
Skills Learned	1.60	.04	.28	40.97	< 2e-16***
Velocity	-2.28	.20	-.07	-11.19	< 2e-16***
Pretest	0.98	.00	.82	160.23	< 2e-16***

Note. * $p < .05$ ** $p < .01$ *** $p < .001$. $R^2 = 0.96$

Since their outcomes were very similar, the results of the linear regression models of the complete dataset confirm a successful split between the data in the testing and training sets. Based on all the data available in this dataset, these model results also reveal the relative *beta* values of the variables. While all the variables were statistically significant, the *beta* values of the hours and skills learned were much larger than that of velocity. As would be expected, the pretest *beta* value indicating the level of background knowledge was also very large. Velocity—the dependent variable in the linear regressions—was negatively correlated with the number of mathematical skills mastered.

Table 11. *Multiple linear regression of the day 1 training dataset.*

Variable	*B*	SE_B	$\boldsymbol{\beta}$	*t-value*	*Sig.*
Hours	0.46	1.28	.01	0.36	0.72
Skills Learned	0.75	0.23	.06	3.19	< 0.001***
Velocity	-0.32	0.22	-.02	-1.49	0.14
Pretest	0.74	0.02	.61	40.65	< 2e-16***

Note. * $p < .05$ ** $p < .01$ *** $p < .001$. $R^2 = 0.38$

Table 12. *Multiple linear regression of the day 1 testing dataset.*

Variable	*B*	SE_B	$\boldsymbol{\beta}$	*t-value*	*Sig.*
Hours	-2.69	1.98	-.04	-.1.36	0.17
Skills Learned	1.32	0.37	.10	3.56	< 0.001***
Velocity	-0.48	0.30	-.04	-1.62	0.10
Pretest	0.75	0.03	0.59	25.47	< 2e-16***

Note. * $p < .05$ ** $p < .01$ *** $p < .001$. $R^2 = 0.35$

After examining the linear regressions of all the data, these results were compared with the multiple linear regressions of the data gathered from the students' first day in course. Although the amount of time spent was significant when the data from all days were included in the complete set, it was not predictive in the linear regression with respect to the total number of mathematical skills that would be ultimately mastered by the students. The number of skills learned, however, was predictive. Velocity continued to be negatively correlated but was not statistically significant.

Logistic Regression Models

Next, logistic regression models were run using the same independent variables as those used in the linear models but with completion/non-completion as the binary outcome rather than the continuous dependent variable of the total number of math skills

gained. These regression models were created using the four datasets noted above as follows:

1. Day 1 data with SRL variables: training set (70%).
2. Day 1 data with SRL variables: testing set (30%).
3. Complete cases day 1 data with SRL and demographic variables: training set (70%).
4. Complete cases day 1 data with SRL and demographic variables: testing set (30%).

Table 13 presents the results of these models while Table 14 presents the predictions resulting from the models. Figures 20–23 show the receiver operating characteristic (ROC) visualization curves of the logistic regression models. The strongest model results were in the day 1 logistic regression model using the complete dataset of student cases. This model performed with an overall prediction accuracy rate of 76% for the data training set and an 82% prediction accuracy rate in the data testing set. The model that based its predictions only on activity log data demonstrated an overall prediction accuracy rate of 71% in both the training and testing datasets.

Table 13. *Logistic regression model comparison. Day 1 Train and Day 1 Test are based on the first day of the whole dataset. Day 1 Train CC and Day 1 Test CC are based on the complete cases.*

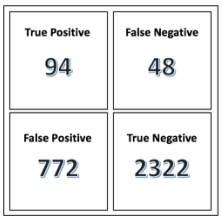
	Model 1 Day 1 Train	Model 2 Day 1 Test	Model 3 Day 1 CC Train	Model 4 Day 1 CC Test
Hours	0.11*	0.01	0.15*	0.21*
	(0.05)	(0.07)	(0.07)	(0.11)
Skills Learned	0.02	0.04**	NA	NA
	(0.01)	(0.01)		
Pretest	0.01 ***	0.01 ***	0.01 ***	0.01***
	(0.00)	(0.00)	(0.00)	(0.11)
ID Verified	NA	NA	1.67***	2.38***
			(0.33)	(0.48)
Region[a]	NA	NA		
N. America			-1.09*	--
			(0.54)	
N. Europe			-1.63*	--
			(0.80)	
S. Asia			-1.35*	--
			(0.67)	
Age[b]	NA	NA		
51-65			1.07*	--
			(0.48)	
Over 65			--	2.67**
				(0.93)
Gender	NA	NA	--	--
AUC	0.72	0.68	0.83	0.84
Correct Predictions	2306	1081	1619	849
Incorrect Predictions	930	306	666	130
Total *N*	3236	1385	2285	979
McFadden R^2	0.142	0.107	0.229	0.281

Note. * $p < .05$ ** $p < .01$ *** $p < .001$. [a, b] Only values with significant levels of at least $p < .05$ are reported.

Table 13 shows the results for all the models. Aside from the pretest results, the number of hours worked on the first day was the most consistent predictor. From the SRL

variables alone, skills learned on Day 1 was predictive. However, when demographic data was included, skills learned was no longer predictive and was dropped from the models. ID verification was a significant predictor, along with region and age. Gender added to the predictive value of the models but was not statistically significant.

Table 14. *Confusion matrices displaying the accuracy of predictions in four different logistic regression models*

	Day 1 Train	Day 1Test
Day 1 SRL Data Only		
Sensitivity	0.662	0.639
Specificity	0.750	0.786
Prediction Average	0.706	0.713

	Day 1 CC Train	Day 1 CC Test
Day 1 Complete Cases With Demographic Info		
Sensitivity	0.820	0.767
Specificity	0.703	0.872
Prediction Average	0.762	0.820

The confusion matrices shown in Table 14 present the actual predictions made by the models. Sensitivity indicates the percentage of time that the model correctly predicted which students would be completers. Specificity indicates the percentage of time the model correctly predicted which students would be non-completers. True positives are the actual number of correctly predicted completers. False positives are individuals whom the model predicted would be completers but were actually non-completers. True negatives are individuals whom the model correctly predicted would be non-completers. False negatives are individuals whom the model predicted would be non-completers but were actually completers. An average correct prediction rate of 71% was achieved using the activity log SRL data alone. The training and testing sets predicted comparably. In the complete cases dataset that included the demographic data, the average correct prediction rate in the training and testing sets was 79%. The testing set predicted slightly better than the training set, achieving a correct prediction rate of 82%.

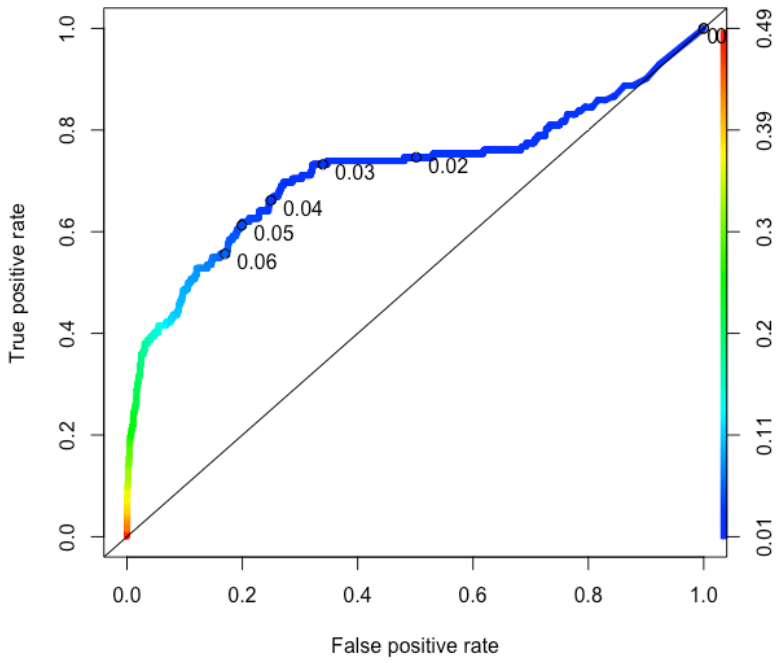


Figure 20. Receiver operating characteristic (ROC) curve for day 1 training model.

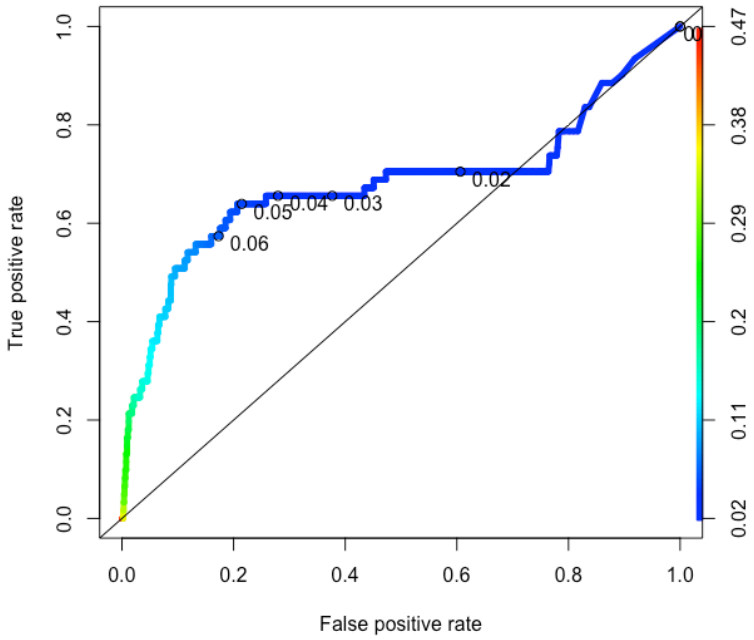


Figure 21. ROC curve for day 1 test model.

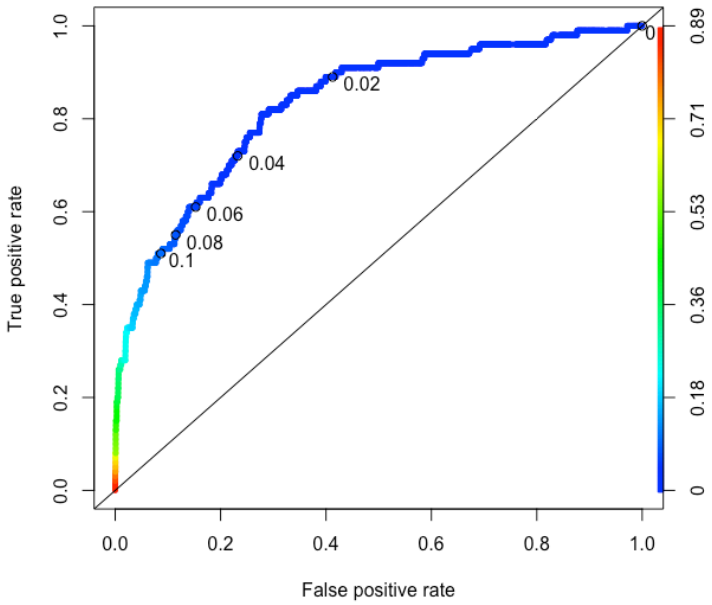


Figure 22. ROC curve for day 1 training model with demographic information included as predictors.

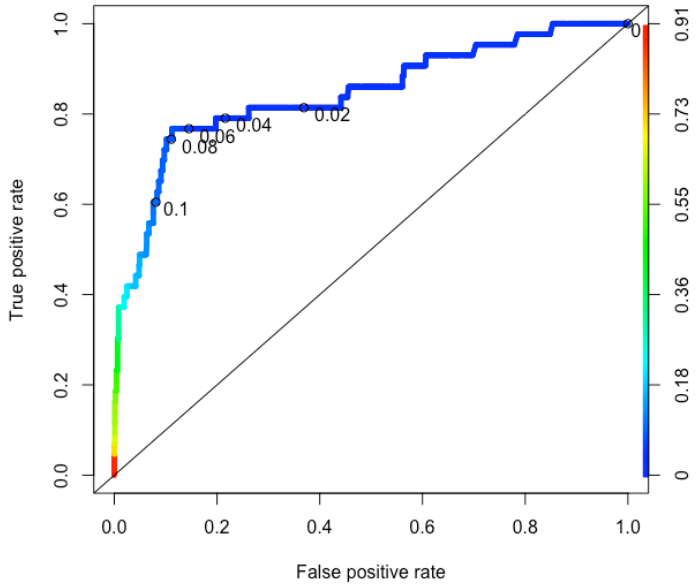


Figure 23. ROC curve for day 1 test model with demographic information included as predictors.

Figures 20–23 above show the receiver operating characteristic (ROC) curves. The ROC curves in Figures 20 and 21 depict the predictions made by the logistic regression models (both training and testing) that used activity log data alone to predict completion by individuals in the college algebra course. The tick-marks on the curves reflect the effective cutoffs for these models. The cutoff point that yielded the most accurate predictions for the training set (Figure 20) was 0.04, whereas the most accurate cutoff point for the test set was 0.05. These are fairly aggressive cutoff points that reflect the unbalanced nature of the dataset (4% completers and 96% non-completers). These cutoffs mean that when the model predicts that a user has a 4% or 5% chance of being a completer, it will predict that the student will complete the course. If the model determines that the chance of that individual completing is less than 4%, it will predict that individual to be a non-completer.

In Figures 22 and 23, the shapes of the curves away from the halfway 50/50 prediction rate that cross the ROC curves diagonally reflect the improvement in the model's prediction rate by the addition of the demographic data. The area-under-the-curve (AUC) rates also reflect this improvement. While the average AUC for the training and testing sets that used activity log data alone was 70% (see Table 13), the average AUC for the models that included the demographic data was 84%.

CHAPTER 5

Discussion

The objective of this research was to examine data generated by students working online in a self-paced mathematics MOOC for evidence that self-regulation plays a role in who completes and who does not complete this course. Predictive models were created to measure the relative importance of these behavioral patterns as evidenced through the data, and these behavioral variables were then combined with demographic information not only to increase the predictive power of the models but also to measure how these variables behave in relation to differences in demographics that also influence completion. There was a further objective for creating these prediction models beyond shedding light on what behavior and demographic variables were influential in this type of MOOC. Although completion is not the objective of every MOOC user, one of the reasons for creating these early prediction models was to assist in the future creation of effective interventions that could contribute to the success of students who do wish to complete.

A Review of the Study Methodology

To accomplish these goals, a three-phase approach was implemented. The first phase was to examine data in the daily activity logs that could contain evidence of self-regulation on the part of the learners and be linked to successful completion of the course. The second phase was to test the predictive power of these behavioral data through linear and logistic regression. The third phase was to combine these variables derived from these data with demographic information gained from a survey conducted by edX to understand how behaviors within the course and demographic characteristics work

together to influence completion. Because one of the objectives of this research was to examine these variables with an eye to interventions that could help students in the future, the final logistic regressions were designed to predict with only data derived from the first day in the course. The purpose in relation to this objective was to test how early signals could be detected that could influence completion.

Seventeen variables were developed either directly from what was recorded by computers through the activity logs or calculated from data by combining these variables (Tables 6 & 7). Because students in a self-paced mathematics MOOC often work in isolation with very little extrinsic pressure to make progress, it seemed reasonable to assume that many of these variables could result from self-regulation. For example, how long a student works in the course each day, or how many days each week they are active, or how quickly they progress from one math skill to another seem to most likely emerge from motivation and impetus that arise from within the student alone when they are working in this kind of MOOC environment.

Three Important Behavioral Variables

In order to choose which of the seventeen behavioral variables should be included in the regression models, correlations with the outcomes of the course were examined. Two outcomes were considered, the total number of mathematics skills learned over the total amount of time the students spent in the course and the binary outcome of course completion/non-completion. They were also evaluated from two perspectives: 1) considering all the days the students worked in the course in relation to the outcome variables, and 2) just considering the first day the students worked in the course. Three variables emerged as most important: time spent working in the course, the number of

math skills learned, and the velocity at which students mastered math skills in the MOOC. These three variables were operationalized in two different ways. The first way took into account all the data generated during the eight month period of the study that could be derived from these variables. The second way only took into account data that were generated on the first day students worked in the course.

Time. For the regressions that took into account all the data in the course, the time variable was operationalized as total hours spent by students in the course. In the regressions that predicted based on just first day data, the time variable was operationalized as the total time the student worked in the course on the first day.

Mathematics skills learned. For the regressions that took into account all the data in the course, the skills learned variable was operationalized as the total number of mathematics skills learned between formative assessments over the entire course. In the regressions that predicted based on just first day data, the skills learned variable was operationalized as the number of mathematic skills that were learned after the pretest on the first day.

Velocity. For the regressions that took into account all the data in the course, the velocity variable was operationalized as the total number of skills learned between formative assessments during the whole time a student worked within the course divided by the total number of hours worked in the course. In the regressions that predicted based on just first day data, the velocity variable was operationalized as the number of skills learned after the pretest on the first day divided by the total number of hours spent on the first day.

Two Different Kinds of Models

The primary purpose of the models that used all the data in the course was exploratory. The relative predictive power of each of these three variables, time spent, skills learned, and velocity could be measured when combined in a single regression model that incorporated the full amount of data available for each of these variables from each student regressed on the total amount of mathematics skills learned in the course (Tables 9-10). As expected, these three variables, along with the pretest functioning as a covariate, were able to capture most of the variance in the data ($R^2 = .96$). The strongest predictor variable in this first linear model was total hours spent in the course ($\beta = .34$) with the next strongest predictor being the number of skills learned ($\beta = .28$). The weakest predictor was velocity and was negatively correlated to the maximum skills learned in the course ($\beta = -.07$). All three of the variables examined were statistically significant ($p < .001$) as part of this model.

In the second set of linear and logistic regression models, the predictive power of the three independent variables, time spent, skills learned, and velocity were measured using just the data generated by students on the first day in the course. In a first set of exploratory regressions it was desirable to see how these variables using just the data from the first day predicted how many skills the student would ultimately earn during their entire time in the course. A set of models regressing each of these variables individually on total number of mathematics skills learned explored this behavior (Figures 17-19). Another set of linear regression models examined how the variables behaved when combined and regressed against the total number of mathematics skills in the course (Tables 11-12). Finally, a set of logistic regression models examined the three

variables using just the data generated on the first day regressed on the binary outcome of completion/non-completion (Table 13).

Time Spent

The most consistently predictive behavioral variable was the time variable. The only variable that was more predictive in any of the regressions was the student background knowledge measured by the score the students achieved on the pretest that was acting as a covariate in all of the models. In the multiple linear regression model using all the data of these three variables that students generated over their entire time in the course, time spent working in the course had the largest *beta* of the three independent variables (β = .34, Table 10). In three of the four of the logistic regression models using just the data generated on the first day in the course, time spent on the first day working in the course was statistically significant with a $p < .05$ (Table 13).

The hours variable was not significant in the linear regression models using just first day data with the total number of skills learned as the dependent variable. This is a reflection of both multicollinearity with the skills learned variable, and the nature of the dependent variable. While dependent variables of total number of skills learned and completion are closely related, they are not the same. Many students in this dataset learned a large number of skills but did not go onto complete. While completion is operationalized at learning 90% or more of the total 419 skills in the course, 8% of non-completers learned 70% or more of the skills but did not go on to complete. So the non-significance of the hours variable in the first day linear regression should be regarded as informative as far as exploring the dataset but should not be regarded as a definitive

explanation of the impact of time worked on the first day on the dependent variable of final completion.

Skills Learned

The second most predictive variable was the skills learned variable. In the regression model that drew from all the data in the course, this variable was the third most predictive after the pretest and time spent in the course (Table 10). When taking into account all the data in the course, time spent and skills earned was highly correlated with a correlation coefficient of .83 (Table 6); however, it was only modestly correlated in the first day data with a correlation coefficient of .48 (Table 7).

Despite the modest correlation between skills learned and time spent on the first day, the two variables responded similarly in the logistic regression models. For example, including or excluding the time spent and skills learned variables resulted in only slight differences in AUC or McFadden's R^2. In Model 1 of Table 13, including both time and skills learned resulted in a model AUC of 0.72 and a McFadden's R^2 of 0.14. The same model excluding hours worked would have produced an AUC of 0.71 and a McFadden's R^2 of 0.13.

Velocity

Velocity was the least predictive of the three variables. In many cases, it was negatively correlated with the outcome variable meaning that the faster a student worked in the course, the less likely that student was to complete the course (Tables 9-12). In the linear regression models using only the first day's data, velocity was not statistically significant and did not add to the predictive value of the logistic regressions and was therefore not included in the final models.

Multicollinearity

One of the challenges of using these pieces of information from the daily activity logs was that many of these variables were highly correlated posing the threat of multicollinearity. For example, time spent working in the course and the number of math skills learned are very closely related as it takes time to work through each math skill. The challenge was to produce parsimonious models that captured the most important differences between completers and non-completers while minimizing multicollinearity. One of the ways multicollinearity was minimized was by keeping the number of variables in the models low and seeking to draw variables from data that would offer different information even if related to a certain degree. To achieve this, the correlations of variables were examined and their behavior in relation to each other was carefully monitored in the linear and logistic regression models.

ID Verification

ID verification was a unique variable that was not part of the data recorded in the daily activity logs, nor was it part of the survey data collected by edX. Rather it was an option offered to students who wish to take the final exam and receive college credit or a certificate of completion. In the MOOC platform edX, ID verification takes place when students pay a nominal fee to have their IDs verified through automated software created for this purpose within the MOOC (edX, Inc., 2017b). Students must provide a photo of themselves and their government issued identification by taking a picture with their computer webcam. ID verification was a more common characteristic of completers than non-completers (15% versus 4%) confirming other research showing ID verification to be a key indicator of completion in MOOCs (Ho et al., 2015). It was also predictive of

completion and was statistically significant ($p < .001$). Only 4% of the students in this sample opted into being ID verified and of those students 16% went on to complete the course.

Important Demographic Variables

When demographic information was added to the logistic regression models, the predictive power of the models increased. These were the demographic variables that emerged as most important in the models.

Age. The mean age of course participants was 30 years old; however, the mean age of completers was 37 years of age. This was an important indicator that this sample was largely made up of non-traditional students and that older students were more likely to complete in this course. Because of this, age was predictive in the logistic regressions with the oldest age groups of 51 to 65 and over 65 being statistically significant ($p < .05$ and $p < .01$ respectively, Table 13). It could be speculated that this is related to the greater self-regulation of these older students. While older students have been shown to have more self-regulation strategies, this has been primarily demonstrated in populations that are much closer to the ages of traditional college students (Usher and Pajares, 2008). There is little research into levels of self-regulation in older adult learners.

Gender. The variable of gender was especially interesting. Slightly more than half of the sample (54%) self-identified as male while 32% self-identified as female. In this sample, 13% self-identified as "other" or did not answer the survey question for gender (Table 2). While males made up 53% of non-completers, they constituted 67% of completers (Figure 6). However, gender was not statistically significant as a predictor in the models tested, although including it as a predictor variable did add to the predictive

power of the model. This predictive power of the gender designation did not seem to come from a difference in behavior between males and females. For example, mean number of hours spent on the first day by male and female completers was almost identical (2.548 for males versus 2.553 for females). The average amount of time spent in the course by male and female non-completers was also very similar (1.505 for males versus 1.499 for females). Likewise, the mean number of math skills learned for male and female completers on the first day was very close with females slightly outperforming males (7.162 skills for males versus 8.195 for females). The mean number of mathematics skills completed by non-completers on the first day was also close with females outperforming males by a narrower margin (4.102 skills for males versus 4.618 skills for females). What made gender predictive of completion in the logistic regression models was only the fact that more males than females completed the course. However, it was a very weak predictor because the behavior of males and females on the first day of the course was almost identical.

Region. The five most common countries of origin for course participants were the United States (44%), India (4%), Canada (3%), Great Britain (3%), and Australia (2%). The geographic location of course participants when grouped by region was weakly predictive. Three regions, North America, northern Europe, and southern Asia were statistically significant (all with $p < .05$) when the larger training dataset was used; however, in the smaller testing set, no regions were statistically significant. The three regions that were statistically significant in the training set were negatively correlated with completion but were also in the top five most common regions from which students

originated (Figure 9 & Table 13). This can viewed as an artifact of the high N's of these regions in highly unbalanced dataset.

Educational Background

One characteristic of students that did not seem to influence completion was educational background. This demographic is operationalized in the edX survey from which all the demographic data for this course was gathered as "highest level of education completed" (edX, Inc., 2017a). A high school diploma was the most common level of education chosen by the entire sample (35%). When broken out by completers and non-completers, approximately half of both completers and non-completers had a college or university degree (46% for completers and 53% for non-completers), and over a quarter of both groups had a high school diploma or less (29% of completers and 27% of non-completers). The distribution of level of education for both groups was very similar, and the amount of education did not add predictive value in any of the models (see Figure 8).

Combining Behavioral and Demographic Information

When the demographic variables were added to the model in the test dataset, the self-regulation variable of time spent working in the course on the first day was statistically significant ($p < .05$). ID verification was also statistically significant ($p < .001$) along with being over age 65 ($p < .01$). Region and gender were not statistically significant; however, they contributed to the overall strength of the model by increasing the prediction accuracy in both the training and the test datasets. The AUC for the logistic regression model that combined behavioral and demographic data was 0.84 and the McFadden's pseudo R^2 was 0.281. This model produced 849 correct predictions with 130

incorrect predictions. It was correct 78% of the time predicting who would complete the course, and it was correct 87% of the time predicting who would not complete and had an average correct prediction rate of 82%.

Overall, adding the variable of ID verification, and the demographic variables of region, age, and gender into the logistic regression models strengthened the predictive power of the models. The AUC increased from 0.68 to 0.84 showing that including demographic variables contributed important information that reflected variation among course participants that was not captured by the behavioral data in the daily activity logs alone (Table 13). Because more variance was captured in the model, the prediction power of the model was also increased from 71% correct predictions to 82% correct predictions. When combined with the demographic variables, the behavioral variable that contributed most strongly to the predictive power of the model was the amount of time students spent in the course on the first day (Table 13). Including the amount of the mathematical skills learned on the first day did not add to the accuracy of the predictions and was not statistically significant and was therefore dropped from the model that combined the demographic data with the daily activity log data.

CHAPTER 6

Conclusion

In addition to looking for variables in student behavior that could point to self-regulation, this study was also focused on early prediction. Reasonably accurate early predictions in any course are valuable because of the prospect of interventions that can be implemented at a point when they might have the greatest impact. This strategy of researching prediction with an eye to intervention is a strategy that is already being deployed for session-based MOOCs (He, Bailey, Rubinstein, & Zhang, 2015).

Pure Accuracy Versus Informative Accuracy

The key goal of this research was to provide insight into what makes a successful completer in a self-paced mathematics MOOC. Because this goal was first and foremost, pure accuracy of prediction was not enough. By merely predicting that all student would not complete, a model could be created that would have an overall correct prediction rate of 96%, but this would provide no information about what it takes to be a completer because the model would be wrong 100% of the time for the all of the completers. By creating a model that has an 80% overall correct prediction rate for both completers and non-completers, key insights into what it takes to be a completer in this environment were illuminated.

The Most Important Predictive Variables

In the end, the most predictive model included only six variables: the score on the pretest, how much time the student spent working on the first day, whether the student was ID verified, the student's geographic region, the student's age, and the student's gender.

How much time a student spent working in the course on the first day was the key behavioral predictor of who would continue working in the course to the end. This supports other research focused on MOOC completion and attrition. In *Learner's Strategies in MOOCs,* Veletsianos, Reich, & Pasquinini found that a key strategy of successful learners is their ability to carve out blocks of time from their busy lives to invest in their educational future (2016). Researchers who have surveyed students that drop out of MOOC courses have reported that "lack of time" is the most commonly cited reason by students regarding why they dropped out (Xiong et al., 2015; Khalil, 2014; Thille et al., 2014; Belanger & Thorton, 2013).

The Importance of Self-Regulation in Self-Paced MOOCs

Because self-paced MOOCs lack the structure of a course schedule with fixed beginning and end dates and a weekly framework to keep students on track, it was theorized by this researcher that self-regulation may play an even greater role in a self-paced MOOC than in session-based MOOCs. This theory motivated the examination of variables in daily activity logs that could be linked to self-regulation and produce predictive results.

Considering the diversity of students working in this college algebra MOOC combined with the structure of a self-paced course, a reasonable explanation for students choosing to work longer in the course on the first day is self-regulation on the part of these learners. When all the days in the course were taken into consideration, students who completed, on average, spent about twice as much time working in the course each day they were active over non-completers even when split by pretest quartiles (Table 4). Self-regulation has been closely linked with effective time management (Zimmerman,

2008; Winne & Hadwin, 1998). For example, in their COPES typology, Winne and Hadwin list time as one of the primary task conditions that must be constantly updated as part of the planning, metacognitive monitoring, and metacognitive evaluating that go into an academic studying task (Winne & Hadwin, 1998). This self-regulation can be either productive, with metacognitive evaluations triggering a greater investment of time and effort, or counter-productive as in this example of an IF-THEN statement provided by Winne and Hadwin:

1. IF time and effort spent on target and
   IF judgement of learning is below standard,
   THEN attribute the negative difference to high task difficulty.
2. IF task difficulty is high,
   THEN quit the task.

Either way, it appears that the amount of time a student is willing to spend on the first day of an online mathematics MOOC may yield insights into positive or negative self-regulation at work within students.

ID verification was also a strong indicator of completion that could be linked to self-regulation. The odds of completing the course increased by a factor of 1.24 ($p <$ .001) when a student opted for ID verification (Table 13). There is reason to believe that this is an indication of self-regulation related to goal setting since there is no other benefit conveyed by ID verification than the option to either receive a certificate of completion or college credit at the end of the course. According to Schunk and Zimmerman, motivation and self-regulation are closely related, and motivation is influenced by setting effective goals. Because ID verification is a self-set goal that is specific to the learning task, it can be expected to have a positive effect on self-regulation influencing choice and

attention toward goal-relevant tasks (working problems within ALEKS), increasing effort, and sustaining persistence toward the goal (2012).

The Importance of MOOCs and Teaching Mathematics at Scale

The significance of studying signal detection of student completion on the first day of a college algebra course is centered on the prospect that providing early detection and intervention to help students is achievable. This is important for several reasons. Knowledge and achievement in mathematics are tied to socioeconomic status not only at the level of the individual but also on a global level (Jurdak, 2014). In the sample studied here, students working within this mathematics course came from 142 countries around the world, and over half the students engaged in this course during the eight month period studied were from outside North America (Table 2). Teaching mathematics at scale successfully is not only going to involve technological innovations in software and artificial intelligence that have been pioneered by MOOC platforms, but will also involve a deep understanding of what students need to bring to the course in terms of self-regulation. Knowledge of demographic variables that can function as barriers to success also needs to be understood if a diverse population of students is going to be served.

Limitations of This Study

There are several limitations connected with this study. The first and most important limitation is that this is a case study of one self-paced mathematics MOOC offered by one university. Self-paced MOOCs is a new area of research that has not been thoroughly studied. Even though single MOOCs afford large samples sizes compared to many other types of educational research, more research on self-paced MOOCs is needed before results can be generalized.

A second limitation related to the first one is the fact that only one mathematics intelligent tutoring system (ITS) was used in this college algebra course. Mathematics ITSs are among the oldest users of artificial intelligence in education (Carnegie Mellon University, 2015). Although, automated mathematics instruction is constantly improving, the content, delivery, and quality of instruction of these systems directly affect student outcomes. Comparing how different ITSs work in the MOOC context of is also important in the progress toward generalizable results.

A third limitation is only one statistical technique was used in the study of this dataset. Logistic regression has been demonstrated to be a robust approach that is especially helpful in handling the unbalanced samples seen in MOOCs (Lauría, Presutti, Guarino, & Sokoloff, 2017). However, this is only one approach and many other machine learning and statistical techniques have been effectively applied to MOOCs and could be useful in exploring this type of data. In addition, interactions were not explored in this study and could yield important information especially when examined in relation to the demographic variables and the behavioral variables.

A fourth limitation is that only data that was readily obtainable was used in this study. The behavioral data was derived from daily activity logs provided by McGraw Hill to this researcher. Although these logs allow for examination on a day-to-day basis, the granularity of this data is limited. It can be assumed that clickstream data would provide additional insights that are not available through daily activity logs. An example of this is the time variable used in this study. While there is a high degree of confidence that this variable is fairly accurate based on its predictive power and statistical significance in the predictive models as well as informal focus groups that were conducted after the

conclusion of this study, webpage timeout issues connected to this variable serve as confounds that could be resolved by having access to more granular data as other research has shown relating clickstream and time data (Douglas & Alemanne, 2007).

A final limitation is that this study only looked at early predictions based on data that was gathered on the first day the student was in the course. Predictions gathered from the data of students as they spend more time in the course could be even more accurate and could still generate early predictions.

Directions for Future Research

Interventions Aimed at Users Who Choose ID Verification. This research represents a limited view into how students work within a self-paced mathematics MOOC; however, the findings can be applied by researchers and practitioners seeking to improve student success. One of the challenges of defining "success" in the MOOC context is the diversity of goals and personal objectives that students bring to MOOCs when they begin a course. Understanding this diversity has been the subject of much of the research and discussion surrounding MOOCs since their inception (Clark, 2016; Ho et al., 2014). For example, many students enroll in MOOCs, but completion is not one of their reasons for participating in the first place (Koller, Ng, Do, & Chen, 2013). So, it is important for practitioners and researchers to understand what constitutes success in the eyes of the MOOC participants themselves to assist them in achieving their own goals and objectives.

One of the most concrete indications of a MOOC participant's intention to complete is ID verification. While most MOOCs allow students to participate for free (at least on a limited level), ID verification involves a monetary fee. In addition, ID

verification adds little if any value to the course for students who do not complete. Since the purpose of ID verification is to be eligible to receive a certificate of completion or college credit as a result of meeting all the requirements of the course, it clearly signals on behalf of the student a desire to finish. Targeting interventions aimed at raising completion rates for students who opt for ID verification not only aligns the goals of the student and the researcher, but it also provides a sandbox for testing interventions that may increase completion rates for other students who would like to finish the course but have not made that goal explicit by opting for ID verification.

While students who do choose ID verification could be targeted for interventions in a self-paced mathematics MOOC without referring to research similar to what has been conducted in this study, the findings of this study provide context and targets that could make these interventions more successful. For example, in this sample, although ID verified students had a higher completion rate (16%) than the sample as whole (4%), the remaining 84% of students who had signaled an intention to complete the course through signing up for ID verification failed to achieve their goal. This research shows that students who complete this course on average work about twice as long on the first day as those who do not. Students who complete and are ID verified also work longer (2.68 hours on the first day) compared to those who are ID verified and who do not complete (1.50 hours on the first day). In addition, ID verified students follow the same pattern as the overall sample in number of mathematics skills learned on the first day: 8.41 skills learned on average for completers versus 4.88 skills on average learned by non-completers. Research into the self-regulation of ID verified completers could lead to successful interventions for the ID verified non-completers who began the course with

the intention to gain credit or a certificate of completion and help these students to meet the goals they have set for themselves.

While the emphasis of this research has been examining the accuracies of the models developed, even the inaccuracies of the models can shed light on ways to increase student success. For example, false positives in the model reflect students who are behaving like completers on the first day but fail to complete the course. This may be another subpopulation in addition to students who have opted for ID verification that may be subjects of further productive research into self-regulation (especially persistence) and possible intervention.

Deeper Research into Self-Regulation. While MOOC user behaviors such as time spent working in the course or ID verification seem to point to self-regulation—especially in a course that is self-paced—the types of self-regulation that are going on to allow these students to complete the course are not clear from purely computer generated data. Investigation into what self-regulation strategies completers are using to be successful in the course has the potential of being a productive area of research that could directly benefit other students who are seeking to complete this type of course but are unsuccessful because they lack the self-regulation of the completers. Qualitative research in this area could be especially useful in terms of interviews, focus groups, or surveys that have the potential of uncovering the internal self-regulation that make students successful when they are working in an online environment without much of the scaffolding that is offered in a face-to-face college or university environment. Other characteristics of students could also be investigated that are closely aligned with self-regulation such as grit and motivation.

Stereotype and Social-Identity Threat. This study could also provide a blueprint for investigations into interventions focused on demographic sub-populations in the course. This study shows that in this self-paced mathematics MOOC, age, gender, and country of origin are more predictive of completion than educational background. Research and interventions aimed at addressing various types of stereotype threat or social-identity threat may be effective in increasing success for students who are being held back by a conscious or sub-conscious sense of lack of welcome or self-confidence.

In the second set logistic regression models, the addition of demographic information increased the number of correct predictions by 11%. Although predictions in self-paced MOOCs like this one have not been well researched, the importance of demographics in influencing course completion in session-based MOOCs has been examined by René F. Kizilcec at Stanford University (2015). The results of the logistic models used in this research show that demographics can play a significant role in self-paced MOOCs as well. In recent research by Kizilcec, his team is showing that targeting sub-populations in MOOCs for intervention can increase completion rate for those whose success in MOOCs is threatened by various types of social-identity threat (Kizilcec, Saltarelli, Reich, & Cohen, 2017).

Female users of the course could also benefit from research into interventions. Although they complete at much lower rates than male students (see Figure 6) their behavior in course does not predict non-completion as demonstrated in all the logistical regressions where gender was not statistically significant as predictor in the models (Table 13). This could point to other factors in play. Stereotype threat among females in mathematics courses is well established in research (e.g. Inzlicht & Ben-Zeev, 2000) and

effective interventions have been implemented to combat stereotype threat (Martens, Johns, Greenberg, & Schimel, 2006). Further research into whether these strategies could be employed at scale in a self-paced mathematics MOOC could result in widespread benefits to females who struggle in math.

Fixed Versus Growth Mindset. One demographic variable that was not predictive of completion in this self-paced mathematics MOOC was educational background (Figure 8). In fact, the distribution of the highest level of education completed was similar between completers and non-completers. Slightly less than a third of both completers (29%) and non-completers (27%) received a high school diploma or less as their highest formal educational achievement. That this is true of completers seems especially significant, as only 4% of everyone to finished the pretest of this self-paced college algebra course went on to complete it. What is positive about this insight is that general educational knowledge in this case does not predict how well a student can do in this course. Sal Kahn, who was part of the inspiration of the first xMOOCs that eventually became Coursera has been very active in promoting the research of Carol Dweck of Stanford University on "growth mindset" (Khan Academy, 2014). This research combines neurobiological research into the ability of the brain to adapt to new challenges and create new connections at the neural level with the importance of encouraging self-regulatory strategies in students rather than emphasizing innate abilities (Myers, Wang, Black, Bugescu, & Hoeft, 2016). Sal Khan and Carol Dweck have pioneered research into messaging at scale that can encourage growth mindset in students who use Khan Academy. Research into whether these messaging strategies work in

mathematics MOOC environments similar to the one used in this study could increase the generalizability of this research.

On the first day that students start working in a self-paced mathematics college algebra MOOC, self-regulatory strategies and demographic conditions are already in play to such a significant degree that it is possible to predict with just first day data who will complete and who will not complete the course with 80% accuracy. However, this does not mean that this research must result in a deterministic attitude toward students. Rather knowing the factors that are the most important elements of this prediction, how a student performed on the pretest, whether the student chose ID verification, the time spent by the student working in the course on the first day, the number of math skills completed on the first day, the age of the student, the gender of the student and the geographical location of the student in the world can help us better understand what can be done to improve mathematics instruction at scale and bring the benefits of mathematics education to an increasingly larger audience.

References

Adamopoulos, P. (2013). *Thirty Fourth International Conference on Information Systems* (pp. 1-21). Milan.

Alario-Hoyos, C., Muñoz-Merino, P. J., Pérez-Sanagustín, M., Kloos, C. D., & Parada, H. A., G. (2016). Who are the top contributors in a MOOC? Relating participants' performance and contributions. *Journal of Computer Assisted Learning, 32*(3), 232-243. doi:10.1111/jcal.12127

Al-Freih, M., Dabbagh, N., & Bannan, B. (2015). Increasing learners' retention and persistence in MOOCs: Designed based research plan. *7th Annual Conference on Higher Education Pedagogy*.

Allen, I. E., & Seaman, J. (2016). *Online report card: Tracking online education in the United States* (Publication). Babson Survey Research Group and Quahog Research Group, LLC.

Allione, G., & Stein, R. M. (2016). Mass attrition: An analysis of drop out from principles of microeconomics MOOC. *The Journal of Economic Education, 47*(2), 174-186. doi:10.1080/00220485.2016.1146096

Amnueypornsakul, B., Bhat, S., & Chinprutthiwong, P. (2014). Predicting attrition along the way: The UIUC model. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 55-59). Doha, Qatar: Association for Computational Linguistics.

Anderson, T., & Dron, J. (2016). The future of e-learning. In C. Haythornthwaite, R. Andrews, J. Fransman, & E. M. Meyers (Eds.), *The SAGE handbook of e-learning research*. Thousand Oaks, CA: SAGE Publications. doi:10.4135/9781473955011

Aparicio, M., Bacao, F., & Oliveira, T. (2017). Grit in the path to e-learning success. *Computers in Human Behavior, 66*, 388-399. doi:10.1016/j.chb.2016.10.009

Armellinini, A., & Rodriguez, B. C. (2016). Are Massive Open Online Courses (MOOCs) pedagogically innovative? *Journal of Interactive Online Learning, 14*(1), 17-28.

Ashenafi, M. M., Ronchetti, M., & Riccardi, G. (2016). Predicting student progress from peer-assessment data. In *Proceedings of the 9th international conference on educational data mining* (pp. 270-275). Boston, MA: International Educational Data Mining Society (IEDMS).

Bacon, L., MacKinnon, L., Anderson, M., Hansson, B., Fox, A., Cecowski, M., . . . Stamatis, D. (2015). Addressing retention and completion in MOOCs - a student-centric design approach. In *E-Learn* (pp. 53-63). Kona, HI.

Baggaley, J. (2013). *Distance Education, 34*(3), 368-378. doi:10.1080/01587919.2013.835768

Baker, R. S., Clarke-Midura, J., & Ocumpaugh, J. (2016). Towards general models of effective science inquiry in virtual performance assessments. *Journal of Computer Assisted Learning, 32*(3), 267-280. doi:10.1111/jcal.12128

Balakrishnan, G. (2013). *Predicting student retention in Massive Open Online Courses using hidden Markov models* (pp. 1-13, Tech.). Berkeley, CA: University of California at Berkeley.

Barber, R., & Sharkey, M. (2012). Course correction: Using analytics to predict course success. In *LAK'12* (pp. 259-262.). ACM.

Belanger, Y., & Thornton, J. (2013). *Bioelectricity: A quantitative approach: Duke University's first MOOC* (pp. 1-21, Tech.). Durham, NC: Duke Center for Instructional Technology.

Belleflamme, P., & Jacqmin, J. (2014). *An economic appraisal of MOOC platforms: Business models and impacts on higher education* (pp. 1-24, Discussion Paper). Louvain-la-Neuve: Center for Operations Research and Econometrics.

Brasher, A., Weller, M., & McAndrew, P. (2016). How to design for persistence and retention in MOOCs? In *EADTU* (p. 2). Rome, Italy: EADTU.

Breazeal, C., Morris, R., Gottwald, S., Galyean, T., & Wolf, M. (2016). Mobile devices for early literacy intervention and research with global reach. In *Proceedings of the third ACM conference on learning @ scale* (pp. 11-20). ACM. doi:10.1145/2876034.2876046

Breslow, L. (2016). MOOC research: Some of what we know and avenues for the future. Retrieved from http://www.portlandpresspublishing.com/sites/default/files/Editorial/Wenner/PPL_Wenner_Ch05.pdf

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom research into edX's first MOOC. *Research & Practice in Assessment, 8*, 13-25.

Bruff, D., Fisher, D., McEwen, K., & Smith, B. (2013). Wrapping a MOOC: Student perceptions of an experiment in blended learning. *Journal of Online Learning and Teaching, 9*(2), 187.

Carnegie Mellon University. (2015). Timeline of cognitive tutor history. Retrieved April 8, 2017, from http://ctat.pact.cs.edu/index.php?id=timeline

Caulfield, M. (2013). XMOOC communities should learn from cMOOCs. Retrieved from http://www.educause.edu/blogs/mcaulfield/xmooc-communities-should-learn-cmoocs

Chacon, F., Spicer, D., & Valbuena, A. (2012). *Analytics in support of student retention and success* (pp. 1-9, Research Bulletin). Louisville, CO: EDUCAUSE: Center for Applied Research.

Chafkin, M. (2013, November 14). Udacity's Sebastian Thrun, Godfather of free online education, changes course. *Fast Company*.

Chaplot, D. S., Rhim, E., & Kim, J. (2015). Predicting student attrition in moocs using sentiment analysis and neural networks. *Proceedings of AIED 2015 fourth workshop on intelligent support for learning in groups*.

Chaplot, D. S., Rhim, E., & Kim, J. (2015). SAP: Student attrition predictor. In *Proceedings of the 8th international conference on educational data mining*. Boston, MA: International Educational Data Mining Society (IEDMS).

Chen, G., Davis, D., Hauff, C., & Houben, G. (2016). Learning transfer: Does it take place in MOOCs? An investigation into the uptake of functional programming in practice. In *L@S* (pp. 409-418). Edinburgh: ACM. doi:10.1145/2876034.2876035

Christensen, G., Steinmetz, A., Alcorn, B., Bennett, A., Woods, D., & Emanuel, E. J. (2013). *The MOOC phenomenon: Who takes Massive Open Online Courses and why?* (Working paper). Available at SSRN 2350964.

Clark, D. (2016, April 11). MOOCs: Course completion is the wrong measure of course success (and a better way has already been suggested). Retrieved November 5, 2016, from https://www.class-central.com/report/moocs-course-completion-wrong-measure/

Clark, R. S. (2016). *Grit within the context of career success: A mixed methods study* (Doctoral dissertation, University of Cincinnati, 2016) (pp. 1-236). Cincinnati, OH: University of Cincinnati.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*(368), 829-836.

Clow, D. (2013). MOOCs and the funnel of participation. In *LAK '13: Proceedings of the third international conference on learning analytics and knowledge* (pp. 185-189). ACM.

Creative Commons. (2016). About the licenses: What our licenses do. Retrieved December 3, 2016, from https://creativecommons.org/licenses/

Cupitt, C., & Golshan, N. (2014). *Participation in higher education online: Demographics, motivators, and grit* (pp. 1-10, Rep.). Perth, Austrailia: National Centre for Student Equity in Higher Education: Curtin University.

Danver, S. L. (Ed.). (2016). OpenCourseWare. In *The SAGE encyclopedia of online education* (p. 886). Thousand Oaks, CA: SAGE Publications.

Danver, S. L. (Ed.). (2016). *The SAGE encyclopedia of online education*. Thousand Oaks, CA: SAGE Publications.

De Barba, P. G., Kennedy, G. E., & Ainley, M. D. (2016). The role of students' motivation and participation in predicting performance in a MOOC. *Journal of Computer Assisted Learning, 32*(3), 218-231. doi:10.1111/jcal.12130

De Barba, P. G., Kennedy, G. E., & Ainley, M. D. (2016). The role of students' motivation and participation in predicting performance in a MOOC. *Journal of Computer Assisted Learning, 32*(3), 218-231. doi:10.1111/jcal.12130

De Freitas, S. I., Morgan, J., & Gibson, D. (2015). Will MOOCs transform learning and teaching in higher education? Engagement and course retention in online learning provision. *British Journal of Educational Technology, 46*(3), 455-471. doi:10.1111/bjet.12268

DeBoer, J., Ho, A. D., Stump, G. S., & Breslow, L. (2014). Changing "course": Reconceptualizing educational variables for Massive Open Online Courses. *Educational Researcher, 43*(2), 74-84. doi:10.3102/0013189X14523038

Deming, D. (2015, February 6). How do traditional and online learning compare? Retrieved November 2, 2016, from https://www.weforum.org/agenda/2015/02/how-do-traditional-and-online-learning-compare/

DeWaard, I., Abajian, S., Gallagher, M. S., Hogue, R., Keskin, N., Koutropoulos, A., & Rodriguez, O. C. (2011). Using mLearning and MOOCs to understand chaos, emergence, and complexity in education. *The International Review of Research in Open and Distance Learning, 12*(7), 94-115.

Dillahunt, T., Wang, Z., & Teasley, S. D. (2014). Democratizing higher education: Exploring MOOC use among those who cannot afford a formal education. *The International Review of Research in Open and Distance Learning, 15*(5).

Dmoshinskaia, N. (2016). *Dropout prediction in MOOCs: Using sentiment analysis of users' comments to predict engagement* (Doctoral dissertation, University of Twente, 2016) (pp. 1-35). Enschede, Netherlands: University of Twente.

Douglas, I. & Alemanne, N. D. (2007). Monitoring participation in online courses. In *8th international conference on informational technology based higher education and training* (pp. 315-320). Kumamoto, Japan.

Downes, S. (2009). Learning networks and connected knowledge. In H. H. Yang (Ed.), *Collective intelligence and e-Learning 2.0: Implications of web-based communities and networking: Implications of web-based communities and networking* (pp. 1-26). Hershey, PA: IGI Global.

Drachsler, H., & Kalz, M. (2016). The MOOC and learning analytics innovation cycle (MOLAC): A reflective summary of ongoing research and its challenges. *Journal of Computer Assisted Learning, 32*(3), 281-290.

Dronkers, J. (2010). *Quality and inequality of education: Cross-national perspectives*. Dordrecht; New York, Springer.

Duckworth, A. (2016). *Grit: The power of passion and perseverance*. New York, NY: Simon and Schuster.

EDUCAUSE learning initiative. (2012, February). 7 things you should know about the flipped classroom. Retrieved from https://net.educause.edu/ir/library/pdf/eli7081.pdf

edX, Inc. (2017). Educational demographics. Retrieved April 2, 2017, from http://edx.readthedocs.io/projects/edx-insights/en/latest/enrollment/Demographics_Education.html

edX, Inc (2017). Verified certificates. Retrieved April 3, 2017, from https://www.edx.org/verified-certificate

Eisenstein, E. L. (1983). The printing revolution in early modern Europe. Cambridge, UK: Cambridge University Press. Duckworth, A. (2016).

Episode1: George Siemens [Interview by D. Bruff, Transcript]. (2016, August 1). In *"Leading Lines" podcast*. Nashville, TN: Vanderbilt Center for Teaching.

Falmagne, J., Koppen, M., Villano, M., Doignon, J., & Johannesen, L. (1990). Introduction to knowledge spaces: How to build, test, and search them. *Psychological Reviw, 97*(2), 201-224.

Ferguson, R., & Clow, D. (2015). Examining engagement: Analyzing learner subpopulations in massive open online courses (MOOCs). In *Proceedings of the 5th International Learning Analytics and Knowledge Conference.* (pp. 16-20). Poughkeepsie, NY. ACM. doi:10.1145/2723576.2723606

Ferguson, R., Coughlan, T., & Herodotou, C. (2016). *MOOCs: What The Open University research tells us* (pp. 1-31, Publication). Milton Keynes, UK: Institute of Educational Technology.

Fischer, G. (2014). Beyond hype and underestimation: Identifying research challenges for the future of MOOCs. *Distance Education, 35*(2), 149-158. doi:10.1080/01587919.2014.920752

Fowler, G. A. (2013, October 8). An early report card on Massive Open Online Courses. *The Wall Street Journal*.

Gašević, D., Kovanović, V., Joksimović, S., & Siemens, G. (2014). Where is research on Massive Open Online Courses headed? A data analysis of the MOOC research initiative. *The International Review of Research in Open and Distance Learning, 15*(5), 134-176.

Gelman, B. U., Beckley, C., Johri, A., Domeniconi, C., & Yang, S. (2016). Online urbanism: Interest-based subcultures as drivers of informal learning in an online community. In *Proceedings of the third ACM conference on learning @ scale* (pp. 21-30). Edinburgh: ACM. doi:10.1145/2876034.2876052

Glance, D. G., & Barrett, P. H. (2014). Attrition patterns amongst participant groups in Massive Open Online Courses. In *Rhetoric and Reality: Critical perspectives on educational technology: Proceedings of ascilite2014* (pp. 12-20). Dunedin, NZ.

Goggins, S. P., Galyen, K. D., Petakovic, E., & Laffey, J. M. (2016). Connecting performance to social structure and pedagogy as a pathway to scaling learning analytics in MOOCs: An exploratory study. *Journal of Computer Assisted Learning, 32*(3), 244-266. doi:10.1111/jcal.12129

Goldwasser, M., Mankoff, C., Manturuk, K., Schmid, L., & Whitfield, K. E. (2016). Who is a student: Completion in Coursera courses at Duke University. *Current Issues in Emerging ELearning, 3*(1), 125-138.

Greene, J. A., Oswald, C. A., & Pomerantz, J. (2015). Predictors of retention and achievement in a Massive Open Online Course. *American Educational Research Journal, 52*(5), 925-955. doi:10.3102/0002831215584621

Gross, A. G., & Harmon, J. E. (2016). *The Internet revolution in the sciences and humanities*. Oxford University Press.

Gütl, C., Rizzardini, R. H., Chang, V., & Morales, M. (2014). Attrition in MOOC: Lessons learned from drop-out students. In *Learning technology for education in the cloud: MOOC and big data* (pp. 37-48). Heidelberg: Springer Cham.

Guo, P., Kim, J., & Rubin, R. (2014). *L@S*. Atlanta, GA: ACM. doi:10.1145/2556325.2566239

Halawa, S., Greene, D., & Mitchell, J. (2014). Dropout prediction in MOOCs using learner activity features. *ELearning Papers, 37*, 7-15.

He, J., Bailey, J., Rubinstein, B. I., & Zhang, R. (2015, January). Identifying At-Risk Students in Massive Open Online Courses. In AAAI (pp. 1749-1755).

Herodotou, C., & Mystakidis, S. (2015). Addressing the retention gap in MOOCs: Towards a motivational framework for MOOCs instructional design. *Proceedings of EARLI 16th biennial conference* (pp. 1-3).

Heutte, J., Kaplan, J., Fenouillet, F., Caron, P., & Rosselle, M. (2014). MOOC user persistence: Lessons from French educational policy adoption and deployment of a pilot course. In *Learning technology for education in the cloud: MOOC and big data* (pp. 13-24). Heidelberg: Springer Cham.

Ho, A. D., Chuang, I., Reich, J., Coleman, C., Whitehill, J., Northcutt, C., . . . Peterson, R. (2015). *HarvardX and MITx: Two years of open online courses* (HarvardX Working Paper No. 10). Boston, MA: HarvardX and MITx. doi:10.2139/ssrn.2586847

Hollands, F. M., & Tirthali, D. (2014). Why do institutions offer MOOCs? *Online Learning, 18*(3), 7-25.

Hollands, F. M., & Tirthalia, D. (2015). *MOOCs in higher education: Institutional goals and paths forward*. New York, NY: Palgrave Macmillan. doi:10.1057/9781137527394

Hone, K. S., & El Said, G. R. (2016). Exploring the factors affecting MOOC retention: A survey study. *Computers & Education, 98*, 157-168. doi:10.1016/j.compedu.2016.03.016

Howarth, J. P., D'Alessandro, S., Johnson, L., & White, L. (2016). Learner motivation for MOOC registration and the role of MOOCs as a university 'taster'. *International Journal of Lifelong Education, Feb* (5), 1-12.

Introduction to the SAGE handbook of e-learning research, second edition [Introduction]. (2016). In C. Haythornthwaite, R. Andrews, J. Fransman, & E. M. Meyers (Eds.), *The SAGE handbook of e-learning research*. Thousand Oaks, CA: SAGE Publications. doi:10.4135/9781473955011

Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science, 11*(5), 365-371.

Jiang, S., Schenke, K., Eccles, J. S., Xu, D., & Warschauer, M. (2016). Females' enrollment and completion in science, technology, engineering, and mathematics Massive Open Online Courses. *ArXiv Preprint*. doi:arXiv:1608.05131

Jiang, S., Williams, A. E., Schenke, K., Warschauer, M., & O'Dowd, D. (2014). Predicting MOOC performance with week 1 behavior. In *Proceedings of the 7th international conference on educational data mining* (pp. 273-275). Boston, MA: International Educational Data Mining Society (IEDMS).

Jordan, K. (2014). Initial trends in enrolment and completion of Massive Open Online Courses. *The International Review of Research in Open and Distance Learning, 15*(1), 133-160.

Jordan, K. (2015). Massive Open Online Course completion rates revisited: Assessment, length and attrition. *International Review of Research in Open and Distributed Learning, 16*(3), 341-358.

Jurdak, M. (2014). Socio-economic and cultural mediators of mathematics achievement and between-school equity in mathematics education at the global level. ZDM Mathematics Education, 46, 1025. doi:10.1007/s11858-014-0593-z.

Kelly, A. P. (2014). *Disrupter, distracter, or what?* (pp. 1-46, Rep.). Bellwether Education Partners.

Kena, G., Hussar, W., McFarland, J., De Brey, C., & Musu-Gillete, L. (2016). *The condition of education 2016* (Publication). Washington, DC: National Center for Education Statistics.

Khalil, H. (2014). MOOCs completion rates and possible methods to improve retention: A literature review. In *EdMedia: World conference on educational media and technology* (pp. 1305-1313). Tampere: EdMedia.

Khalil, M., & Ebner, M. (2016). Learning analytics in MOOCs: Can data improve students retention and learning? In *EdMedia: World conference on educational media and technology*. Vancouver.

Kich, M. (2015, December 22). ASU's Global Freshman Academy is a complete bust: Is anyone actually surprised? [Web log post]. Retrieved November 26, 2016, from https://academeblog.org/2015/12/22/asus-global-freshman-academy-is-a-complete-bust-is-anyone-actually-surprised/

Kizilcec, R. F., & Halawa, S. (2015). Attrition and achievement gaps in online learning. In *L@S*. New York, NY: ACM. doi:10.1145/2724660.2724680

Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. Computers & Education, 104, 18-33.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in Massive Open Online Courses. In *Learning analytics & knowledge conference*. New York: ACM.

Kizilcec, R. F., Saltarelli, A. J., Reich, J., & Cohen, G. L. (2017). Closing global achievement gaps in MOOCs. Science, 355(6322), 251-252.

Klobas, J. E. (2014). Measuring the success of scaleable open online courses. *Performance, Measurement, and Metrics, 15*(3), 145-162. doi:10.1108/PMM-10-2014-0036

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 60-65). Doha, Qatar: Association for Computational Linguistics.

Koller, D., Ng, A., Do, C., & Chen, Z. (2013). Retention and intention in Massive Open Online Courses: In depth. *Educause Review,* 1-12.

Konnikova, M. (2014, November 08). Will MOOCs be Flukes? Retrieved from http://www.newyorker.com/science/maria-konnikova/moocs-failure-solutions

Kop, R. (2011). The challenges to connectivist learning on open online networks: Learning experiences during a Massive Open Online Course. *The International Review of Research in Open and Distance Learning, 12*(3), 19-37.

Kop, R., Fournier, H., & Mak, J. S. (2011). A pedagogy of abundance or a pedagogy to support human beings? Participant support on Massive Open Online Courses. *The International Review of Research in Open and Distance Learning, 12*(7), 74-93.

Kovacs, G. (2016). Effects of In-Video Quizzes on MOOC lecture viewing. In *Proceedings of the third ACM conference on learning @ scale* (pp. 31-40). Edinburgh: ACM. doi:10.1145/2876034.2876041

Laurillard, D. (2014). *Anatomy of a MOOC for teacher CPD* (pp. 1-33, Publication). Moscow: UNESCO Institute for IT in Education.

Laurillard, D. (2014, February 16). Five myths about MOOCs. Retrieved from https://www.timeshighereducation.com/comment/opinion/five-myths-about-moocs/2010480.article

Lee, M. J., Kirschner, P. A., & Kester, L. (2016). Learning analytics in massively multi-user virtual environments and courses. *Journal of Computer Assisted Learning, 32*(3), 187-189. doi:10.1111/jcal.12139

Lee, M. M., Reeves, T. C., & Reynolds, T. H. (2015). *MOOCs and open education around the world* (C. J. Bonk, Ed.). New York, NY: Routledge.

Lewin, T. (2012, March 4). Instruction for masses knocks down campus walls. *The New York Times*.

Lewin, T. (2013, February 20). Universities abroad join partnerships on the web. Retrieved from http://www.nytimes.com/2013/02/21/education/universities-abroad-join-mooc-course-projects.html

Lewin, T. (2015, April 22). Promising full college credit, Arizona State University offers online freshman program. Retrieved from http://www.nytimes.com/2015/04/23/us/arizona-state-university-to-offer-online-freshman-academy.html

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

Literat, I. (2015). Implications of Massive Open Online Courses for higher education: Mitigating or reifying educational inequities? *Higher Education Research & Devlopment,* 1-15. doi:10.1080/07294360.2015.1024624

Liyanagunawardena, T., Lundqvist, K., Parslow, P., & Williams, S. (2014, September 22). *MOOCs and retention: Does it really matter?* Lecture presented in University of Reading, Reading, UK.

Mackness, J., Mak, S. F., & Williams, R. (2010). The ideals and reality of participating in a MOOC. *Proceedings of the 7th International Conference on Networked Learning* (pp. 266-274).

Mak, S. F., Williams, R., & Mackness, J. (2010). Blogs and forums as communication and learning tools in a MOOC. *Proceedings of the 7th International Conference on Networked Learning* (pp. 275-284).

Martin, F. G. (2012). Will Massive Open Online Courses change how we teach? *Communications of the ACM, 55*(8), 26-28. doi:10.1145/2240236.2240246

Martins, A., Johns, M., Greenberg, J., & Schimel, J. (2006). Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Social Psychology, 42*(2), 236-243.

Massive study on MOOCs: Harvard, MIT report provides new insights on an evolving space. (2015, April 1). *Harvard Gazette*, 1-3.

Mathewson, T. G. (2015, December 22). Under 1% of Global Freshman Academy students eligible for ASU credit. Retrieved November 28, 2016, from http://www.educationdive.com/news/under-1-of-global-freshman-academy-students-eligible-for-asu-credit/411241/

Mazoue, J. (2013, October 7). Five myths about MOOCs. Retrieved from http://er.educause.edu/articles/2013/10/five-myths-about-moocs

McAuley, A., Stewart, B., Siemens, G., & Cormier, D. (2010). *The MOOC Model for Digital Practice* (p. 1, Issue brief). Charlottetown: University of Prince Edward Island.

McGraw Hill Education. (2016). ALEKS. Retrieved from https://www.aleks.com/about_aleks/knowledge_space_theory

McGraw-Hill Education's perspective on adaptive learning [Interview by M. Feldstein]. (2016, October 24). Retrieved November 29, 2016, from (https://www.youtube.com/watch?v=N6aYR0uCe9o)

Mitra, S. (2016). The future of learning. In *Proceedings of the third ACM conference on learning @ scale* (pp. 61-62). Edinburgh: ACM. doi:10.1145/2876034.2876053

MoocGuide. (2016). History of MOOCs. Retrieved from https://moocguide.wikispaces.com/1.+History+of+MOOC%27s

MOOCs on the Move [Interview by Knowledge@Wharton]. (2012, November 17). Retrieved October 28, 2016, from http://knowledge.wharton.upenn.edu/article/moocs-on-the-move-how-coursera-is-disrupting-the-traditional-classroom/

Moore, C., & Greenland, S. J. (2016). Insights for future assessment policies and procedures planning and design by open access online higher education providers. In *DEANZ2016* (pp. 112-116). Hamilton, NZ: The University of Waikato.

Myers, C. A., Wang, C., Black, J. M., Bugescu, N., & Hoeft, F. (2016). The matter of motivation: Striatal resting-state connectivity is dissociable between grit and growth mindset. *Social cognitive and affective neuroscience. 11*(10). 1521-1527.

Ng, A. & Widom, J. (2014). Origins of the modern MOOC (xMOOC). In F. M. Hollands & D. Tirthali (Eds.), *MOOCs: Expectations and reality. Full report* (pp. 34-47). New York, NY: Center for Benefit-Cost Studies in Education, Teachers College, Columbia University.

Oakley, B., Poole, D., & Nestor, M. A. (2016). Creating a sticky MOOC. *Online Learning, 20*(1), 1-12.

O'Rourke, E., Peach, E., Dweck, C. S., & Popović, Z. (2016). Brain points: A deeper look at a growth mindset incentive structure for an educational game. In *Proceedings of the third ACM conference on learning @ scale* (pp. 41-50). Edinburgh: ACM. doi:10.1145/2876034.2876040

Papa, R. (2014). *Media rich instruction: Connecting curriculum to all Learners*. Springer.

Pappano, L. (2012, November 2). The Year of the MOOC. *The New York Times*.

Parr, C. (2013, October 17). Mooc creators criticise courses' lack of creativity: Original vision lost in scramble for profit and repackaging of old ideas, say pair. Retrieved from https://www.timeshighereducation.com/news/mooc-creators-criticise-courses-lack-of-creativity/2008180.article

Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaïane, O. R. (2009). Clustering and sequential pattern mining of online collaborative learning data. *Knowledge and Data Engineering, IEEE Transactions on, 21*(6), 759-772.

Perna, L. W., Ruby, A., Boruch, R. F., Wang, N., Scull, J., Ahmad, S., & Evans, C. (2014). Moving through MOOCs: Understanding the progression of users in Massive Open Online Courses. *Educational Researcher, 43*(9), 421-432. doi:10.3102/0013189X14562423

Peterson, C., Matthews, M. D., Kelly, D. R., & Duckworth, A. L. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087-1101. doi:10.1037/0022-3514.92.6.1087

Phan, T., McNeil, S. G., & Robin, B. R. (2016). Students' patterns of engagement and course performance in a Massive Open Online Course. *Computers & Education, 95*, 36-44. doi:10.1016/j.compedu.2015.11.015

Pinkus, A. (2015). Toward a one world schoolhouse: Interview with Sal Kahn. *Independent School*, *Winter*, 42-46.

Poe, M. T. (2011). *A history of communications*. Cambridge: Cambridge University Press.

Porter, E. (2014, June 17). A smart way to skip college in pursuit of a job. *The New York Times*.

Poulin, R., & Straut, T. (2016). *WCET distance education enrollment report 2016* (pp. 1-50, Report). Boulder, CO: WICHE Cooperative for Educational Technologies.

Pursel, B. K., Zhang, L., Jablokow, K. W., Choi, G. W., & Velegol, D. (2016). Understanding MOOC students: Motivations and behaviours indicative of MOOC completion. *Journal of Computer Assisted Learning, 32*(3), 202-217. doi:10.1111/jcal.12131

Rayyan, S., Fredericks, C., Colvin, K. F., Liu, A., Teodorescu, R., Barrantes, A., . . . Pritchard, D. E. (2016). A MOOC based on blended pedagogy. *Journal of Computer Assisted Learning, 32*(3), 187-290. doi:10.1111/jcal.12126

Reich, J. (2015). Rebooting MOOC research. *Science, 347*(6217), 34-35. doi:10.1126/science.1261627

Reich, J., Stewart, B., Mavon, K., & Tingley, D. (2016). *Proceedings of the Third ACM Conference on Learning @ Scale* (pp. 1-10). Edinburgh: ACM. doi:10.1145/2876034.2876045

Rice Online. (2016). Overview of Coursera data and analytics. Retrieved from http://online.rice.edu/media/files/files/3814f9d6/Overview_Of_Coursera_Data_and_Analytics.pdf

Rogers, T., Dawson, S., & Gasevic, D. (2016). Learning analytics and the imperative for theory-driven research. In C. Haythornthwaite, R. Andrews, J. Fransman, & E. M. Meyers (Eds.), *The SAGE handbook of e-learning research*. Thousand Oaks, CA: SAGE Publications. doi:10.4135/9781473955011

Rosé, C. P. (2014). Shared task on prediction of dropout over time in massively open online courses. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 39-41). Doha, Qatar: Association for Computational Linguistics.

Ruiperez-Valiente, J. A., Yudelson, M., Fancsali, S. E., & Berman, S. R. (2016). Using multiple accounts for harvesting solutions in MOOCs. In *Proceedings of the third ACM conference on learning @ scale* (pp. 63-70). Edinburgh: ACM. doi:10.1145/2876034.2876037

Sandeen, C., Jarrat, D., & Parkay, C. (2016, December 3). *To MOOC or not to MOOC: Strategic lessons from the pioneers*. Address presented at American Council on Education.

Scorza, J. (2016, May). The promise of online learning. *HR Magazine*, *61*(4), 25-26.

Schunk, D. H., & Zimmerman, B. J. (2012). *Motivation and self-regulated learning: Theory, research, and applications.* New York: Routledge.

Sears, A. (2016, June 29). Unlocking stackable global credentials. Retrieved November 2, 2016, from http://www.christenseninstitute.org/blog/unlocking-stackable-global-credentials/

Semenova, T. (2016). *Participation in Massive Open Online Courses: The effect of learner motivation and engagement on achievement* (pp. 1-16, Working paper). Moscow: National Research University Higher School of Economics (HSE).

Severance, C. (2012, July 20). IEEE Interview: Teaching the World -- Daphne Koller and Coursera. Retrieved from http://www.dr-chuck.com/csev-blog/2012/07/ieee-interview-teaching-the-world-daphne-koller-and-coursera/

Severance, C. (2012). Teaching the world: Daphne Koller and Coursera. *IEEE Computer Society, 45*(8), 8-9. doi:10.1109/MC.2012.278

Shah, D. (2016, January 12). Six universities in talks to pilot credits for MOOCs. Retrieved from https://www.class-central.com/report/moocs-global-credit-transfer/

Shah, D. (2016, November 16). MOOC Trends in 2016: MOOCs No Longer Massive. Retrieved December 7, 2016, from https://www.class-central.com/report/moocs-no-longer-massive/

Shah, D. (2016, October 31). MOOC course Report: November 2016. Retrieved from https://www.class-central.com/report/mooc-course-report-november-2016/

Sharkey, M., & Sanders, R. (2014). A process for predicting MOOC attrition. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 50-54). Doha, Qatar: Association for Computational Linguistics.

Shea, P. (2014). Introduction. *Online Learning, 18*(3), 1-6.

Siemens, G. (2005). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology and Distance Learning*. Retrieved from http://www.itdl.org/journal/jan_05/article01.htm

Siemens, G., Gašević, D., & Dawson, S. (2015). *Preparing for the digital university: A review of the history and current state of distance, blended, and online learning* (Publication). Athabasca, AB: Athabasca University.

Sinha, T., Jermann, P., Li, N., & Dillenbourg, P. (2014). Your click decides your fate: Inferring information processing and attrition behavior from MOOC video clickstream interactions. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 3-14). Doha, Qatar: Association for Computational Linguistics.

Sinha, T., Li, N., Jermann, P., & Dillenbourg, P. (2014). Capturing "attrition intensifying" structural traits from didactic interaction sequences of MOOC learners. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 42-49). Doha, Qatar: Association for Computational Linguistics.

Skrypnyk, S., De Vries, P., & Hennis, T. (2015). Reconsidering retention in MOOCs: The relevance of formal assessment and pedagogy. In *Proceedings of the European MOOC stakeholder summit* (pp. 166-173). Mons, Belgium.

Smith Jaggars, S. (2015, April 7). No, online classes are not going to help America's poor kids bridge the achievement gap. *The Washington Post*.

Sokolik, M., & Bárrcena, E. (2015). What constitutes an effective language MOOC? In E. Martin-Monje (Ed.), *Language MOOCs: Providing learning, transcending boundries*. Walter de Gruyter GmbH.

Stein, R. M., & Allione, G. (2014). *Mass attrition: An analysis of drop out from a principles of microeconomics MOOC* (pp. 1-20, Working paper No. 14-031). Philadelphia, PA: Penn Institute for Economic Research.

Straumsheim, C. (2014, June 2). Harvard and MIT release MOOC student data set. Retrieved from https://www.insidehighered.com/quicktakes/2014/06/02/harvard-and-mit-release-mooc-student-data-set

Thelin, J. R. (2009). *The race between education and technology*. Cambridge, MA: Harvard University Press.

Thille, C., Schneider, E., Kizilcec, R. F., Piech, C., Halawa, S. A., & Greene, D. K. (2014). The future of data–enriched assessment. *Research & Practice in Assessment, 9*, 5-16.

Tierney, W. G. (1992). An anthropological analysis of student participation in college. *Journal of Higher Education, 63*(6), 603-618.

Toussaint, M., & Brown, V. (2016). Increasing college students' engagement and success rates in developmental mathematics with web-based technologies. In *EdMedia: World conference on educational media and technology* (1st ed., Vol. 2016, pp. 652-658). Waynesville,, NC: AACE.

Townsley, L. (2016). Using a MOOC format for a precalculus course. *PRIMUS, 26*(6), 618-630. doi:10.1080/10511970.2016.11 53544

U. S. Department of Education. (2016). Fast facts: Distance learning. Retrieved November 2, 2016, from https://nces.ed.gov/fastfacts/display.asp?id=80

Usher, E. L., & Pajares, F. (2008). Self-efficacy for self-regulated learning: A validation study. Educational and Psychological Measurement, 68(3), 443-463.

Van der Sluis, F., Ginn, J., & Van der Zee, T. (2016). Explaining student behavior at scale: The influence of video complexity on student dwelling time. In *Proceedings of the third ACM conference on learning @ scale* (pp. 51-60). Edinburgh: ACM. doi:10.1145/2876034.2876051

Vardi, M. Y. (2012). Will MOOCs destroy academia? *Communications of the ACM, 55*(11), 5. doi:10.1145/2366316.2366317

Veletsianos, G., & Shepherdson, P. (2016). A systematic analysis and synthesis of the empirical MOOC literature published in 2013–2015. *International Review of Research in Open and Distributed Learning, 17*(2).

Vygotskiĭ, L. S. (1978). *Mind in society: The development of higher psychological processes.* (M. Cole, Trans.). Cambridge: Harvard University Press.

Veletsianos, G., Reich, J., & Pasquini, L. A. (2016). The life between big data log events: Learners' strategies to overcome challenges in MOOCs. *AERA Open, 2*(3), 1-10. doi:10.1177/2332858416657002

Waks, L. J. (2016). *The evolution and evaluation of Massive Open Online Courses: MOOCs in motion*. Springer.

Walsh, M. L. (2016). Competency-Based Education. In S. L. Danver (Ed.), *The SAGE encyclopedia of online education* (pp. 217-220). Thousand Oaks, CA: SAGE Publications.

Weng, F., & Chen, L. (2016). A nonlinear state space model for identifying at-risk students in open online courses. In *Proceedings of the 9th international conference on educational data mining* (pp. 527-532). Boston, MA: International Educational Data Mining Society (IEDMS).

What's AHEAD: Key trends in higher education. (2014). *Poll #1: Massive Open Online Courses (MOOCs)* (pp. 1-4, Rep.). Philadelphia, PA: Penn AHEAD (Alliance for Higher Education and Democracy).

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. *Metacognition in educational theory and practice*, 93, 27-30.

Whitmer, J., Schiorring, E., & James, P. (2014). Patterns of persistence: What engages students in a remedial English writing MOOC? In *The 4th international learning analytics and knowledge conference (LAK'14)* (pp. 279-280). New York: ACM. doi:10.1145/2567574.2567601

Witten, I. H., & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with Java implementations.* San Francisco, CA: Morgan Kaufmann.

Wladis, C., Wladis, K., & Hachey, A. C. (2014). The Role of Enrollment Choice in Online Education: Course Selection rationale and course difficulty as factors affecting retention. *Online Learning, 18*(3), 29-42.

Wolfe, A. (2015, June 5). Daphne Koller on the future of online education. *The Wall Street Journal*.

Xiong, Y., Li, H., Kornhaber, M. L., Suen, H. K., Pursel, B., & Goins, D. D. (2015). Examining the relations among student motivation, engagement, and retention in a MOOC: A structural equation modeling approach. *Global Education Review, 2*(3), 23-33.

Yeager, C., Hurley-Dasgupta, B., & Bliss, C. A. (2013). CMOOCs and global learning: An authentic alternative. *Journal of Asynchronous Learning Networks, 17*(2), 133-147.

Yeager, C., Hurley-Dasgupta, B., & Bliss, C. A. (2013). CMOOCs and global learning: An authentic alternative. *Journal of Asynchronous Learning Networks, 17*(2), 133-147.

Yuzer, T. V., & Kurubacak, G. (Eds.). (2014). *Handbook of research on emerging priorities and trends in distance education: Communication, pedagogy, and technology*. Hershey, PA: Information Science Reference.

Zheng, S. (2015). Retention in MOOCs: Understanding users' motivations, perceptions and activity trajectories. In *CHI 2015 Crossings, Seoul, Korea* (pp. 247-250). New York: ACM. doi:10.1145/2702613.2702628

Zheng, S., Han, K., Rosson, M. B., & Carroll, J. M. (2016). The role of social media in MOOCs: How to use social media to enhance student retention. In *L@S* (pp. 419-428). New York, NY: ACM. doi:10.1145/2876034.2876047

Zimmerman, B. J., (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal.* 45(1), 166-183.