WHAT MAKES A GREAT MOOC? AN INTERDISCIPLINARY ANALYSIS OF STUDENT RETENTION IN ONLINE COURSES

Completed Research Paper

Panagiotis Adamopoulos
Department of Information, Operations and Management Sciences
Leonard N. Stern School of Business, New York University
44 W 4th St, New York, NY 10012
padamopo@stern.nyu.edu

Abstract

Massive Open Online Courses (MOOCs) have experienced rapid expansion and gained significant popularity among students and educators. Although the broad acceptance of MOOCs, there is still a long way to go in terms of satisfaction of students' needs, as witnessed in the extremely high drop-out rates. Working toward improving MOOCs, we employ the Grounded Theory Method (GTM) in a quantitative study and explore this new phenomenon. In particular, we present a novel analysis using a real-world data set with user-generated online reviews, where we both identify the student, course, platform, and university characteristics that affect student retention and estimate their relative effect. In the conducted analysis, we integrate econometric, text mining, opinion mining, and machine learning techniques, building both explanatory and predictive models, toward a more complete analysis. This study also provides actionable insights for MOOCs and education, in general, and contributes to the related literature discovering new findings.

Keywords: Massive Open Online Courses, MOOCs, online learning, student retention, grounded theory method, econometric analysis, choice models, text mining, opinion mining, predictive modeling, user-generated content, user satisfaction, utility

Introduction

Massive Open Online Courses (MOOCs) are a recent development in the area of e-learning and distant education that gains significant popularity among both students and educators. MOOCs are larger in scale than traditional courses, have no restrictions on individual participation, they are globally distributed across a variety of networks, and aim at revolutionizing the way education happens. Such massive online courses are offered in a wide range of topics, spanning the Humanities, Social Sciences, Mathematics, Engineering, Computer Science, and other disciplines. Besides, dozens of top universities, including many international and Ivy League institutions, are offering a large number of MOOCs. The aforementioned factors, together with the open nature of the courses and the lack of need for physical presence, attract a huge number of students from a wide variety of backgrounds. Interestingly, some prominent MOOCs, such as various classes from Stanford University, edX, Coursera, and Udacity, have attracted tens of thousands of participants. For instance, as of November 2012 more than 1,900,241 students from 196 countries have enrolled in at least one course by Coursera.

Even though MOOCs have been broadly accepted, there is still plenty of room for improvement as far as the actual needs of students are considered. This is evident if we take into consideration that the student retention rates are very low (Clow 2013; Downes 2010; Knowledge@Wharton 2013; Lewin 2013). At the same time, the huge popularity in starting students shows that the time is ripe for MOOCs (Fox and Patterson 2012; Pappano 2012). Nevertheless, it is the duty of the academic community to shed light on the problems of MOOCs, trying to both understand their causes and provide actionable solutions, in order open education to achieve its potential and not fail. Working toward this direction, we explore one of the most important, if not the most important, problems of MOOCs. In this study, we develop a hedonic-like approach to identify important concepts that affect online course retention and estimate their relative importance based on textual reviews submitted by students on special online communities. In particular, we employ the Grounded Theory Method (GTM) (Eisenhardt 1989; Glaser and Strauss 1967) on quantitative data, a less frequently applied paradigm, which provides a flexible way of conducting research that prioritizes exploration of the given phenomenon in a predominantly inductive theory development paradigm (Birks et al. 2013) in Information Systems (IS). Overall, we use radically different approaches and interdisciplinary perspectives to provide a holistic and deep view of a novel phenomenon of scientific and societal interest. Our goal is to understand the behavioral process that leads to the students' choices and, thus, we take a causal perspective since there are factors that collectively determine, or cause, the agents' choices (Train 2003). In addition, we aim at contributing to the related literature and improving course design and, thus, online and distance education. Finally, we hope to inspire future research and open up new streams of scientific inquiry for the IS field.

In this paper, we identify the important determinants that affect retention in online courses and conduct an innovative analysis to study their relative importance. We employ GTM in a quantitative study that integrates econometric, text mining, opinion mining, and predictive modeling techniques toward a more complete analysis of the information captured by user-generated content. Additionally, we provide actionable insights with important implications for both MOOCs and education, in general, and contribute to the related literature on student retention, course design, and educational policy by discovering new findings.

Related Work

While Massive Open Online Courses (MOOCs) are a relatively recent development, the ideas behind MOOCs can be traced as back as the early 1960s, when Fuller (1962) proposed industrial scale educational technology. In particular, MOOCs succeeded the trend of Open Education movement that focuses on how open education tools, resources, and knowledge can improve the quality of education (Iiyoshi and Kumar 2008). The open education tools in combination with the latest technological advancements allow professors and educational institutions to offer MOOCs with massive number of students (Fox and Patterson 2012) and, at the same time, they allow students to overcome geographic and economic barriers (Russell et al. 2013) and pursue their own learning objectives (Armstrong 2012). From the beginning of such developments, many practitioners and researchers, such as Harasim et al. (1995), have put great emphasis on the use of learning networks for knowledge construction, discussing how people learn in

large open networks which offer extensive diversity and connectivity (Mackness et al. 2010). Exploring different dimensions of MOOCs, Cormier and Siemens (2010) and Masters (2011) discussed how the roles of the professor (instructor) have changed, which now include amplifying, curating, aggregating, filtering or selecting, modeling, and staying present, while Xu and Jaggars (2013) examined the extent to which students perform as well online as they do face-to-face, and Mak et al. (2010) studied the importance of blogs and discussion forums as communication and learning tools in a MOOC. In other streams of research, Joglekar et al. (2013) and Singh et al. (2013) proposed methods for automatically providing feedback and evaluating the assignments of students, Piech et al. (2013) developed algorithms for estimating and correcting for grader biases and reliabilities in peer assessment in MOOCs, and Sadigh et al. (2012) discussed automated exercise generation. In addition, regarding the retention problem, Cormier and Siemens (2010) and Russell et al. (2013) asked from where the key value of a course in the educational systems comes and Mackness et al. (2010) posed the question of how to design a course which will provide participants with positive experiences. Toward this direction, Vihavainen et al. (2012) recently established a positive statistical significant correlation between the difficulty of a MOOC and its educational value. Nevertheless, despite the wide scientific interest, much more work should be done to improve MOOCs. In particular, researchers should shed light on the problems of MOOCs and try to solve the problem of student retention as witnessed in the high drop-out rates (Clow 2013; Lewin 2013).

However, the problem of student retention is not pertaining only to MOOCs. In traditional education, the first national retention study in the United States surveyed 25 universities and reported a dropout rate of 45% (MacNeely 1938). Examining this phenomenon, Tinto (1975; 1987) proposed the student integration model to study college dropout and explain the process that motivates individuals to leave colleges and universities before graduating while Bean (1985) proposed the student attrition model to study the conscious, openly discussed student intention to leave an institution, coupled with actual attrition, emphasizing academic, social, and personal outcomes of the selection or socialization of students at an institution. Besides, Avakian et al. (1982) studied race and sex differences in student retention at an urban university, Boudreau and Kromrey (1994) examined the relationship between the completion of a freshman orientation course and retention, whereas other researchers, such as Johnson (1996) and Pesron and Christensen (1996), studied retention for specific groups of students. In addition, Moore and Miller (1996) examined how the use of multimedia affects students retention and learning. Furthermore, the literature also includes descriptions of strategies that specific colleges and universities have found helpful in retaining students (Bedford and Durkee 1989; Hyman 1995). However, it is also evident that the findings reported in the literature are often conflicting (Cabrera et al. 1992). For instance, Glass and Garrett (1995) reported that among community college students retention and grade point average (GPA) are not related to age, gender, race, employment status, college major, or college attended. On the other hand, Murtaugh et al. (1999) found that attrition increases with age and decreases with increasing high school GPA and first-quarter GPA. Statistically significant associations of retention with ethnicity/race, orientation courses, residency, and college at first enrollment were also noted. Considering these reports and the often conflicting findings, one can conclude that student attrition is a continuing problem and not much has been done to effectively improve retention during the last centuries (Glass Jr and Garrett 1995). Additionally, the majority of these studies did not examine any individual course characteristics but focused on identifying the most promising students to recruit; a practice that is very different, in principle, from the massive and open nature of MOOCs.

As far as the methodology of this study is considered, an appropriate method for exploration of the given phenomenon and theory building relevant to the IS discipline is the Grounded Theory Method (GTM) (Eisenhardt 1989; Glaser and Strauss 1967). In this regard, GTM provides a flexible way of conducting research that prioritizes exploration of the given phenomenon in a predominantly inductive theory development paradigm (Birks et al. 2013). GTM enables researchers to provide a rich description of the given phenomenon based on a systematic approach through observations, archival data, interviews, and other sources (Corbin and Strauss 2007). The greatest advantage of GTM is the logic of discovery, rather than that of verification in data analysis, which is essential to the delicate question of theory building in grounded research (Vaast and Walsham 2011). The emergent principle of grounded theory is manifested in the belief that both the outcome (grounded theory) and the research design should be emergent (Corbin and Strauss 2007). Despite the misconceptions about GTM (Suddaby 2006; Urquhart and Fernandez 2013), its roots in quantitative methods make GTM a powerful tool also for quantitative research (Birks et al. 2013) allowing to conduct pioneering research moving between induction and

deduction (Corbin and Strauss 2007) with both flexibility and rigor (Birks et al. 2013). In the new era of wide access to data, researchers can effectively integrate GTM into their work while working within the research paradigm with which they are most comfortable (Birks et al. 2013).

Moreover, this study also draws on a new stream of research that applies text-mining techniques to reviews and provides significant opportunities to explore important sources of information. The most common techniques search for statistical patterns and trends in the text in order to distinguish between positive and negative reviews (Dave et al. 2003) or use classifiers that determine whether a particular feature or concept is implicitly discussed (Ghani et al. 2006). For example, Hu and Liu (2004a; 2004b) summarized all the customer reviews of a product employing association rule mining to find frequent n-grams to be used as candidate features and Lee and Bradlow (2007) presented an automatic procedure for obtaining conjoint attributes and levels that list the explicit pros and cons of a product. In a close stream of research using the sentiment of user-generated content, Pang et al. (2002) studied the rating-inference problem by classifying movies and, then, Das and Chen (2007) examined bulletin boards and showed that the aggregate sentiment predicts the stock index. Moreover, Archak et al. (2007; 2011) demonstrated how textual data can be used for both predicting future changes in sales and learning consumers' relative preferences for different product features. Finally, Pang and Lee (2008) offered a comprehensive survey of the research in the field of sentiment analysis. In general, the employed techniques are similar to those used in conjoint analysis (Green and Srinivasan 1990) and preference measurement (Netzer et al. 2008), which also aim at determining how people value different features in a product or service, but they differ in the source of data used for the analysis.

Recognizing the significance of user-generated content and online reviews, various important aspects have been studied in relevant streams of literature. Research on computer mediated communication suggests that online community members communicate information about product or service evaluations with intent to influence others' purchase decision as well as provide social information about contributing members themselves (Postmes et al. 2001). Search time (Goldsmith and Horowitz 2006) and risk (Hennig-Thurau et al. 2004) reduction have been also identified as motives for using consumer opinion platforms. Also, recent studies show that personal opinions found on internet forums are more likely to be perceived trustworthy, less risky, and more relevant to the consumer than other sources (e.g. official webpages) because users perceive the source to be similar to themselves (Bickart and Schindler 2001). Finally, the hypothesis that user-generated reviews affect user choices has also received strong support in empirical studies (Dhar and Chang 2009; Ghose et al. 2012; Liu 2001; Muchnik et al. 2013).

Data Analysis and Coding of Variables

In this section, we explain the methodological ideas and the research model we employed. Also, we describe the data we collected from various MOOC platforms and CourseTalk.org (2012), a website for online course reviews, and discuss each variable used in the subsequent analysis. In particular, we collected qualitative and quantitative data about 133 courses offered by 30 universities and 6 providers (platforms). The courses were offered by the following providers: Canvas Network, Codecademy, Coursera, edX, Udacity, and Venture Lab, and belong to a wide range of academic disciplines: Business, Computer Science, Engineering, Humanities, Mathematics, and Science.[1] More specifically, in order to study the retention of students in MOOCs, we collected and analyzed 1163 textual reviews submitted online by 842 students that participated in at least one course. In addition, following the GTM guidelines, after actively participating in various MOOCs, observing students, and directly interacting with participants, additional concepts emerged related to *student*, *course*, *university*, and *platform* characteristics. These concepts that affect student retention in MOOCs were identified through several iterations and the corresponding data was collected. Hence, for each course we manually gathered additional data about the university which offers the course, the platform where the course is hosted, the academic discipline(s) of the course, the average difficulty of the course in the posted reviews, whether the course is self-paced or has a calendar-based schedule, the estimated workload in hours/week, the duration of the course in weeks, whether there is automated feedback or peer assessment, if team projects or final exams are required, whether a (paid)

[1] The Science discipline includes courses in the following sub-disciplines: Medicine, Healthcare, Physical Sciences, Biology, and Food & Nutrition. Also, a course may belong to more than one academic discipline. Besides, we do not consider any courses non-affiliated with a formal educational institution, such as those offered by Code School LLC and SkillShare Inc.

textbook is suggested, and whether students who successfully complete the course receive a certificate. Besides, for each university we estimate its relative ranking for the academic discipline of each course offered. In particular, the universities are ranked based on the QS World University Rankings (2012).[2] Finally, for a large number of students, we also collected data about their sex and whether they attend a formal educational institution. Studying the course retention problem, the dependent variable of interest in our analysis is the self-reported progress of each student in each course, and can be naturally ordered as following: Course Not Completed (i.e. dropped), Course Partially Completed (e.g. complete 70% of the course or complete the course without submitting the assignments), and Course Successfully Completed.[3] In the next session, we describe in detail the textual analysis we conducted and the variables that we estimated in the explanatory numerical analysis.

Textual Analysis and Integration of Methods

Using text and opinion mining methods to identify a) the important features that affect student satisfaction and retention of MOOCs and b) the corresponding evaluation of each student about these features, we proceed according to the following steps:

- We mine MOOC features on which students have commented,
- identify opinion sentences in each review and the corresponding opinion words, and
- decide whether each opinion sentence is positive or negative and estimate the corresponding sentiment score for each feature.

To perform these tasks, we made use of both data mining and natural language processing techniques. In order to identify the concepts that characterize a MOOC and concern students, we followed an approach similar to the one proposed by Hu and Liu (2004a; 2004b). First, we found those features on which many people have expressed their opinions. The course features we are interested in are nouns or noun phrases in review sentences and, hence, we used part-of-speech tagging (Manning and Schütze 1999). Some pre-processing of words, which includes removal of stop-words and non-English words, stemming, and fuzzy matching, was also performed. Then, a transaction file was created where each transaction corresponds to the identified nouns and noun phrases included in a sentence. This file was subsequently used in association mining (Margahny and Mitwaly 2005) to find all frequent item-sets (a set of words or phrases that frequently occurs together in some sentences). Those nouns and noun phrases are the features we are interested in. In particular, in order to identify the important features, we used the Apriori algorithm (Margahny and Mitwaly 2005) with 1% minimum support. Further, in order to remove unlikely features, we used both compactness pruning and redundancy pruning (Hu and Liu 2004b).

Then, we implemented two alternative approaches in order to identify the expressed opinion about each feature discussed in the posted reviews. The first approach, an *orientation analysis mechanism*, consists of the extraction of opinion words and the identification of their orientation. Opinion words were extracted using the corresponding frequent features and the semantic orientation of each opinion word was identified with the help of WordNet (Fellbaum 1998). In detail, we identified and extracted the adjectives (opinion words) from those sentences that contained one or more identified features (opinion sentences); we limited the opinion word extraction only to opinion sentences, as we are only interested in users' opinions on these features. Then, to predict both the orientation of the adjectives in the collected reviews and the orientations of the opinion sentences, we selected a set of common adjectives as a seed list and proceeded as in (Hu and Liu 2004a).[4] The alternative approach consists of identifying the sentiment for each feature in each opinion. To perform this task we used a *sentiment analysis mechanism* (AlchemyApi 2012). The two alternative mechanisms for opinion mining produce, in principal, similar

[2] If a university is not included in the QS World University Rankings, then it is relatively ranked as next to the university with the lowest known ranking.

[3] When submitting a review, the students could only select one of the predefined options regarding their progress (i.e. Taking Course Now, Course Not Completed, Course Partially Completed, and Course Successfully Completed). In the conducted analysis, we did not include any reviews from students currently attending the corresponding course.

[4] Opinion word extraction is related to previous work on subjectivity (Bruce and Wiebe 1999) which has established a positive statistical significant correlation between the expression of an opinion and the presence of adjectives (Hu and Liu 2004b). Words that encode a desirable state (e.g. beautiful, awesome) have a positive orientation, while words that represent undesirable states have a negative orientation (e.g. disappointing).

results. Because of space restrictions, we focused on the sentiment analysis mechanism and we mainly used the orientation analysis mechanism to check the robustness of our study.

Additionally, in order to evaluate and verify the results of the aforementioned procedure and also enhance our data with rich interpretative accounts, we appointed two graduate students and in collaboration we manually investigated in total 20% of the textual reviews. Following the GTM guidelines and without having any predefined framework in mind, we conducted a systematic manual exploration of the online reviews from different standpoints and tried to make sense of the user-generated content. Employing an iterative approach, through constant comparison of the course reviews, the already identified features of MOOCs, and the relative relationships of these features, we coded the data identifying the most important factors that affect the course retention decision. The codes were documented using memos taking the form of text narratives and diagrams. At each iteration of this process, we were sampling more reviews identifying potentially new emerging factors that affect a student's decision to drop a course or not, abstracting out the concepts from the codes, and contrasting any new features with the existing ones. In parallel with this passive process of analyzing online user-generated data and in an effort to embrace multiple sources of data, we also employed a more active engagement as researchers. In particular, we enrolled as students and participated in several MOOCs by submitting assignments, engaging in discussions on the bulletin board, and reviewing online material of the courses. Proceeding as previously described, new factors and interactions were identified and integrated with the already identified concepts through constant comparison. This active engagement yielded a richer description and different perspectives of the phenomenon. At the end of each iteration of this process, we conducted a corresponding econometric analysis and built predictive models to assess the explanatory and predictive power of the concepts that had already been identified and evaluate the potential need for further analysis and coding of variables. This process of sampling, iterative manual coding, and constant comparison was repeated until saturation had been reached and no new concepts came up through examining additional reviews or making new observations. During the final step of this process, the emergent concepts were grouped into different categories according to their relationships and differences. Based on this approach, we came up with a joint list that contained the factors our algorithm had revealed; the employed categorization was somehow different, grouping tightly themed topics into one single category (e.g. videos and lectures under course material). Hence, we updated our algorithm to group these specific features together. Additionally, apart from matching the automated features, new features emerged through the manual iterative coding process.

The MOOC features that were identified based on the textual analysis and for which the corresponding opinion of each student (i.e. *student course evaluation*) was quantified using the sentiment analysis mechanism, were grouped into the following concepts (variables): *Assignments*, *Course Material*, *Discussion Forum*, and *Professor*. Moreover, using the aforementioned research methods, additional concepts emerged that are related to *student* (i.e. gender, matriculated student), *course* (i.e. assignments, professor, discussion forum, course material, self-paced, difficulty, workload, weeks, certificate, peer assessment, team projects, suggested textbook, paid textbook, final exam, academic discipline), *university* (i.e. ranking, courses offered), and *platform* (i.e. Canvas Network, Udacity, Venture Lab, Coursera, edX, Codecademy) characteristics; these variables were manually collected. Table 1 summarizes all the variables, which were used in the analysis, and shows the corresponding descriptive statistics computed over all the observations in our data set. Interestingly, Table 1 shows that the average sentiment of the students for the course evaluation variables (i.e. assignments, professor, discussion forum, course material) is positive but of small magnitude. In other words, this initial observation confirms that, despite the broad acceptance of MOOCs among the students and the positive sentiment, there are still opportunities for further enhancements.

Table 1. Variables and Descriptive Statistics

Variable	Description	Mean	Std.Dev.	Min	Max
Course Progress	Whether the student did not complete, partially completed, or successfully completed the course	1.6708	0.5648	0	2
Assignments	The sentiment of the individual review for course assignments	0.1094	0.1367	-1	1
Professor	The sentiment of the individual review for the professor(s)	0.1158	0.1410	-1	1

Discussion Forum	The sentiment of the individual review for the discussion forum	0.1105	0.1396	-1	1
Course Material	The sentiment of the individual review for the course material	0.1154	0.1346	-1	1
Self-paced	Whether the course is self-paced (dummy)	0.0879	0.2821	0	1
Difficulty	The average difficulty of the course	2.8838	0.5387	1	5
Workload	The estimated workload in hours/week	6.0714	2.6795	0	17.5
Weeks	Duration of the course in weeks	7.8968	3.3293	0	15
Certificate	Whether students who successfully complete the course receive a certificate (dummy)	0.7136	0.4522	0	1
Peer Assessments	Whether there is peer assessment in the course (dummy)	0.2218	0.4157	0	1
Team Projects	Whether there is any team project in the course (dummy)	0.0091	0.9474	0	1
Suggested Textbook	Whether there is a suggested textbook (dummy)	0.3353	0.4723	0	1
Suggested Paid Textbook	Whether there is a suggested paid textbook (dummy)	0.1625	0.3691	0	1
Final Exam/Project	Whether there is a final exam or project in the course (dummy)	0.3715	0.4834	0	1
Interested Users	The number of user on the review platform interested in attending the course	24.7933	18.1152	0	74
Course Reviews	The number of reviews posted for the course	45.2630	41.2266	1	126
University Ranking	Relative ranking of the university	12.1786	7.0543	1	30
Courses Offered	The number of courses offered by the university	16.4372	9.4899	1	34
Business & Management	Whether the course belongs to the Business & Management discipline (dummy)	0.1470	0.3543	0	1
Computer Science	Whether the course belongs to the Computer Science discipline (dummy)	0.3972	0.4895	0	1
Engineering	Whether the course belongs to the Engineering discipline (dummy)	0.0353	0.1845	0	1
Humanities	Whether the course belongs to the Humanities discipline (dummy)	0.3525	0.4780	0	1
Mathematics	Whether the course belongs to the Mathematics discipline (dummy)	0.0404	0.1970	0	1
Science	Whether the course belongs to the Science discipline (dummy)	0.1273	0.3334	0	1
Canvas Network	Whether the course is offered on the Canvas Network platform (dummy)	0.0009	0.0293	0	1
Udacity	Whether the course is offered on the Udacity platform (dummy)	0.0598	0.2315	0	1
Venture Lab	Whether the course is offered on the Venture Lab platform (dummy)	0.0069	0.0827	0	1
Coursera	Whether the course is offered on the Coursera platform (dummy)	0.8783	0.3305	0	1
edX	Whether the course is offered on the edX platform (dummy)	0.0559	0.2298	0	1
Codecademy	Whether the course is offered on the Codecademy platform (dummy)	0.0026	0.0507	0	1
Student Gender	Whether the student is male (dummy)	0.5994	0.4903	0	1
Matriculated Student	Whether the student attends a formal educational institution (dummy)	0.2415	0.4285	0	1

Figure 1 summarizes the proposed research model illustrating the categories that emerged in our analysis. In particular, following the GTM guidelines and triangulating different data sources, we suggest that the

decision of a student to drop a course is affected by the student course evaluation (e.g. evaluation of the professor from individual student), the course characteristics (e.g. difficulty, academic discipline), university characteristics (e.g. university ranking), platform characteristics (e.g. platform usability), and student characteristics (e.g. gender).

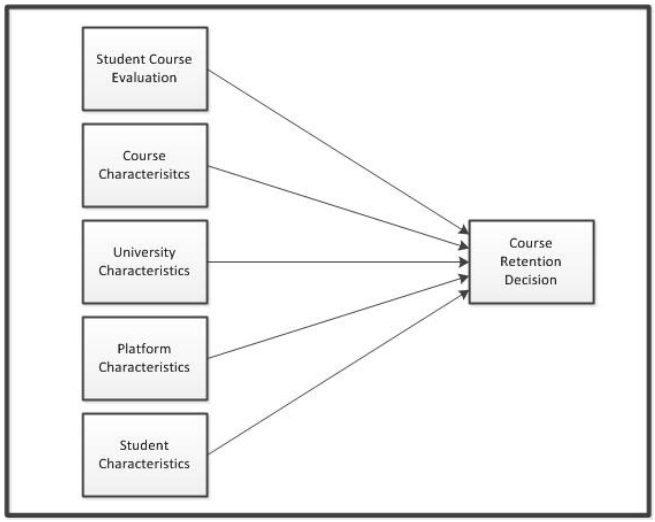

Figure 1. A Model of Online Course Retention.

Explanatory Econometric Analysis

In this section, we present the results of our *explanatory* econometric analysis, which examines the importance of each factor in student retention in MOOCs (in the next session, we will describe our predictive model, based on machine learning techniques). Through our analysis, we employ ordered choice models in a hedonic-like framework, where we use text and opinion mining to model qualitative opinions in quantitative evaluation scores, and we aim to provide a better understanding of how students' decisions are affected by the different factors.

In prior literature on traditional education, various methods have been employed to study student retention including simple cross-tabulation (Avakian et al. 1982), two-sample comparisons (Naretto 1995), linear regression, logistic regression, probit analysis (Dey and Astin 1993), and Markov processes (Heiberger 1993). In some prior works, the research methods applied to the problem of student retention have been criticized for issues related to their consistency, efficacy, and suitability (Dey 1997; Murtaugh et al. 1999; Nandeshwar et al. 2011); this might be the explanation for the conflicting findings reported in literature. For instance, the use of linear regression in our study, because of the unboundness problem and the assumption of a continuous dependent variable instead of a discrete non-numeric outcome, could lead to erroneous findings (Greene 2003). In our analysis, we take advantage of the nature of our data and we model course retention as a countable, discrete, ordered choice by giving meaningful numeric values to student choices and estimate the corresponding probabilities; each student chooses among not to complete a course, to partially complete it (e.g. complete 70% of the course or complete the course without submitting the assignments), or to successfully complete it. Our goal is to understand the behavioral process that leads to the agent's choice and, thus, we take a causal perspective since there are factors that collectively determine, or cause, the agent's choice (Train 2003). Since there are also factors not observed by the researcher, such as unobservable abilities of students, the agent's choice is not deterministic and cannot be predicted exactly. Instead, the probability of any particular outcome is derived in a consistent, invariant, and efficient (given the model) way (Greene 2003). Based on the above discussion and since an underlying rational decision process takes place when a student decides to complete a course or not, we use ordered choice models to efficiently model student retention. The model platform is an underlying random utility model or latent regression model (McFadden 1974),

$$y_i^* = \boldsymbol{\beta}' \boldsymbol{x_i} + \varepsilon_i, i = 1, \dots, n,$$

in which the continuous latent utility or measure, y_i^*, is observed in discrete form through a censoring

mechanism:

$$\begin{aligned} y_i = 0 \text{ if } & -\infty < y_i^* \leq \mu_0 \quad \text{(course not completed)}, \\ = 1 \text{ if } & \mu_0 < y_i^* \leq \mu_1 \quad \text{(course partially completed)}, \\ = 2 \text{ if } & \mu_1 < y_i^* < +\infty \text{ (course successfully completed)}. \end{aligned}$$

The vector $\boldsymbol{x_i}$ is the set of the covariates of our model and $\boldsymbol{\beta}$ is the vector of parameters to be estimated. In order to identify the factors that affect the decision of the users and examine their relative importance, we fit different specifications for the course retention choice model that correspond to iterations described in the previous section. In particular, Model 1 studies the effect of specific *student course evaluation* (i.e. sentiment of the student with respect to assignments, professor, discussion forum, and course material), *course* (i.e. difficulty, workload, weeks, certificate), and *university* (i.e. university ranking) *characteristics*. Then, Model 2 uses the additional information of academic discipline (course characteristic) and Model 3 controls also for the characteristics and idiosyncrasies of the various *platforms* (platform characteristics). Finally, Model 4 uses the interactions between various course characteristics, controls for the popularity of the course, and the prominence of the university on MOOCs, Model 5 uses also the gender of the student, and Model 6 controls for students attending a formal educational institution; a smaller number of observations was available for Model 5 and Model 6. Fitting different specifications also allows us to test the robustness of our models. The results obtained using the two alternative methods for opinion identification, described in the previous section, are similar and, henceforth, only results using the sentiment analysis mechanism are presented, because of the space limitations. Table 2 presents the results obtained using the different specifications for the logit ordered choice model; similar results were also obtained using the probit order choice model specification.

Table 2. Course Retention Ordered Choice Model

Estimates	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Constant	**3.9598*****	**4.9278*****	**6.0277*****	**13.4146*****	**15.3925*****	**14.6814*****
	(0.6170)	(0.8286)	(1.6078)	(2.9766)	(3.1672)	(4.2600)
Assignments	1.5263	**2.2288***	**2.1729***	**2.2127***	2.1999	1.0635
	(1.2364)	(1.2806)	(1.2832)	(1.3021)	(1.3903)	(2.3073)
Professor	**2.3389***	**2.1863***	**2.1834***	**2.2163***	**2.1886***	2.7785
	(1.2383)	(1.2560)	(1.2542)	(1.2580)	(1.3067)	(2.0393)
Discussion Forum	**-3.8139****	**-3.7593****	**-3.6273****	**-3.7215****	**-4.0392****	**-5.0402***
	(1.5970)	(1.6440)	(1.6461)	(1.6489)	(1.7538)	(2.7933)
Course Material	0.6782	0.5295	0.5350	0.5793	0.7241	2.6901
	(1.2708)	(1.2726)	(1.2770)	(1.2585)	(1.3247)	(2.2039)
Self-paced	-0.4641	-0.4639	-0.3851	**-10.716*****	**-16.8596***	**-17.9977***
	(0.5377)	(0.5564)	(0.6850)	(3.3558)	(8.8453)	(10.8333)
Difficulty	**0.3041***	-0.0887	-0.0958	**-2.7487*****	**-2.9125*****	**-2.2418***
	(0.1584)	(0.1853)	(0.1890)	(0.7493)	(0.7920)	(1.2399)
Workload	**-0.1664*****	**-0.0924***	-0.0764	-0.2103	-0.4135	-0.7555
	(0.0437)	(0.0471)	(0.0493)	(0.2650)	(0.2914)	(0.4843)
Weeks	**-0.1296*****	**-0.1203*****	**-0.1252*****	**-0.6288*****	**-0.6171****	-0.3342
	(0.0340)	(0.0379)	(0.0391)	(0.2364)	(0.2453)	(0.3713)
Certificate	**0.3799****	0.0116	-0.0937	0.2114	0.0026	-0.3241
	(0.1677)	(0.1918)	(0.2086)	(0.2491)	(0.2714)	(0.4108)
Peer Assessments	**1.7670*****	**1.3055*****	**1.3606*****	**0.9291*****	**1.1171*****	0.5559
	(0.2706)	(0.2915)	(0.3011)	(0.3438)	(0.3610)	(0.5152)
Team Projects	**-1.9212****	-1.4399	**-1.5771***	-1.2295	**-1.9028****	**-4.8089*****
	(0.8723)	(0.8762)	(0.8896)	(0.9216)	(0.9461)	(1.7193)
Suggested Textbook	0.0691	**1.0653*****	**1.1038*****	**1.3616*****	**1.4586*****	**1.7157*****
	(0.2855)	(0.3455)	(0.3476)	(0.3991)	(0.4279)	(0.6053)
Suggested Paid Textbook	0.0380	**-1.2164*****	**-1.1827*****	**-1.3849*****	**-1.6432*****	**-1.9530*****
	(0.3475)	(0.4275)	(0.4348)	(0.4652)	(0.4948)	(0.7013)
Final Exam/Project	**1.2159*****	**0.8153*****	**0.7559*****	**0.4593***	0.3408	0.1223
	(0.1985)	(0.2154)	(0.2301)	(0.2594)	(0.2717)	(0.3748)
University Ranking	**-0.0321*****	**-0.0270****	**-0.0390****	**-0.0415***	**-0.0653*****	-0.0567
	(0.0119)	(0.0124)	(0.0179)	(0.0229)	(0.0245)	(0.0345)

Business & Management		**0.8299****	**0.8536****	**1.0338*****	**1.1646*****	0.8711
		(0.3468)	(0.3483)	(0.3737)	(0.3987)	(0.5535)
Computer Science		**0.5235***	0.4959	**0.8257****	0.4778	-0.5413
		(0.3087)	(0.3154)	(0.3982)	(0.4321)	(0.6218)
Engineering		0.0010	-0.0081	0.1848	0.0111	-0.4097
		(0.5490)	(0.5529)	(0.6026)	(0.6462)	(1.0279)
Humanities		**-0.8842****	**-0.8464****	**-0.7931****	**-0.9107****	**-1.1150***
		(0.3593)	(0.3654)	(0.4035)	(0.4326)	(0.6303)
Mathematics		-0.6146	-0.6582	-0.5737	-0.6943	-0.5447
		(0.4121)	(0.4187)	(0.4426)	(0.4628)	(0.6091)
Science		0.3471	0.4381	0.6384	0.2107	-0.2312
		(0.3718)	(0.3805)	(0.4318)	(0.4494)	(0.6042)
Self-paced×Difficulty				**4.0867*****	5.1084	4.8426
				(1.1802)	(3.3601)	(4.1269)
Difficulty×Workload				**0.1042***	**0.1667****	**0.2401***
				(0.0624)	(0.0708)	(0.1227)
Difficulty×Weeks				**0.2367*****	**0.2000****	0.0425
				(0.0869)	(0.0882)	(0.1306)
Workload×Weeks				-0.0248	-0.0200	-0.0013
				(0.0209)	(0.0222)	(0.0373)
Interested Users				0.0074	0.0109	0.0191
				(0.0078)	(0.0082)	(0.0117)
Course Reviews				0.0052	0.0048	-0.0058
				(0.0037)	(0.0039)	(0.0059)
Courses Offered				-0.0153	-0.0181	-0.0031
				(0.0134)	(0.0138)	(0.0197)
Student Gender					0.1702	0.2342
					(0.1853)	(0.2879)
Matriculated Student						-0.4539
						(0.2870)
Platform dummies	-	-	Yes	Yes	Yes	Yes
N	1163	1163	1163	1163	1043	441
Restricted log likelihood	-834.490	-834.490	-834.490	-834.490	-775.724	-310.695
K	17	23	28	35	36	37
Log likelihood function	-701.310	-676.181	-674.883	-664.827	-609.834	-260.460
Chi squared	266.3575	316.6166	319.2124	339.325	331.780	100.470
Significance level	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
McFadden Pseudo R^2	0.1595	0.1897	0.1912	0.2033	0.2139	0.1617
Inf.Cr.AIC	1436.6	1398.4	1405.8	1399.7	1291.7	590.9
AIC/N	1.235	1.202	1.209	1.203	1.238	1.340

Note: ***, **, * → Significance at 1%, 5%, 10% level.

As Table 2 illustrates, all the examined specifications are statistically significant (we reject the hypothesis that all the slope coefficients are zero with $p<0.001$) and do not violate the proportional parallel lines assumption (Agresti 2002). Also, the *academic disciplines* and the *platform* dummy variables provide significant explanatory power to our model. As we can see, the extended logit model (Model 5), with *course* characteristics, *academic disciplines, platform* indicators, and the gender of the *student*, provides the best fit to our data with McFadden pseudo R^2 = 0.2139, which represents a very good fit for an ordered choice model (McFadden 1978).

Interpretation of the coefficients in the ordered choice models is more complicated than the ordinary regression setting. There is no natural conditional mean function in the model to analyze (this is a characteristic of discrete choice models), the outcome variable is merely a label for the non-quantitative outcomes, and the coefficients of the logit model are usually larger than the probit because of the functional form (Greene 2003). In order to attach meaning to the parameters, the partial effects should be examined. Table 3 presents the partial effects on the probability of a student successfully completing a course at means using the aforementioned models of logit ordered choices; very similar results were also obtained using the probit order choice model specification.

Table 3. Partial effects on Prob[Y=Course Successful Completion] at means						
Estimates	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
Assignments	0.2592	**0.3791***	**0.3700***	**0.3698***	0.3872	0.1711
Professor	**0.3973***	**0.3719****	**0.3718***	**0.3704***	**0.3852***	0.4471
Discussion Forum	**-0.6478****	**-0.6394*****	**-0.6177****	**-0.6219****	**-0.7109****	**-0.8111***
Course Material	0.1152	0.0901	0.0911	0.0968	0.1275	0.4329
Self-paced	-0.0873	-0.0874	-0.0715	**-0.8981*****	**-0.7859*****	**-0.8293*****
Difficulty	**0.0516***	-0.0151	-0.0163	**-0.4593*****	**-0.5126*****	**-0.3607***
Workload	**-0.0282*****	**-0.0157***	-0.0130	-0.0352	-0.0728	-0.1216
Weeks	**-0.0220*****	**-0.0205*****	**-0.0213*****	**-0.1051*****	**-0.1086****	-0.0538
Certificate	**0.0675****	-0.0020	-0.0158	0.0363	0.0005	-0.0493
Peer Assessments	**0.2253*****	**0.1792*****	**0.1854*****	**0.1330*****	**0.1666*****	0.0801
Team Projects	**-0.4363****	-0.3213	-0.3552	-0.2663	**-0.4356****	**-0.7699*****
Suggested Textbook	0.0116	**0.1638*****	**0.1693*****	**0.2000*****	**0.2275*****	**0.2374*****
Suggested Paid Textbook	0.0064	**-0.2491****	**-0.2416****	**-0.2850*****	**-0.3539*****	**-0.4048*****
Final Exam/Project	**0.1889*****	**0.1306*****	**0.1217*****	**0.0742***	0.0588	0.0196
University Ranking	**-0.0055*****	**-0.0046****	**-0.0066****	**-0.0069***	**-0.0115*****	-0.0091
Business & Management		**0.1184*****	**0.1214*****	**0.1383*****	**0.1604*****	**0.1161***
Computer Science		**0.0863***	0.0820	**0.1314****	0.0814	-0.0883
Engineering		0.0002	-0.0014	0.0294	0.0020	-0.0735
Humanities		**-0.1609****	**-0.1538****	**-0.1412***	**-0.1685****	-0.1985
Mathematics		-0.1205	-0.1303	-0.1100	-0.1419	-0.0998
Science		0.0547	0.0678	**0.0924***	0.0355	-0.0392
Self-paced×Difficulty				**0.6829*****	0.8991	0.7793
Difficulty×Workload				**0.0174***	**0.0293****	**0.0386***
Difficulty×Weeks				**0.0396*****	**0.0352****	0.0069
Workload×Weeks				-0.0042	-0.0035	-0.0002
Interested Users				0.0009	0.0019	0.0031
Course Reviews				0.0009	0.0009	-0.0009
Courses Offered				-0.0028	-0.0032	-0.0005
Student Gender					0.0302	0.0386
Matriculated Student						-0.0781
Platform dummies	-	-	Yes	Yes	Yes	Yes

Note: Marginal Effects for dummy variables are $\Pr[y|x = 1] - \Pr[y|x = 0]$.
Note: ***, **, * → Significance at 1%, 5%, 10% level.

As Table 3 shows, using Model 5, which provides the best fit to our data, *Professors* have the largest positive significant effect (0.39, p<0.1) on the probability of a student to successfully complete a course. *Assignments* (0.39) and *Course Material* (0.13) also have positive effects on the successful completeness of a course whereas the *Discussion Forum* has a negative effect (-0.71, p<0.1). In other words, the more satisfied (i.e. positive sentiment) a student is with the professor, the teaching material, and the assignments, the more probable that s/he will successfully complete the course. However, a negative effect for the satisfaction of the user with the discussion forum was estimated. Examining the partial effects of *Discussion Forum* on the other possible outcomes, the sentiment of the student about the discussion forum has a small positive effect (0.09) on the probability of the student not to complete the course but a stronger positive effect on the probability of the student to partially complete the course (0.62, p<0.05). One possible explanation which we investigated is that mostly students who are not satisfied or have difficulties with the course and are more likely not to successfully complete a course are using the online discussion forum and then comment on this in their reviews, but no statistically significant difference was observed on the distributions. Another possible explanation would be that students who are more likely not to complete a course have low expectations for the discussion forums and, hence, they express more easily a positive sentiment, but no statistically significant difference in the mean values among the different groups was found. Furthermore, Table 3 also shows that self-paced courses compared to courses that follow a specific timetable have a negative effect (-0.79, p<0.01). In

addition, the *difficulty* of the course (-0.51, p<0.01), the *workload* (hours/week) (-0.07), and its duration in *weeks* (-0.11, p<0.05) have a negative effect on student retention. Besides, peer assessment has a positive effect (0.17, p<0.01) compared to automated feedback indicating that better technological solutions for automatically providing feedback and evaluating the assignments of the students are still needed. Although, final exams and projects make the courses more engaging and have a positive effect (0.06), team projects that require the active collaboration with other students do not have the same effect (-0.44, p<0.05). Interestingly, there is a positive effect (0.23, p<0.01) when there is a suggested textbook on the syllabus that the students can refer to; however, there is a negative effect (-0.35, p<0.01) if this is a paid textbook to which most of the students do not have access. Moreover, whether a certificate is awarded upon the successful completion of a course has also a small but positive effect (0.0005). Further, the better a university is considered and the higher it is ranked, the more likely that a student will successfully complete a course (-0.01, p<0.01). In addition, the popularity of the course (i.e. number of users interested in the course on the review platform and number of posted course reviews) and the prominence of the university (i.e. number of courses offered by the specific university) do not have a significant effect. Also, based on the interaction terms, for more difficult courses, self-paced courses (0.90), longer duration in weeks (0.04, p<0.05) and more workload (0.03, p<0.05) have also positive effects on the probability of a student to successfully complete a course. Besides, Table 3 illustrates that courses which belong to the disciplines of Business and Management, Computer Science, and Science have also a positive significant effect in contrast to other disciplines. Finally, there were not found any statistically significant effects based on the gender of students, whether a student attends a formal educational institution, or the various MOOC platforms.

Predictive Modeling

The explanatory study that we described above revealed what factors influence course retention and their relative importance. In this section, we switch from explanatory modeling to *predictive* modeling. In other words, the main goal now is not to explain which factors affect the course retention, but to examine how well we can predict whether a student will successfully complete a course. This method will also provide us evidence of whether there are other concepts discussed in the user-generated reviews that we have not identified in the proposed approach. For this *multiclass classification* problem, we train a Random Forests classifier (Breiman 2001) employing a holdout validation scheme (Provost and Fawcett 2013) with 80/20 split of data, and we evaluate each model based on the F1-score measure.[5]

Table 4. Classification Performance of the Predictive Model

Student Course Evaluation	Course Characteristics	University Characteristics	Academic Disciplines	Platform Characteristics	Student Characteristics	Review Nouns Sentiment	Review Corpus	Random Forest Classifier
-	-	-	-	-	-	-	-	0.5887*
✓	-	-	-	-	-	-	-	0.6170
✓	✓	-	-	-	-	-	-	0.7728
✓	✓	✓	-	-	-	-	-	0.7973
✓	✓	✓	✓	-	-	-	-	0.8004
✓	✓	✓	✓	✓	-	-	-	0.8128
✓	✓	✓	✓	✓	✓	-	-	0.8108
✓	✓	✓	✓	✓	✓	✓	-	0.7825
✓	✓	✓	✓	✓	✓	-	✓	0.8084

Note: * The Baseline corresponds to predicting the most frequent class in the training set (i.e. "Course Successfully Completed").

Table 4 presents the results of our experimental evaluation. The first line shows the *Baseline* method and corresponds to predictions made using the most popular class (i.e. "Course Successfully Completed") in our train set. Each of the next lines shows the experimental performance of our predictive model using the

[5] Other models that were tested and yielded similar results include decision trees (Breiman et al. 1984), one vs. one multiclass strategy with support vector classification (Chang and Lin 2011), and one vs. all multiclass strategy with logistic regression (Yu et al. 2011) and ridge regression.

corresponding features (i.e. columns of Table 4). The first result that we observe is that using a classifier with the features that emerged based on our textual analysis (i.e. Assignments, Professor, Discussion Forum, and Course Material), labeled here as *Course Evaluation Features*, outperforms the standard *Baseline* method by 4.8% in terms of the F1-score (i.e. 0.62 F1-score). Besides, Table 4 illustrates that the *course characteristics* features (i.e. self-paced, difficulty, workload, etc.) provide significant predictive power to our model (i.e. 0.77 F1-score). Moreover, there is further increase in performance using the university ranking (*university characteristics*), the academic disciplines (*course characteristics*), and the platform dummies (*platform characteristics*). However, there is no further improvement using *student characteristics*. The best performance (i.e. 0.81 F1-score) outperforms the standard *Baseline* by 38.1% and was achieved using all the aforementioned features.

Finally, in order to test whether there are any features that affect student retention and have not yet been identified in our textual analysis, and hence to examine whether we could further improve the performance of our models, we also use the estimated sentiment for each noun and noun group included in the review (labeled as *Review Nouns Sentiment*) as well as the pre-processed corpus of the reviews. Table 4 shows that the performance of the predictive model was reduced and that these extra features introduced more noise than predictive power. This means that we have successfully identified all the features that affect student retention and are discussed in user-generated online course reviews.

Discussion

With the rapid expansion of Massive Open Online Courses (MOOCs), special online communities have emerged in order to reduce the search costs associated with the proliferation of platforms and courses. In such online communities, the user-generated course reviews are an important and unique source of information for both students and researchers who want to better understand the phenomenon of MOOCs. Working toward the direction of improving MOOCs and reducing attrition, in order to conduct a more in-depth analysis of the information captured by user-generated content, we integrate multiple research methods. In this paper,

- We explore one of the most important problems of MOOCs and open education identifying both the concepts that affect online course retention and their relative impact.
- We illustrate the use of the Grounded Theory Method in a quantitative research, a less frequently applied paradigm, and integrate state-of-the-art econometric, text mining, opinion mining, and machine learning techniques, building both explanatory and predictive models.
- We contribute to the related literature discovering new rich findings and provide actionable insights to solve a problem of scientific and societal interest about a breakthrough phenomenon.

In particular, seeking to provide a deep understanding of the phenomenon and develop theory grounded in empirical observations, the Grounded Theory Method (GTM) is the most appropriate research method for this study. Despite the misconceptions about GTM (Suddaby 2006; Urquhart and Fernandez 2013), its roots in quantitative methods make GTM a powerful tool also for quantitative research by prioritizing the exploration of the given phenomenon and allowing to conduct research with both flexibility and rigor (Birks et al. 2013; Corbin and Strauss 2007). The greatest advantage of GTM is the logic of discovery, rather than that of verification, which is essential to the delicate question of grounded theory building (Vaast and Walsham 2011). Another advantage of GTM as inductive methodology for theory generation is that it allows the discovery of causal relationships and not just correlations. Besides, many IS researchers have already relied upon GTM to develop new theory, while its use could be further promoted through the integration with explanatory and predictive techniques. This integration of GTM with econometric and machine learning techniques could bring radically different approaches and new theories to the IS field due to the innovative possibilities of "big data." Moreover, such methods can lead to scalable and reproducible research and more consistent results alleviating also researcher biases. This integration also allows researchers to consistently quantify qualitative data in a continuous numerical scale (e.g. opinion mining and sentiment analysis) and evaluate the predictive and explanatory power of the emergent theory.

In this study, as GTM requires, we avoided using specific theories as the starting point or formulating the hypotheses in advance (preconceived hypotheses result in a theory that is ungrounded from the data). Nonetheless, hypothesis generation in GTM is essentially the statement of probabilities that explain latent patterns of social behavior (Glaser 1998) and, thus, fairly fits with the goals of choice models. Overall, we

let the data speak, not by imposing beforehand a conceptual framework upon it, but through an inductive approach, which avoids pre-existing knowledge bias, and some back and forth between a conceptual and an empirical plane. In the same way, we let also the theory emerge from the data and become gradually refined through comparison of existing theories.

Furthermore, we decided to use a real-world data set with user-generated information in order to conduct our analysis. In contrast to data collected from surveys or course evaluations, which usually involves a low number of responses about only few courses and only touches on some aspects of a course while aspects not obvious to faculty are ignored, this user-generated information provides a holistic view and can be effectively used to improve students' experience with MOOCs by understanding some of their most important problems. Also, the specifics of the constructed data set allow for a more granular level of analysis. Besides, another major advantage of our data set is that course reviews are submitted by students that consciously registered in the corresponding course and intended to successfully complete it. Hence, students' reviews can be used to estimate better models of course retention compared to the use of average statistics (e.g. per cent passed the course based on course start).

Implications for Research and Theoretical Integration

Based on the prior literature on student retention in traditional education and the initial discussion presented in this paper, the aforementioned results both quantify and verify the conceptualized high significance of professors and instructors in MOOCs (Cormier and Siemens 2010; Masters 2011). Also, the finding that the likelihood of dropping a course depends on the academic discipline is partially supported by prior research on community colleges affirming that while some areas attract students with a strong ability to adapt to online coursework, others attract students who do not adapt well (Xu and Jaggars 2013). Similarly, using a different classification on broad fields of education, Olsen (2008) found that Australian fulltime students in Health, Engineering, Management, Architecture and Education stayed the course better than students in other disciplines, but only aggregate statistics across all students were used and no marginal effects. Further, our results are also supported by qualitative studies suggesting that students prefer to take difficult courses in a traditional setting rather than online (Jaggars 2012) and that demographics variables do not predict the likelihood of dropping from online courses (Glass Jr and Garrett 1995; Pantages and Creedon 1978; Willging and Johnson 2004). Also, our findings are in accordance with prior research on the effect of difficulty (Xenos et al. 2002), workload (Tresman 2002), and team projects (Knight et al. 2007) on course retention in traditional educational settings. In addition, our finding about discussion forums may seem to partially contradict that retention in traditional education (not individual courses) could be improved by increasing the integration of students into the academic and social systems of the college (Tinto 1987). However, one possible explanation is a selection effect according to which students who are more likely not to successfully complete a course have low expectations for the discussion forums and, hence, they express more easily or more often a positive sentiment, but no statistically significant difference was found. Another possible explanation is that discussion forums can reduce student retention because of the diversity among the students. In particular, variations in levels of expertise and in individual online behavior can limit connectivity and lead to the formation of groups. Group formation in turn can reduce possibilities for autonomy, openness and diversity, which in turn can reduce the opportunities for connectivity, connectedness and engagement, and so on (Mackness et al. 2010). Finally, our findings contradict prior results on self-pacing versus instructor-pacing courses (Morris et al. 1978) while we provide richer explanations accounting for many interaction effects and controlling for various factors.

Because of the novelty of MOOCs and the different levels of analysis in the prior literature on traditional education, most of our findings cannot be directly compared to the results of previous studies. However, the aforementioned findings can also be confirmed using the actual conversations of the students on the MOOC platforms. For instance, discussing automated feedback various students have commented that "the autograder was very frustrating at times", whereas other students mentioned that "the weekly projects made it very fun and having to grade other students gave me a glimpse into how others are solving the same problem". Besides, many reviews discuss also the inefficiencies of discussion forums explaining that students "dropped this class as it was unbalanced and the team did not respond to issues in a well-rounded manner." Also, as far as the final projects are considered, there were reviews mentioning that "the course project has everyone excited about it as we are addressing real life issues and solving them in our own way." In addition, the positive effect of textbooks is also depicted on numerous

posts referring to MOOCs that "follow a very good textbook written by the instructor." Similar comments are also common for other course components of MOOCs (e.g. "the professor is a great teacher", "the assignments were hard but the lecture material was excellent and easy to follow"). Finally, even though a small fraction of the students expressed different preferences across the various MOOC platforms, they did not consider these differences to be significant but "more of a personal choice".

Managerial Implications and Policy

This study also provides several insights with significant implications for course design, the use of technology in learning environments, and educational policy. Based on the results, in order to increase the engagement of students and reduce dropout rates, online courses should be of average difficulty for the students (i.e. less than 2.9 on a scale of 1 to 5) and require a moderate workload (i.e. about 6 hours/week) without spanning many weeks (i.e. less than 8 weeks). Also, courses should have a calendar-based schedule and final exams or projects. Moreover, the use of textbooks is beneficial for the students since they can refer to standard offline sources. However, the use of paid textbooks has the opposite results since it restricts the access to important information for a large portion of the students. In addition, peer assessment results in increased engagement compared to automatic grading, which indicates that better technological solutions are still needed. Interestingly, our results also confirm that the current form of the certifications provided in MOOCs is of limited usefulness to the students. The certificates should be enhanced offering both knowledge verification and student identification. Besides, public policy could also evaluate the credit-equivalency of MOOCs in order to increase retention. Further, our results about discussion forums indicate that platforms need to better facilitate knowledge construction and communication among the students. For instance, the introduction of wikis can supplement and enhance the discussion forums by providing a reference for codified knowledge separated from the miscellaneous communication. Finally, the finding that traditional university rankings do not significantly affect course retention in MOOCs indicates that the content of a course is becoming more important. Universities that offer MOOCs can build their reputation based on the content they offer and, thus, new significant players might emerge in education. As a consequence, college and university rankings should be updated in order to take into consideration these new factors when they evaluate educational institutions.

Apart from the implications in education and the methodological ideas of GTM, the presented analytical approach has also important managerial implications in other domains, such as social media and online commerce. For instance, the proposed approach can be effectively used to identify important attributes of various products and dynamically measure and monitor the corresponding sentiment on social media, estimate their relative effect on sales and brand perception, actively manage dissatisfied customers on real-time, and accurately predict churns. Such research also raises keys issues concerning the systematic discovery of new theories and interesting patterns through the analysis of "big data."

Limitations and Future Research

Nevertheless, we should be careful interpreting the aforementioned findings since MOOCs are only a recent development and all the providers, students, and platforms are still adjusting to this new phenomenon. For instance, after this initial adjusting period, students may start focusing also on other aspects of the courses. This hypothesis is partially supported by the fact that the main characteristics of the courses that students evaluate in their reviews are not very different from the ones evaluated in a traditional educational setting. Thus, future research should re-examine the described problem and study the adoption process among the students. Another limitation of our work is that the course retention variable was self-reported by the students and not directly acquired from the course providers. Additionally, in the quantitative analysis we only use data about students that submitted at least one course review and not all the students that attended any of these courses. Besides, no specific features of the platforms, such as usability and technical characteristics, were included in the conducted analysis; to partially address this limitation we use indicator variables for the platform that hosts each course. Also, the problem of course retention should be studied vis-à-vis the satisfaction of the students and the corresponding course evaluation.

Moreover, future research is needed to look into how online socialization influences the dropout decision of students in MOOCs. Prior work on traditional education has illustrated that social life has large significant effects on institutional fit for each class (Pascarella 1980; Spady 1970; Tinto 1975) and has

indicated that students have a much greater effect on the attitudes of other students than do faculty members, considering them the primary agents of socialization in this type of academic environment (Bean 1985). Also, future work should look into the interactions between the individual and the academic and social systems. In particular, researchers should closely study individual student goals (e.g. importance of successfully completing the course), attributes (e.g. ability), educational background and academic performance (e.g. academic major, grade-point average, and academic attainments), family background (e.g. social status attributes), faculty contact and interaction (e.g. type and frequency of contact), as well as environmental factors (e.g. finances, employment rates).

Conclusions

Massive Open Online Courses (MOOCs) have remarkably expanded during the last years offering a rapidly growing number of courses across various platforms. However, the very high drop-out rates indicate that much more should be done in order to satisfy the actual educational needs of the students. Tackling this important problem, we employ the Grounded Theory Method (GTM) on quantitative data, a less frequently applied paradigm. In particular, we present an innovative analysis using a real-world data set with user-generated online reviews, where we both identify the different *student*, *course, platform,* and *university* characteristics that affect student retention and study their relative effect. An important aspect of the conducted analysis is the integration of econometric, text mining, opinion mining, and machine learning techniques, building both *explanatory* and *predictive* models, toward a more complete analysis of the information captured by user-generated content. Extending the methodological ideas of GTM to explanatory quantitative analysis and going beyond descriptive statistics of coded verbal data, we both contribute to the related literature discovering new rich findings and provide actionable insights with implications for both MOOCs and education. The use of radically different approaches and fresh interdisciplinary perspectives in this paper offers the potential to inspire future research and open up new streams of scientific inquiry for the IS field.

The findings of our analysis illustrate that *Professor(s)* is the most important factor in online course retention and has the largest positive effect on the probability of a student to successfully complete a course. The sentiment of students for *Assignments* and *Course Material* also has positive effects on the successful completeness of a course whereas the *Discussion Forum* has a positive effect on the probability to partially complete a course. Furthermore, self-paced courses have a negative effect, compared to courses that follow a specific timetable. In addition, the *difficulty*, the *workload,* and the *duration* of a course have a negative effect. On the other hand, for the more difficult courses, self-paced timetable, longer duration in weeks, and more workload have a positive effect on the probability to successfully complete a course. Besides, final exams and projects, open textbooks, and peer assessment have also positive effects. Moreover, whether a certificate is awarded upon the successful completion of a course also affects retention. Additionally, the better a university is considered (i.e. higher ranking), the more likely that a student will successfully complete a course. Further, our results illustrate that the courses which belong to the academic disciplines of Business and Management, Computer Science, and Science have a positive significant effect in contrast to courses in other disciplines (i.e. Engineering, Humanities, and Mathematics). Finally, attrition was not found to be related with student characteristics (i.e. gender, formal education). The aforementioned findings were also verified by the predictive study we conducted. The employed predictive modeling techniques also indicate that we have successfully identified all the features that affect course retention and are discussed in user-generated course reviews.

The proposed approach and the corresponding findings have several implications for the universities and platforms offering online courses and can be used for real-time detection of dissatisfied students as well as the design of better and more engaging courses that will increase retention rates. In particular, the results suggest that the course characteristics (e.g. estimated difficulty, workload, duration, whether there is automated grading, etc.) are important determinants of students' satisfaction and suggest useful guidelines for course design. For instance, MOOCs in general should have a specific instructor-based timetable, but for the most difficult courses students should be allowed to follow their own pace. Also, the findings suggest that there is room for improvement in the current form of certifications which should be redesigned in order to become more useful for the students and better motivate them to successfully complete the corresponding course. Moreover, the positive effect of the sentiment of students for the discussion forum on the probability to partially complete the course illustrates that better mechanisms or

complementary technologies, such as wikis, are still needed in order to successfully advice, assist, connect, and motivate the students. In addition, the finding that courses in the disciplines of Business and Management, Computer Science, and Science are more likely to be successfully completed may suggest that either specific types of courses are better suited for online education or are better accepted by the students; an issue that should be addressed by future research. Finally, apart from the implications in education and the methodological ideas of GTM, we also discuss the managerial implications of the proposed analytical approach in other domains, such as social media and online commerce.

As part of the future work, we will study the dynamics of student online communities that participate in discussion forums of MOOCs. Furthermore, we will study how different factors affect the popularity of MOOCs and the helpfulness (Ghose and Ipeirotis 2011) of online MOOC reviews in order to reduce the search costs for students and enable them to make more informed decisions. Moreover, we will also study the student dropout from online education and not just from individual online courses. Finally, the growing number of available courses for each academic discipline and topic will soon make it even harder for students to locate the best MOOCs for them, and thus we plan to develop a recommender system (Adamopoulos and Tuzhilin 2013b; Adomavicius and Tuzhilin 2005) to deliver to students personalized unexpected (Adamopoulos 2013; Adamopoulos and Tuzhilin 2011; Adamopoulos and Tuzhilin 2013a) recommendations for courses, which fit their interests and educational needs, based on their profiles and the submitted course reviews (Aciar et al. 2006; Terzi et al. 2011).

Acknowledgments

The author is grateful to Anindya Ghose, Natalia Levina, Vilma Todri, and Alexander Tuzhilin as well as the track chairs, associate editors, and three anonymous reviewers at the International Conference on Information Systems (ICIS 2013) for their constructive comments and valuable feedback on an earlier version of the manuscript.

References

Aciar, S., Zhang, D., Simoff, S., and Debenham, J. 2006. "Recommender System Based on Consumer Product Reviews," in: *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*. Washington, DC, USA: IEEE Computer Society, pp. 719-723.

Adamopoulos, P. 2013. "Beyond Rating Prediction Accuracy: On New Perspectives in Recommender Systems," *Proceedings of the seventh ACM conference on Recommender Systems (RecSys'13)*, Hong Kong, China: ACM.

Adamopoulos, P., and Tuzhilin, A. 2011. "On Unexpectedness in Recommender Systems: Or How to Expect the Unexpected," *Workshop on Novelty and Diversity in Recommender Systems (DiveRS 2011), at the 5th ACM International Conference on Recommender Systems (RecSys'11),* P. Castells, J. Wang, R. Lara and D. Zhang (eds.), Chicago, Illinois, USA: ACM, pp. 11-18.

Adamopoulos, P., and Tuzhilin, A. 2013a. "On Unexpectedness in Recommender Systems: Or How to Better Expect the Unexpected," *NYU Working Paper No. 2451/31832*), June 19, 2013, pp. 1-50.

Adamopoulos, P., and Tuzhilin, A. 2013b. "Recommendation Opportunities: Improving Item Prediction Using Weighted Percentile Methods in Collaborative Filtering Systems," *Proceedings of the seventh ACM conference on Recommender Systems (RecSys'13)*, Hong Kong, China: ACM.

Adomavicius, G., and Tuzhilin, A. 2005. "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *Ieee Transactions on Knowledge and Data Engineering* (17:6), Jun, pp. 734-749.

Agresti, A. 2002. *Categorical Data Analysis*. Wiley-interscience.

AlchemyApi. 2012. "Orchestr8, Llc." from http://www.alchemyapi.com/

Archak, N., Ghose, A., and Ipeirotis, P.G. 2007. "Show Me the Money!: Deriving the Pricing Power of Product Features by Mining Consumer Reviews," in: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. San Jose, California, USA: ACM, pp. 56-65.

Archak, N., Ghose, A., and Ipeirotis, P.G. 2011. "Deriving the Pricing Power of Product Features by Mining Consumer Reviews," *Management Science* (57:8), Aug, pp. 1485-1509.

Armstrong, J.S. 2012. "Natural Learning in Higher Education," in *Encyclopedia of the Sciences of Learning*. Springer, pp. 2426-2433.

Avakian, A.N., Mackinney, A.C., and Allen, G.R. 1982. "Race and Sex-Differences in Student Retention at an Urban University," *College and University* (57:2), pp. 160-165.
Bean, J.P. 1985. "Interaction Effects Based on Class Level in an Explanatory Model of College-Student Dropout Syndrome," *American Educational Research Journal* (22:1), pp. 35-64.
Bedford, M.H., and Durkee, P.E. 1989. "Retention: Some More Ideas," *NASPA Journal* (27:2), pp. 168-171.
Bickart, B., and Schindler, R.M. 2001. "Internet Forums as Influential Sources of Consumer Information," *Journal of interactive marketing* (15:3), pp. 31-40.
Birks, D.F., Fernandez, W., Levina, N., and Nasirin, S. 2013. "Grounded Theory Method in Information Systems Research: Its Nature, Diversity and Opportunities," *European Journal of Information Systems* (22:1), Jan, pp. 1-8.
Boudreau, C.A., and Kromrey, J.D. 1994. "A Longitudinal-Study of the Retention and Academic-Performance of Participants in Freshmen Orientation Course," *Journal of College Student Development* (35:6), Nov, pp. 444-449.
Breiman, L. 2001. "Random Forests," *Machine Learning* (45:1), Oct, pp. 5-32.
Breiman, L., Friedman, J., Stone, C.J., and Olshen, R.A. 1984. *Classification and Regression Trees*. Chapman & Hall/CRC.
Bruce, R.F., and Wiebe, J.M. 1999. "Recognizing Subjectivity: A Case Study in Manual Tagging," *Natural Language Engineering* (5:2), pp. 187-205.
Cabrera, A.F., Castañeda, M.B., Nora, A., and Hengstler, D. 1992. "The Convergence between Two Theories of College Persistence," *The Journal of Higher Education* (63:2), pp. 143-164.
Chang, C.C., and Lin, C.J. 2011. "Libsvm: A Library for Support Vector Machines," *Acm Transactions on Intelligent Systems and Technology* (2:3), p. 27.
Clow, D. 2013. "Moocs and the Funnel of Participation," in: *Proceedings of the Third International Conference on Learning Analytics and Knowledge*. Leuven, Belgium: ACM, pp. 185-189.
Corbin, J., and Strauss, A. 2007. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications, Incorporated.
Cormier, D., and Siemens, G. 2010. "The Open Course: Through the Open Door--Open Courses as Research, Learning, and Engagement," *Educause Review* (45:4), pp. 30-32.
CourseTalk.org. 2012. "Coursetalk.Org." from http://coursetalk.org/
Das, S.R., and Chen, M.Y. 2007. "Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web," *Management Science* (53:9), Sep, pp. 1375-1388.
Dave, K., Lawrence, S., and Pennock, D.M. 2003. "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," in: *Proceedings of the 12th international conference on World Wide Web*. Budapest, Hungary: ACM, pp. 519-528.
Dey, E.L. 1997. "Working with Low Survey Response Rates: The Efficacy of Weighting Adjustments," *Research in Higher Education* (38:2), 1997/04/01, pp. 215-227.
Dey, E.L., and Astin, A.W. 1993. "Statistical Alternatives for Studying College-Student Retention - a Comparative-Analysis of Logit, Probit, and Linear-Regression," *Research in Higher Education* (34:5), Oct, pp. 569-581.
Dhar, V., and Chang, E.A. 2009. "Does Chatter Matter? The Impact of User-Generated Content on Music Sales," *Journal of Interactive Marketing* (23:4), 11//, pp. 300-307.
Downes, S. 2010. "Learning Networks and Connective Knowledge," *Collective intelligence and e-learning* (2), pp. 1-26.
Eisenhardt, K.M. 1989. "Building Theories from Case-Study Research," *Academy of Management Review* (14:4), Oct, pp. 532-550.
Fellbaum, C. 1998. *Wordnet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
Fox, A., and Patterson, D. 2012. "Crossing the Software Education Chasm," *Communications of the Acm* (55:5), May, pp. 44-49.
Fuller, R.B. 1962. *Education Automation: Freeing the Scholar to Return to His Studies: A Discourse before the Southern Illinois University, Edwardsville Campus Planning Committee, April 22, 1961*. Southern Illinois University Press.
Ghani, R., Probst, K., Liu, Y., Krema, M., and Fano, A. 2006. "Text Mining for Product Attribute Extraction," *ACM SIGKDD Explorations Newsletter* (8:1), pp. 41-48.
Ghose, A., and Ipeirotis, P.G. 2011. "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics," *Ieee Transactions on Knowledge and Data Engineering* (23:10), Oct, pp. 1498-1512.
Ghose, A., Ipeirotis, P.G., and Li, B. 2012. "Designing Ranking Systems for Hotels on Travel Search

Engines by Mining User-Generated and Crowdsourced Content," *Marketing Science* (31:3), May-Jun, pp. 493-520.
Glaser, B.G. 1998. *Doing Grounded Theory: Issues and Discussions*. Sociology Press.
Glaser, B.G., and Strauss, A.L. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine de Gruyter.
Glass Jr, J.C., and Garrett, M.S. 1995. "Student Participation in a College Orientation Course, Retention, and Grade Point Average," *Community College Journal of Research and Practice* (19:2), pp. 117-132.
Goldsmith, R.E., and Horowitz, D. 2006. "Measuring Motivations for Online Opinion Seeking," *Journal of Interactive Advertising* (6:2), pp. 1-16.
Green, P.E., and Srinivasan, V. 1990. "Conjoint-Analysis in Marketing - New Developments with Implications for Research and Practice," *Journal of Marketing* (54:4), Oct, pp. 3-19.
Greene, W. 2003. *Econometric Analysis*. Pearson Education.
Harasim, L., Hiltz, S.R., Teles, L., and Turoff, M. 1995. *Learning Networks: A Field Guide to Teaching and Learning on-Line*. MIT press.
Heiberger, R.M. 1993. "Predicting Next Year's Enrollment: Survival Analysis of University Student Enrollment Histories," *Proceedings of the American Statistical Association, Social Statistical Section*, pp. 143-148.
Hennig-Thurau, T., Gwinner, K.P., Walsh, G., and Gremler, D.D. 2004. "Electronic Word-of-Mouth Via Consumer-Opinion Platforms: What Motivates Consumers to Articulate Themselves on the Internet?," *Journal of interactive marketing* (18:1), pp. 38-52.
Hu, M., and Liu, B. 2004a. "Mining and Summarizing Customer Reviews," in: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 168-177.
Hu, M., and Liu, B. 2004b. "Mining Opinion Features in Customer Reviews," in: *Proceedings of the National Conference on Artificial Intelligence*. pp. 755-760.
Hyman, R.E. 1995. "Creating Campus Partnerships for Student Success," *College and University* (71:2), pp. 2-8.
Iiyoshi, T., and Kumar, S.V. 2008. *Opening up Education: The Collective Advancement of Education through Open Technology, Open Content, and Open Knowledge*. Mit Press.
Jaggars, S.S. 2012. "Beyond Flexibility: Why Students Choose Online Courses in Community College." American Educational Research Association Annual Meeting, Vancouver, Canada.
Joglekar, M., Garcia-Molina, H., and Parameswaran, A. 2013. "Evaluating the Crowd with Confidence," in: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. Chicago, Illinois, USA: ACM, pp. 686-694.
Johnson, R. 1996. "The Adult Student: Motivation and Retention," *American Music Teacher* (46:2), pp. 16-19.
Knight, D.W., Carlson, L.E., and Sullivan, J. 2007. "Improving Engineering Student Retention through Hands-on, Team Based, First-Year Design Projects," *Proceedings of the International Conference on Research in Engineering Education*.
Knowledge@Wharton. 2013. "Moocs on the Move: How Coursera Is Disrupting the Traditional Classroom." *Innovation and Entrepreneurship Articles* Retrieved September 5, 2013, from http://knowledge.wharton.upenn.edu/article.cfm?articleid=3109
Lee, T., and Bradlow, E.T. 2007. "Automatic Construction of Conjoint Attributes and Levels from Online Customer Reviews," *University Of Pennsylvania, The Wharton School Working Paper*).
Lewin, T. 2013. "Universities Abroad Join Partnerships on the Web," in: *The New York Times* February 21, 2013: p. A18.
Liu, Y. 2001. "Word-of-Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," *Journal of marketing* (70:3), pp. 74-89.
Mackness, J., Mak, S., and Williams, R. 2010. "The Ideals and Reality of Participating in a Mooc," *Networked Learing Conference*: University of Lancaster, pp. 266-275.
MacNeely, J.H. 1938. *College Student Mortality*. US Government Printing Office.
Mak, S., Williams, R., and Mackness, J. 2010. "Blogs and Forums as Communication and Learning Tools in a Mooc," in: *Networked Learing Conference*. pp. 275-285.
Manning, C.D., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT press.
Margahny, M., and Mitwaly, A. 2005. "Fast Algorithm for Mining Association Rules," in: *the conference*

proceedings of AIML, CICC, pp (36-40) Cairo, Egypt. pp. 19-21.
Masters, K. 2011. "A Brief Guide to Understanding Moocs," *The Internet Journal of Medical Education* (1:2).
McFadden, D. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior," in: *Frontiers in Econometrics*. Academic Press, pp. 105-142.
McFadden, D. 1978. *Quantitative Methods for Analyzing Travel Behavior of Individuals: Some Recent Developments*. London: Croom Helm.
Moore, R., and Miller, I. 1996. "How the Use of Multimedia Affects Student Retention and Learning," *Journal of College Science Teaching* (25:4), pp. 289-293.
Morris, E.K., Surber, C.F., and Bijou, S.W. 1978. "Self-Pacing Versus Instructor-Pacing - Achievement, Evaluations, and Retention," *Journal of Educational Psychology* (70:2), pp. 224-230.
Muchnik, L., Aral, S., and Taylor, S.J. 2013. "Social Influence Bias: A Randomized Experiment," *Science* (341:6146), August 9, 2013, pp. 647-651.
Murtaugh, P.A., Burns, L.D., and Schuster, J. 1999. "Predicting the Retention of University Students," *Research in Higher Education* (40:3), Jun, pp. 355-371.
Nandeshwar, A., Menzies, T., and Nelson, A. 2011. "Learning Patterns of University Student Retention," *Expert Systems with Applications* (38:12), 11//, pp. 14984-14996.
Naretto, J.A. 1995. "Adult Student Retention: The Influence of Internal and External Communities," *NASPA journal* (32:2), pp. 90-97.
Netzer, O., Toubia, O., Bradlow, E.T., Dahan, E., Evgeniou, T., Feinberg, F.M., Feit, E.M., Hui, S.K., Johnson, J., Liechty, J.C., Orlin, J.B., and Rao, V.R. 2008. "Beyond Conjoint Analysis: Advances in Preference Measurement," *Marketing Letters* (19:3-4), Dec, pp. 337-354.
Olsen, A. 2008. "Staying the Course: Retention and Attrition in Australian Universities." *Paper for the Australian Universities International Directors' Forum* Retrieved September 5, 2013, from http://trove.nla.gov.au/work/153059543?q&versionId=166810055
Pang, B., and Lee, L. 2008. *Opinion Mining and Sentiment Analysis*. Now Pub.
Pang, B., Lee, L., and Vaithyanathan, S. 2002. "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques," in: *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*. Association for Computational Linguistics, pp. 79-86.
Pantages, T.J., and Creedon, C.F. 1978. "Studies of College Attrition - 1950-1975," *Review of Educational Research* (48:1), pp. 49-101.
Pappano, L. 2012. "The Year of the Mooc," in: *The New York Times*. November 4, 2012: p. ED26
Pascarella, E.T. 1980. "Student-Faculty Informal Contact and College Outcomes," *Review of Educational Research* (50:4), pp. 545-595.
Person, D.R., and Christensen, M.C. 1996. "Understanding Black Student Culture and Black Student Retention," *Journal of Student Affairs Research and Practice* (34:1), pp. 70-79.
Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. 2013. "Tuned Models of Peer Assessment in Moocs," *6th International Conference on Educational Data Mining*, Memphis, Tennessee, USA.
Postmes, T., Spears, R., Sakhel, K., and De Groot, D. 2001. "Social Influence in Computer-Mediated Communication: The Effects of Anonymity on Group Behavior," *Personality and Social Psychology Bulletin* (27:10), pp. 1243-1254.
Provost, F., and Fawcett, T. 2013. *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*. O'Reilly Media.
Rankings, Q.W.U. 2012. "Qs Quacquarelli Symonds, Ltd." Retrieved September 5, 2013, from http://www.topuniversities.com
Russell, D.M., Klemmer, S., Fox, A., Latulipe, C., Duneier, M., and Losh, E. 2013. "Will Massive Online Open Courses (Moocs) Change Education?," in: *CHI '13 Extended Abstracts on Human Factors in Computing Systems*. Paris, France: ACM, pp. 2395-2398.
Sadigh, D., Seshia, S.A., and Gupta, M. 2012. "Automating Exercise Generation: A Step Towards Meeting the Mooc Challenge for Embedded Systems," *Proceedings of the Workshop on Embedded Systems Education (WESE), ESWeek*.
Singh, R., Gulwani, S., and Solar-Lezama, A. 2013. "Automated Feedback Generation for Introductory Programming Assignments," *SIGPLAN Not.* (48:6), pp. 15-26.
Spady, W.G. 1970. "Dropouts from Higher Education: An Interdisciplinary Review and Synthesis," *Interchange* (1:1), pp. 64-85.
Suddaby, R. 2006. "From the Editors: What Grounded Theory Is Not," *Academy of management journal* (49:4), pp. 633-642.

Terzi, M., Ferrario, M.A., and Whittle, J. 2011. "Free Text in User Reviews: Their Role in Recommender Systems," *Workshop on Recommender Systems and the Social Web at the 5th ACM International Conference on Recommender Systems (RecSys'11)*, Chicago, Illinois, USA: ACM, pp. 45-48.

Tinto, V. 1975. "Dropout from Higher Education: A Theoretical Synthesis of Recent Research," *Review of educational research* (45:1), pp. 89-125.

Tinto, V. 1987. *Leaving College: Rethinking the Causes and Cures of Student Attrition*. ERIC.

Train, K.E. 2003. *Discrete Choice Methods with Simulation*. Cambridge University Press.

Tresman, S. 2002. "Towards a Strategy for Improved Student Retention in Programmes of Open, Distance Education: A Case Study from the Open University Uk," *The International Review of Research in Open and Distance Learning* (3:1).

Urquhart, C., and Fernandez, W. 2013. "Using Grounded Theory Method in Information Systems: The Researcher as Blank Slate and Other Myths," *Journal of Information Technology* (28), 01/29/online, pp. 224–236.

Vaast, E., and Walsham, G. 2011. "Grounded Theorizing for Electronically Mediated Social Contexts," *European Journal of Information Systems* (22:1), pp. 9-25.

Vihavainen, A., Luukkainen, M., and Kurhila, J. 2012. "Multi-Faceted Support for Mooc in Programming," in: *Proceedings of the 13th annual conference on Information technology education*. Calgary, Alberta, Canada: ACM, pp. 171-176.

Willging, P.A., and Johnson, S.D. 2004. "Factors That Influence Students' Decision to Dropout of Online Courses," *Journal of Asynchronous Learning Networks* (8:4), pp. 105-118.

Xenos, M., Pierrakeas, C., and Pintelas, P. 2002. "A Survey on Student Dropout Rates and Dropout Causes Concerning the Students in the Course of Informatics of the Hellenic Open University," *Computers & Education* (39:4), Dec, pp. 361-377.

Xu, D., and Jaggars, S. 2013. "Adaptability to Online Learning: Differences across Types of Students and Academic Subject Areas." *CCRC* Working Paper No. 54. Retrieved September 5, 2013, from http://ccrc.tc.columbia.edu/media/k2/attachments/adaptability-to-online-learning.pdf

Yu, H.F., Huang, F.L., and Lin, C.J. 2011. "Dual Coordinate Descent Methods for Logistic Regression and Maximum Entropy Models," *Machine Learning* (85:1-2), Oct, pp. 41-75.