Replicating 21 Findings on Student Success in Online Learning

JUAN MIGUEL L. ANDRES, University of Pennsylvania
RYAN S. BAKER, University of Pennsylvania
GEORGE SIEMENS, University of Texas at Arlington
DRAGAN GAŠEVIĆ, University of Edinburgh
CATHERINE A. SPANN, University of Texas at Arlington

There has been a considerable amount of research over the last few years devoted towards studying what factors lead to student success in online courses, whether for-credit or open. However, there has been relatively limited work towards formally studying which findings replicate across courses. In this paper, we present an architecture to facilitate replication of this type of research, which can ingest data from an edX Massively Open Online Course (MOOC) and test whether a range of findings apply, in their original form or slightly modified using an automated search process. We identify 21 findings from previously published studies on completion in MOOCs, render them into production rules within our architecture, and test them in the case of a single MOOC, using a post-hoc method to control for multiple comparisons. We find that nine of these previously published results replicate successfully in the current data set and that contradictory results are found in two cases. This work represents a step towards automated replication of correlational research findings at large scale.

1. INTRODUCTION

Replication, the reproduction of a previous study in order to investigate the agreement between the current results and those of the original study (Brandt et al., 2014), is highly important in scientific research. A study can be deemed reproducible if an independent team is able to follow its published method as closely as possible from start to finish and obtain a result similar to, if not exactly the same as, the original result (Brandt et al., 2014). As such, replication is a critical step in the process of scientific inquiry, enabling researchers to better understand the reliability, validity, and merit of a study's findings.

However, despite the importance of replication studies, they remain rare in the social sciences, with only 1.07% of published psychology studies from 2007 to 2012 representing an attempt at replication (Makel, Plucker, & Hegarty, 2012, p. 537). Replication is even rarer in education research. A recent survey of the 100 education journals with the highest 5-year impact factor ratings in 2013 found that only 0.13% of the studies published involved replication (Makel & Plucker, 2014). There are several reasons for this; many educational research studies are difficult to reproduce due to issues of cost, as well as differences between populations and idiosyncrasies of the match between content and current instructional conditions.

That said, the problem of replication is more serious than simply a failure to conduct best practice. Instead, it leads to a surprisingly large proportion of spurious results being widely believed. One of the best estimates of how problematic the failure to replicate is was provided by the Open Science Collaboration (OSC, 2015), who replicated 100 experimental and correlational studies from three psychology journals. The study compared significance and effect sizes between the original studies and their replications. The study reported that 64% of the replication studies failed to obtain a statistically significant result. Beyond this, "replication effects were half the magnitude of original effects (OSC, 2015, p. 944)." This is a sobering finding, which brings to light the importance of replication research and the need to validate previous findings. Without replication, exploratory studies are taken as fact, which can have effects varying from useless to dangerous, depending on the scope of people it affects and the gravity of its effect.

However, a new source of data provides the opportunity to improve on the status quo in at least one area of education: online learning. While modern practice in randomized controlled trials often involves recruiting a large and representative sample (Glennerster & Takavarasha, 2013), and the recruitment and research processes are expensive to conduct at scale (Feuer, Towne, & Shavelson, 2002), recruiting and studying large samples is considerably less painful in online learning platforms already used at scale. Commercial platforms for K-12 education are used by tens or hundreds of thousands of students (cf. Koedinger & Corbett, 2006; Koedinger, McLaughlin, & Heffernan, 2010). Perhaps the largest opportunities for replication research, however, come from Massive Open Online Courses (MOOCs). MOOC platforms are used by millions of learners around the world who obtain free access to a wide variety of online course topics taught by professors from prestigious universities (Yuan & Powell, 2013). While MOOC populations are typically biased towards individuals living in developed countries who already have substantial educational attainment (Yuan & Powell, 2013), this limitation is surely not greater than the long-term reliance by researchers on subject pools of undergraduates enrolled in psychology courses at a small set of prestigious universities (Rozin, 2001).

Within this paper, we focus on research that attempts to predict student MOOC completion, i.e., obtaining a certificate for completing the course. Performance in MOOCs is typically measured by calculating a learner's assignment and test scores (Breslow et al., 2013). These scores are calculated at the end of the course and combined via a formula to determine whether or not a learner will receive a certificate and thus, complete the course. The completion cut-off is usually defined by the course instructor (cf. Belanger & Thornton, 2013; MOOC @ Edinburgh, 2013).

We focus on this type of research for several reasons. First, it is widely considered to be an area of significant concern for MOOCs. MOOCs have been criticized for their severely high attrition rates (Clow, 2013), with only about 3-10% of students successfully completing the MOOCs in which they register (Yang, Sinha, Adamson, & Rosé, 2013; Jordan, 2014). The process of attrition in MOOCs has been likened to a funnel of participation (Clow, 2013), where learners pass through the four stages of awareness, registration, activity, and progress, each stage characterized by severe drop-offs. In Clow's model, *awareness* occurs when potential participants learn about the MOOC. A small proportion of these potential participants then engage in *registration*, signing up to take the course. A small proportion of registrants enter the phase of *activity,* actively participating in the MOOC. Finally, only a small proportion of active registrants make *progress* at their learning within the MOOC or complete their intended course.

Second, failure to complete courses is a problem that is potentially actionable – it may be possible to design interventions that increase the proportion of students who succeed in MOOCs. For instance, in one study that sought to investigate forum participation, participants were randomly given different badges for posting in the course's discussion forums (Anderson, Huttenlocher, Kleinberg, & Leskovec, 2014). The study found that some of these badges eventually improved forum participation. In another study, a random sample of students who had *stopped-out*, i.e., stopped participating in a MOOC, were sent emails aimed at bringing them back to the MOOC. The students who received these intervention emails were significantly more likely to return to the class than students who did not receive the emails (Whitehill, Williams, Lopez, Coleman, & Reich, 2015).

Third, there is a considerable volume of published research on this problem, making it an attractive context to study replication in. To give just a few examples, Crossley and colleagues (2013) investigate the relationship between discussion forum features, such as the length and frequency of the students' posts within the forum, and MOOC completion. Wang (2014) examined the relationship between course completion and student motivation as reported in a pre-course survey. DeBoer and colleagues (2013) correlated course completion to the amount of time spent on different online course resources, such as time spent on the forums and time spent on assignments. Other studies have also looked at the relationship of completion with prior knowledge (Kennedy, Coffrin, de Barba, & Corrin, 2015), demographics and social network participation (Engle, Mankoff, & Carbrey, 2015), and interaction with the instructor (Hone & El Said, 2016). All these are evidence that, indeed, research concerning MOOC completion is an active area for researchers as well as practitioners and one in need of a replication study.

Despite the general agreement that course completion in MOOCs is an important problem, it is, however, important to acknowledge that course completion is not the only measure of success in MOOCs. A number of recent studies have begun looking into the longitudinal impact of MOOCs upon course completion (Wang, Paquette, & Baker, 2014; Radford, Coningham, & Horn, 2015). These studies recognize that post-MOOC success can be difficult to measure, as it will depend on the learner's own goals and domain, as many MOOC learners do not consider course completion to be their primary goal (Belanger & Thorton, 2013). Career advancement and being able to join and contribute to communities of practice have been cited among the primary goals of MOOC learners (Radford, Robles, Cataylo, Horn, Thorton, & Whitfield, 2014; Wang et al., 2014). However, due to the multiple operationalizations of post-MOOC success and the relatively sparser literature in this area, this article focuses on the more-studied question of what factors are associated with MOOC completion.

In the following sections, we discuss the research that is incorporated into our model. Next, we study the modeling framework and how it is used to study replication. This framework was developed using a production-system framework, which represents existing findings in a fashion that human researchers and practitioners can understand. The framework can be parametrically adapted to different contexts, where slightly different variations of the same findings may hold. We discuss the course and data set in which we examined these issues, and then detail which of the previous findings hold true within this data set, attempting to replicate 21 previously published findings. We conclude with a discussion of future work, and how the work presented here can serve as a template for a new type of replication research in education.

2. METHODOLOGY

2.1 Initial Data Set & Demographics

We analyzed the 21 previous findings within the context of data from the 2015 MOOC Big Data in Education MOOC (BDEMOOC), offered through edX by Teachers College, Columbia University. BDEMOOC covered the concepts and methods of the emerging field of educational data mining (Baker & Siemens, 2014), and was designed to be roughly equivalent to a graduate-level course. The MOOC had a total of 6,566 registrants. Of the cohort, 1,333 participants completed part or all of at least one assignment, 516 had at least 1 post in the discussion forum, and 166 completed the MOOC and earned a certificate.

Of the students registered, 1,088 participants took a pre-course survey, which contained questions about MOOC-specific motivational variables, such as familiarity with MOOCs as a platform and interest in the course content. The survey also included a set of questions geared towards the measurement of learner goal orientation (such as learning and performance goals), and academic efficacy (Wang, 2014). Of the survey respondents, 65% were male and 35% were female. A majority of these survey respondents fell within the age range of 25 to 44 years old (25-34 y/o: 32%, 35-44 y/o: 27%). Most of the respondents had either a 4-year college degree (27%), a master's degree (44%), or a doctoral degree (17%), and worked for a large non-profit (14%) or for-profit (13%) company in the education sector.

BDEMOOC spanned 8 weeks. Weekly sessions were composed of 5 to 7 lecture videos and a corresponding assignment requiring students to practice methods learned that week using spreadsheets and data mining tools. Assignments were created and presented to the students using the Cognitive Tutor Authoring Tools (Aleven et al., 2015). This framework offered step-by-step guidance to students, including both hints and messages regarding specific misconceptions, as the students attempted to solve the assignment problems. The course also assigned weekly collaborative assignments that encouraged discussion among students about what they had learned that week. Students and teaching staff participated in forum discussions accompanying weekly sessions. In order to earn a certificate in the MOOC, students needed to earn a final grade of at least 70%. Final grades were calculated by averaging the 6 assignments with the highest scores out of the 8 offered to students.

With its intelligent-tutor based assignments, weekly collaborative assignments, and high level of expertise and content, BDEMOOC was a somewhat atypical MOOC; any findings which replicate from more standard MOOCs can be thought to be quite robust.

2.2 Research Synthesis

The initial step in studying the replicability of findings in MOOCs was to compile a list of previous findings. MOOC literature is still in its infancy, with relatively few publications occurring before 2010 (see discussion in McAuley, Stewart, Siemens, & Downes, 2010). As such, the initial search conducted examined only work published in and after 2010. Within this first pass on conducting multiple replications at once, we focused on findings that related some aspect of the student's attributes and behaviors to course completion. For example, studies that investigated characteristics other than those of the students (i.e., platform, course, or university characteristics) and studies that investigated outcomes other than engagement and course completion were dropped from the analysis. During the literature review, we encountered findings that required the use of specific analytical tools. Where possible, we contacted the researchers and obtained copies of these analytical tools; analyses requiring tools not readily available to the researchers were dropped from the review and set aside for future work. The study focused on behaviors seen in the system and motivational surveys for which data was available. From this search, 68 papers were reviewed; the findings investigated in this study were drawn from 8 published articles. Twenty-one findings in total were obtained and analyzed. It is important to note that this paper does not attempt to be fully comprehensive in analyzing predictors of course completion; by explicitly studying these 21 findings, however, this paper represents the largest-scale replication analysis (in terms of number of findings studied) that we are currently aware of in the field of education.

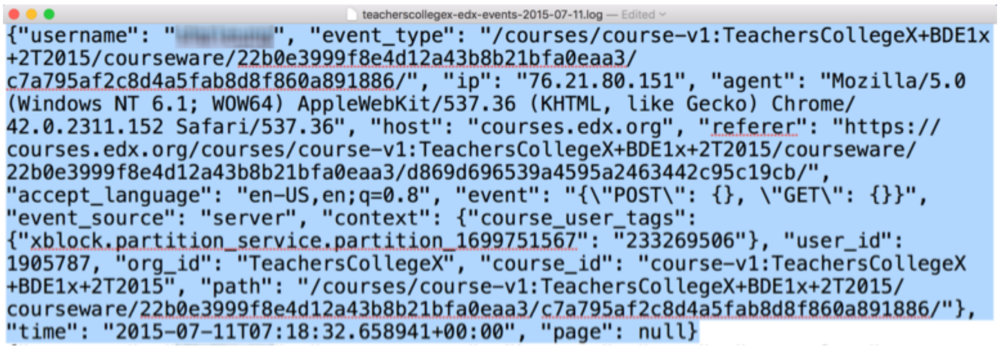
The study included three papers that looked at student attributes derived from pre-course survey responses. One paper found that participants taking the MOOC for credit were more likely to complete the course (Clow, 2013). Other papers found that being motivated by course content and having high self-efficacy (Wang, 2014), as well as being certain one would master the skills to be taught in the MOOC (Wang & Baker, 2015) were associated with completion.

The current study also included five papers that investigated different student features and behaviors within the discussion forums. These papers found that writing longer posts (Crossley et al., 2015; Yang et al., 2013), writing more often (Crossley et al., 2015; Yang, Wen, Howley, Kraut, & Rosé, 2015), starting a thread, receiving replies on one's thread, and replying to others' threads (Ramesh, Goldwasser, Huang, Daumé, & Getoor, 2013; Yang et al., 2013; Yang et al., 2015), and just generally spending more time in the forums (DeBoer, Ho, Stump, Pritchard, Seaton, & Breslow, 2013) were significantly associated with course completion. Crossley and colleagues (2015) also found a range of linguistic features associated with successful completion of MOOCs, such as the use of more concrete and more sophisticated words, and the use of more bigrams and trigrams.

The findings from the Wang (2014), Wang & Baker (2015), and Crossley et al. (2015) studies all came from the previous iteration of BDEMOOC on Coursera. In his study introducing the funnel of participation in MOOCs, Clow (2013) conducted his investigation on data from three open, online learning environments: iSpot, a social learning community geared towards learning about nature observations, Cloudworks, a professional learning community for educators and educational researchers, and openED, a business and management MOOC (p. 186). The two studies from Carnegie Mellon University (Yang et al., 2013; Yang et al, 2015) explore MOOC dropout rates, confusion, and forum features extracted from two Coursera MOOCs: one on Algebra and the other on Microeconomics. The study by Ramesh and colleagues (2013) evaluated the models they created using data from a Coursera MOOC entitled *Surviving Disruptive Technology*, which had 1,665 participants engaged in the forums, and 826 completers. Finally, the study by De Boer and colleagues (2013) explored the impact of resource use and the students' background characteristics on achievement within an edX MOOC entitled *Circuits and Electronics*.

2.3 edX Interaction Log Data Scrub

Log data were obtained from BDEMOOC, representing 1,252,306 student actions within the system. The raw edX interaction logs present data in an attribute-value object format, an example of which can be seen in Figure 1. Each mouse click within the MOOC generates one transaction in the logs. Each transaction is treated as an object, and each object has multiple attributes (e.g., username, timestamp, event source). This format allows for the logging of hierarchical attributes (i.e., attributes within attributes) on multiple sublevels, which can impede analysis. As such, the raw edX interaction logs required pre-processing in order to get into a more analyzable format. A parser was developed in order to conduct this pre-processing. The parser accepts as input any number of log files, and returns as output a single tab-delimited text file containing all transactions. Tab was chosen as the delimiting character because discussion forum post contents can contain any number of symbols in them, like the comma and semicolon, which are the more common delimiters. Pre-processing the logs aided in the next step of feature engineering. This parser can now be re-used with other edX courses.



```
{"username": "[redacted]", "event_type": "/courses/course-v1:TeachersCollegeX+BDE1x
+2T2015/courseware/22b0e3999f8e4d12a43b8b21bfa0eaa3/
c7a795af2c8d4a5fab8d8f860a891886/", "ip": "76.21.80.151", "agent": "Mozilla/5.0
(Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/
42.0.2311.152 Safari/537.36", "host": "courses.edx.org", "referer": "https://
courses.edx.org/courses/course-v1:TeachersCollegeX+BDE1x+2T2015/courseware/
22b0e3999f8e4d12a43b8b21bfa0eaa3/d869d696539a4595a2463442c95c19cb/",
"accept_language": "en-US,en;q=0.8", "event": "{\"POST\": {}, \"GET\": {}}",
"event_source": "server", "context": {"course_user_tags":
{"xblock.partition_service.partition_1699751567": "233269506"}, "user_id":
1905787, "org_id": "TeachersCollegeX", "course_id": "course-v1:TeachersCollegeX
+BDE1x+2T2015", "path": "/courses/course-v1:TeachersCollegeX+BDE1x+2T2015/
courseware/22b0e3999f8e4d12a43b8b21bfa0eaa3/c7a795af2c8d4a5fab8d8f860a891886/"},
"time": "2015-07-11T07:18:32.658941+00:00", "page": null}
```

Figure 1. *Example of raw edX interaction log file*

2.4 Feature Engineering

The next step was to operationalize the attributes and behaviors investigated in the findings examined in this study. In order to replicate previous findings on the current data set, this step required mapping and replicating the variables seen in those previous papers within the BDEMOOC data.

Feature engineering and the next step of building respective production rules were done simultaneously on an iterative basis. That is, the variable found in one finding were engineered and the finding was turned into a production rule for execution. Once the production rule could be run and analyzed (see next section), the variables used in the next finding were engineered and the finding was turned into a production rule for execution, and so on.

2.5 Production Rule System and Validation

The current study conducted its replication analysis through the development of a production-system framework that represented existing findings in a fashion that human researchers and practitioners can easily understand, but which can be parametrically adapted to different contexts, where slightly different variations of the same findings may hold.

The production rule system was built on Jess, an expert system programming language (Friedman-Hill, 2002). All findings were programmed into if-else production rules following the format, "If a student who is <`attribute`> does <`operator`>, then <`outcome`: completes or does not complete>." Attributes are pieces of information about a student. Operators are actions a student does within the MOOC. Outcomes are, in the case of this study, whether or not the student in question completed the MOOC. Using this production rule format, this study was able to capture the set of student attributes and actions and combinations of them, and relate it to whether the student completed or not. Not all production rules had both attributes and operators. Production rules that look at survey responses, for example, had only attributes (e.g., whether or not the participant says they are likely to follow the course pace) and outcomes (i.e., whether or not the participant completed the MOOC). Conversely, some production rules involving forum posts had only operators (e.g., whether or not the participant posted on the forums more frequently than the average) and outcomes. The production rule approach was chosen for its feasibility, its ability to directly represent findings, and its high degree of interpretability, attributes that previously made this approach common in efforts to make human-understandable models and theories of cognition (cf. Anderson, Matessa, & Lebiere, 1997; Laird, Newell, & Rosenbloom, 1987).

Table 1 presents a list of the previous findings we attempted to replicate within the BDEMOOC data set, divided into three groups: findings involving data drawn from pre-course survey responses (Rules 1-6), findings involving data drawn from clickstream data concerning time spent on specific activities within the MOOC (Rules 7-8), and findings involving data drawn from the participants' discussion forum posts (Rules 9-21).

Table 1. *Findings on MOOC completion that were included in the current replication study.*

#	If	Then	Source
1	Participant indicated they are taking the course for credit	Likely to earn certificate	Clow, 2013
2	Participant indicated that their interest in MOOC features is a motivation in taking the course	Not likely to earn a certificate	Wang, 2014; Wang & Baker, 2015
3	Participant indicated that their interest in course content is a motivation in taking the course	Likely to earn certificate	Wang, 2014
4	Participant indicated that they are certain they will master skills to be taught in the course	Likely to earn certificate	Wang & Baker, 2015
5	Participant indicated they had high self-efficacy	Likely to earn certificate	Wang, 2014
6	Participant indicated that they were likely follow the pace set by the course instructor	Likely to earn certificate	Wang & Baker, 2015
7	Participant spends more time in forums than average	Likely to earn certificate	DeBoer et al., 2013
8	Participant spends more time on assignments than average	Likely to earn certificate	DeBoer et al., 2013
9	Participant's average length of posts is longer than the course average	Likely to earn certificate	Crossley et al., 2015; Yang et al., 2013
10	Participant posts on the forums more frequently than average	Likely to earn certificate	Crossley et al., 2015; Yang et al., 2015
11	Participant responds more frequently to other participants' posts than average	Likely to earn certificate	Yang et al., 2013
12	Participant starts a thread	Likely to earn certificate	Yang et al., 2013
13	Participant starts threads more frequently than average	Not likely to earn certificate	Yang et al., 2015
14	Participant has respondents on threads they started	Likely to earn certificate	Ramesh et al., 2013
15	Participant has respondents on threads they started greater than average	Likely to earn certificate	Ramesh et al., 2013
16	Participant uses more concrete words than average	Likely to earn certificate	Crossley et al., 2015
17	Participant uses more bigrams than average	Likely to earn certificate	Crossley et al., 2015
18	Participant uses more trigrams than average	Likely to earn certificate	Crossley et al., 2015
19	Participants uses more meaningful words than average	Likely to earn certificate	Crossley et al., 2015
20	Participant uses more sophisticated words than average	Likely to earn certificate	Crossley et al., 2015
21	Participant uses a wider variety of words than average	Likely to earn certificate	Crossley et al., 2015

Some production rules were parameterized, for example to determine cut-offs. In these cases, grid search was used to find the variant with the largest effect size, as in (Baker, Gowda, & Corbett, 2011). For example, in the production rule that looked at the participants' intent to follow the pace set by the instructor (Table 1, Rule 4), participants gave answers on a scale of 0 to 5. Instead of considering only scores of 5, $\chi^2(1, N{=}1088) = 0.044$, $p = 0.834$, or only both scores of 4 and 5, $\chi^2(1, N{=}1088) = 0.026$, $p = 0.872$, as representing student certainty, the final parameter looked at scores of 3 and above, $\chi^2(1, N{=}1088) = 4.704$, $p = 0.030$. The same threshold was used for the production rule on self-efficacy (Table 1, Rule 5). In the case of Rules 14 and 15, Rule 14 was the original finding, i.e., participants having respondents on their threads in the discussion forum. However, when the production rule did not return significant findings, we created Rule 15 as a variation of the rule, i.e., participants having more respondents on their threads than average.

Each production rule returned two counts: 1) the confidence (Agrawal, Imielinski, & Swami, 1993), or the number of participants who fit the rule (i.e., meets both the `if` and the `then` statements), and 2) the conviction (Brin, Motwani, Ullman, & Tsur, 1997), the production rule's counterfactual, or the number of participants who did not fit the rule, but still meet the rule's outcome (i.e., does not meet `if` statement, but meets the `then` statement). For example, in the production rule, "If a student posts more frequently than the average student, then they are more likely to complete the MOOC," the two counts returned will be the number of participants that posted more than the average and completed the MOOC, and the number of participants who posted less than average *but still* completed the MOOC.

A chi-square test of independence was conducted on each pair of results, i.e. comparing the confidence to the conviction. The chi-square test was used in order to determine whether the two values are significantly different from each other, and in doing so, determine whether the production rule or its counterfactual significantly generalized to the current data set. Since 21 tests were conducted (one per finding), Benjamini & Hochberg's (1995) post-hoc correction method was used to weed out findings that were likely to be spurious, due to running many tests. This method produces a substitute for p-values, termed q-values, driven by controlling the proportion of false positives obtained via a set of tests. Whereas a p-value expresses that 5% of all tests may include false positives, a q-value indicates that 5% of significant tests may include false positives. As such, this method does not guarantee each test's significance, but guarantees a low overall proportion of false positives, preventing the substantial over-conservatism found in methods such as the Bonferroni correction (cf. Perneger, 1998).

3. RESULTS

The analysis was comprised of the replication of 21 findings relating to participant characteristics or behavior, and MOOC completion. Six production rules looked at pre-course survey responses. These rules were only applied to the 1,088 participants who had completed the survey. Participants who had failed to do so were excluded from the analyses of these production rules. Fourteen production rules examined discussion forum behaviors and content features. Only the 516 participants who had posted at least once in the forums were included in the analyses of these rules. Finally, one production rule looked at total time spent on assignments. Only the 1,333 participants who started at least one assignment were included.

The 21 production rules can be found in Table 2, where each row presents one previously published finding, broken down into an if-then production rule, followed by whether the attempt to replicate this finding was statistically significant, and a reference to the published previous work that this finding was drawn from. The significant production rules (after controlling for multiple comparisons) are marked with an asterisk. This signifies that a previously published finding replicated. Statistically significant counterfactuals are marked with double asterisks. This signifies that the *opposite* of the previously published result was obtained (in this case, the actual result for this data set is listed in the table, rather than the original finding).

Table 2. *Production rule analysis results*. Statistically significant results in agreement with previous findings denoted by *. Statistically significant results representing the opposite of previous findings denoted by **. Statistically significant results representing re-parameterized versions of the previous findings have their sources denoted by ***.

#	If	Then	Chi-square	Source
1	On survey: Taking for credit	Likely to earn certificate	$\chi^2(1, N=1088) = 0.350$, $p = 0.554$	Clow, 2013
2	On survey: Interested in MOOC features	Not likely to earn a certificate	$\chi^2(1, N=1088) = 1.467$, $p = 0.226$	Wang, 2014; Wang & Baker, 2015

3	On survey: Interested in course content	Likely to earn certificate	$\chi^2(1, N=1088) = 2.582$, $p = 0.108$	Wang, 2014
4	On survey: Certain will master skills to be taught in course	Likely to earn certificate	$\chi^2(1, N=1088) = 4.704$, $p = 0.030$	Wang & Baker, 2015***
5	On survey: Has high self-efficacy	Likely to earn certificate	$\chi^2(1, N=1088) = 4.608$, $p = 0.032$	Wang, 2014***
6	On survey: Will likely follow pace*	Likely to earn certificate	$\chi^2(1, N=1088) = 12.472$, $p < 0.001$	Wang & Baker, 2015
7	Participant spends more time in forums than average*	Likely to earn certificate	$\chi^2(1, N=516) = 136.814$, $p < 0.001$	DeBoer et al., 2013
8	Participant spends more time on assignments than average*	Likely to earn certificate	$\chi^2(1, N=1333) = 50.053$, $p < 0.001$	DeBoer et al., 2013
9	In forums: Length of posts is longer than average	Likely to earn certificate	$\chi^2(1, N=516) = 3.875$, $p = 0.049$	Crossley et al., 2015; Yang et al., 2013
10	In forums: Number of posts is greater than average*	Likely to earn certificate	$\chi^2(1, N=516) = 102.728$, $p < 0.001$	Crossley et al., 2015; Yang et al., 2015
11	In forums: Number of responses to others is greater than average*	Likely to earn certificate	$\chi^2(1, N=516) = 74.214$, $p < 0.001$	Yang et al., 2013
12	In forums: Starts thread	Likely to earn certificate	$\chi^2(1, N=516) = 0.004$, $p = 0.951$	Yang et al., 2013
13	In forums: Starts thread less frequently than average**	Not likely to earn certificate	$\chi^2(1, N=516) = 63.577$, $p < 0.001$	Yang et al., 2015
14	In forums: Has respondents on thread	Likely to earn certificate	$\chi^2(1, N=516) = 2.067$, $p = 0.150$	Ramesh et al., 2013
15	In forums: Has respondents on thread greater than average*	Likely to earn certificate	$\chi^2(1, N=516) = 52.479$, $p < 0.001$	Ramesh et al., 2013***
16	In forums: Uses more concrete words	Likely to earn certificate	$\chi^2(1, N=516) = 3.537$, $p = 0.060$	Crossley et al., 2015
17	In forums: Uses more bigrams than average*	Likely to earn certificate	$\chi^2(1, N=516) = 8.357$, $p = 0.004$	Crossley et al., 2015
18	In forums: Uses more trigrams than average*	Likely to earn certificate	$\chi^2(1, N=516) = 9.580$, $p = 0.002$	Crossley et al., 2015
19	In forums: Uses less meaningful than average**	Likely to earn certificate	$\chi^2(1, N=516) = 13.821$, $p < 0.001$	Crossley et al., 2015
20	In forums: Uses more sophisticated words than average*	Likely to earn certificate	$\chi^2(1, N=516) = 11.643$, $p < 0.001$	Crossley et al., 2015
21	In forums: Uses more variety of words than average	Likely to earn certificate	$\chi^2(1, N=516) = 2.838$, $p = 0.092$	Crossley et al., 2015

As shown in the table, only 9 of the 21 previous findings were replicated in the current data set. Two of the 21 previous findings actually had their counterfactual come out statistically significant, i.e., they had the opposite result as in previously published literature.

4. DISCUSSION

Nine production rules replicated significantly within the current data. From the pool of survey-related rules, only Rule 6, which states that if students intend to follow the pace set by the instructor, then they are likely to complete the course and earn a certificate, significantly replicated. This rule was drawn from a study that analyzed survey and log data from the previous iteration of BDEMOOC (Wang & Baker, 2015), which looked at the participants' motivations in taking the course. Upon inspecting the current data set further, we found that many participants had accessed videos and resources for only a limited number of specific modules within the course. This behavior implies that these participants had enrolled with the goal of gathering resources on specific topics they were interested in, rather than covering the course comprehensively. This may have been the reason that participants who had enrolled with the motivation to complete the course, and had expressed so in the pre-course survey, were more likely to complete.

Rules 7 and 8 looked at the total amounts of time spent in the forums and on assignments, respectively. Both rules replicated significantly within the current data set, agreeing with previous findings that spending more time on these activities is characteristic of course completion. These finding indicate that spending more time with the course content, either through engaging in or observing the discussions in the forums or through engaging with the course assignments, is associated with completion. This is likely for multiple reasons. More motivated participants are likely to spend more time and are also more likely to complete. But to the degree that the materials are themselves beneficial to learning (the goal of educational materials), spending more time with them may increase the change of successful performance and completion. In an environment such as MOOCs, where students have the freedom to disengage at any point in the course, knowing that time spent in the discussion forums is associated with remaining engaged till completion indicates that attention should be spent on designing engaging and positive discussion forum experiences.

Rules 10, 11, and 15 look at posting behaviors: the rules state that if students post more frequently than average, respond to other students' threads more frequently than average, and have more respondents on their own threads than average, they are more likely to earn a certificate. Participants posted frequently on the discussion forums for a number of reasons. For instance, many participants posted to ask questions, whether about the lecture content itself or about running analyses on the course's suggested tool or to share additional resources to augment the weekly modules. In this specific MOOC, question posts almost always got responses from either the course instructor, TA, or other participants. These posts allowed participants to discuss the weekly modules and deepen their understanding of the content outside of the lecture videos and quizzes. These interactions, and certainly, the behavior of posting and responding frequently on the forums implies an understanding of the topics being discussed or, at the very least, an interest to learn. This pattern may be different on MOOCs with less engaged instructors, an area for future investigation.

Finally, Rules 17, 18, and 20 look at linguistic features that were derived from the students' discussion forum posts, analyzed using the Tool for the Automated Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015) and the Tool for the Automatic Analysis of Cohesion (TAACO; Crossley, Kyle, & McNamara, in press). The rules state that if students use more bigrams (combinations of two words) than average, more trigrams (combinations of three words) than average, or more sophisticated words than average in their posts, they are more likely to complete. Lexical sophistication involves the "depth and breadth of lexical knowledge (Kyle & Crossley, 2015, p.759);" the measure used here was derived from word occurrence in a range of large-scale corpuses. Each of these indices has been found to be correlated with several other measures of writing quality (Crossley et al., 2012). The findings in Rules 17, 18, and 20, thus, imply that lengthier and more sophisticated posts are associated with remaining engaged in the course. While some students may have used sophisticated language in off-topic discussion (cf. Comer, Baker, & Wang, 2015), generally more sophisticated language may be associated with positive understanding of the course content. The three features are drawn from a longer list of linguistic features that were correlated with course completion in the original study (Crossley et al., 2015).

Two production rules were significant, but in the reverse direction from what was reported in the original papers they came from. Rule 13 was drawn from a study where annotated confusion scores were used to predict a number of forum and confusion features, including the number of forum threads initiated (Yang et al., 2015). Each forum post was given a 1-4 Likert scale confusion score by 5 coders with reasonably high inter-coder reliability, and the average was used as each respective post's confusion grade. However, within this analysis, they determined that if students started threads more frequently than average, then they were less likely to complete and earn a certificate (Rule 13), and that students who make more posts are more likely to obtain a certificate (Rule 10, also seen in Crossley et al., 2015). In this paper, we do not replicate their hand-coded confusion variable for feasibility reasons, but examine these two additional findings (Rule 10 and Rule 13) from that paper. In our analysis, we found that starting threads *less* frequently than average is significantly related to a lesser likelihood of course completion. Students start threads for reasons other than confusion, for instance due to being interested in the subject matter. Rule 19 was part of a set of linguistic features that were correlated with course completion (Crossley et al., 2015). The rule originally stated that if students used more meaningful words (i.e., words with higher association to other words) in their discussion forum posts, they were more likely to complete the course. Our analysis, however, found that using *fewer* meaningful words was significantly related to course completion. This may be because words graded as "meaningful" in TAALES may actually be less relevant to course content than other words. TAALES grades meaningfulness based on how related a word is to other words (Kyle & Crossley, 2015, p.762). As such, words like "computer" are likely to be more *meaningful* compared to field-specific terms like "regression". Using fewer meaningful words could thus mean that participants were using field-specific terms in their discussion posts. Being able to converse using field-specific terms would imply better understanding of the content being taught in the course.

The 11 other production rules were not statistically significant, indicating a failure to replicate. Interesting among these findings is that most of the production rules that were based on pre-course survey responses and linguistic features did not replicate in the current data set. They are interesting because most of these production rules were drawn from the three studies that used data from the previous iteration of BDEMOOC (Wang, 2014; Wang & Baker, 2015; Crossley et al., 2015). That is, even with the same intended audience, taught with the same learning design, and following the same progression of content, previously discovered findings did not turn up significant in the second iteration of the course. This finding is surprising, to say the least. One possibility is that the difference between running a MOOC in edX and Coursera is larger than previously anticipated. There are several differences in the design of these two platforms, such as better support for complex assignments in edX, and more sophisticated discussion forum threading in Coursera. Additionally, there may have been differences in the population taking these MOOCs, as Coursera reaches a wider, more general audience. Additional evidence that these populations were different comes from the lack of posts that were abusive towards the instructor in the edX run of this course, unlike the Coursera iteration (cf. Comer et al., 2015). Still, the lack of replication of multiple findings across such similar courses is a warning sign on how far MOOC research findings should be treated as generalizing. This result further stresses the importance of conducting replication studies within MOOC research.

5. IMPLICATIONS FOR THE DESIGN OF MOOCS

We believe that this paper's findings can aid in discussion of how to best support instructors and designers of MOOCs, providing a basis for future design and intervention by contributing to the literature on what findings apply and do not apply across different courses. Research on MOOCs is in its infancy, and a better understanding of how broadly findings apply may be of value.

The pre-course surveys that were administered in BDEMOOC asked participants about their motivations going into the MOOC. As mentioned earlier in the discussion, researchers have increasingly been looking into success measures of MOOCs beyond just course completion, particularly since not all learners have the goal of completing their course. As such, data on whether a participant indicates that they are likely to follow the pace of the course may be useful for distinguishing participants who want to complete the course from those who only wish to access specific course resources. Thus, based on this finding replicating within the current data set, course instructors can potentially use this survey response as an early identifier of participants whose goal is to complete the course.

The second set of replicated results find that participants who spend longer periods of time in either the discussion forums or on their assignments are likely to complete the course. Based on these findings, course instructors can begin considering how to encourage participants in their courses to spend more time on these two components. Instructors can encourage more active participation in the discussion forums (which will be discussed in greater detail below) or modify the course design to encourage participants to spend more time with the assignments. One possible modification could be the embedding of an assignment rubric that promotes self-reflection. This design could encourage participants to think more about the assignment and their performance, potentially leading to deeper understanding and better performance.

The final set of replicated results look at posting behaviors in the discussion forums. There is a growing body of literature that investigates the role of the forums in course completion. These studies look at the type of participation, such as passively reading the posts of others or actively engaging in conversation, and the language used. Based on the replicated findings indicating which posting behaviors are correlated significantly with course completion, course instructors can look into ways of encouraging participants to engage more actively in the forums. It will be valuable to encourage students both to ask questions and to work towards resolving other participants' questions, as both behaviors have been found to be associated with better student outcomes. Some MOOCs have begun requiring participation in the forums or have implemented *reputation scores*, where posts are voted up or down by other course participants. In fostering conversations in the forums, course instructors can also work towards encouraging conversations around course content, and curbing off-topic conversations.

6. CONCLUSION & NEXT STEPS

In this paper, we investigate the degree to which previously published findings on MOOC course completion replicate in new data. This was achieved through the development of a production system framework that was used to attempt the replication of 21 previously published findings on MOOC completion on a new data set. These 21 productions rules were drawn from 8 studies that sought to address the high attrition rate in MOOCs. Of these 21 findings, 9 were successfully replicated in the current data set (2 were statistically significant in the opposite direction). Through the analysis conducted, this study contributes to the slowly growing literature on replication in the field of education research. It is our hope that research of this nature can eventually result in faster and easier replication of published findings, at scale. One limitation to this study is that it is only conducted in one specific MOOC. However, as mentioned earlier, BDEMOOC was a somewhat atypical MOOC, and any findings which replicate from more standard MOOCs can be thought to be quite robust. In general, we will have more evidence on these findings when they are replicated in a greater number of MOOCs.

The study also contributes to the more efficient analysis of edX data through the creation of the first version of a pre-processing parser. The parser was developed in order to transform raw edX logs into tab-delimited text files, a format that is easier to both understand and analyze. edX and other researchers interested in using and analyzing edX data will be able to use the parser on edX data. We anticipate that some minor modifications will be needed by the parser in order to accept additional log syntax not present in the current data set.

Our next steps include extending our work published here in several ways. First, we plan to expand the current set of variables being modeled, both in terms of predictor (independent) variables and outcome (dependent) variables. Our first efforts do not yet include findings involving data from performance on assignments or behavior during video-watching, two essential activities in MOOCs which have been extensively researched in the last three years. To accomplish this goal, we intend to conduct a more comprehensive literature review. The findings in published papers can then be turned into production rules for replication on the current data set.

Second, we plan to expand to a greater range of data. Initially, we plan to apply the production rules to data from other edX courses. This should be a straightforward process, as the pre-processing parser was built to accept edX-format data. Once the pre-processed data has undergone feature engineering, the production rule system should execute seamlessly. With a large pool of courses, we can go beyond simple replication to studying how factors like course design, target and actual population, domain, and instructor pedagogy influence the applicability of these findings.

Eventually, we intend to expand to data from different online learning platforms. More resources will be needed for the creation of pre-processing parsers for each platform, if none are already available or if log data is not already in an analyzable format (in general, this task would be facilitated by the adoption of a logging standard such as the MoocDB standard proposed by Veeramachaneni, Dernoncourt, Taylor, Pardos, & O'Reilly, 2013). This will enable us to study the findings we have seen more generally still, studying how the different design features of different platforms drive differences in the factors associated with student success.

The analysis conducted in this study mined the features that were tested for replicability from previous literature on engagement and completion in MOOCs. These findings were then transformed into "if-then" production rules, which were fed into an expert system that determined the percentage of participants who met both condition and result of each rule. This approach, by definition, is bottom-up in that production rules characterized student behaviors in the MOOC as a means of evaluating course completion.

An alternative approach to looking at this phenomenon is through the lens of the structural learning theory (SLT; Scandura, 2003, 2014). SLT starts by assuming what is necessary for a student to do and learn in order to successfully interact with a learning environment. In the case of this study, SLT would look at steps required towards completing a MOOC course. Each operation within this framework is broken down into a sequence, condition, or iteration of actions. The resulting model is a representation of what a student would have to do or learn to complete any given task and at any point during the MOOC. We believe that both approaches are valid and reasonable, and may each serve as a complement to the other.

The long-term goal of this program of research is to take the initial steps towards building a theory on student success in online learning that can aid in supporting learners across different platforms and contexts. In order to be optimally useful and generative, next-generation theory on online learning needs to be able to recognize varied aspects of the learner and their behavior, and what to do in response to this information. Or, as suggested by Scandura (2014, p. 237), "students with different degrees of expertise need different kinds of help at various times during the course of learning." As such, tracking students' progress will be essential to providing support and instruction adapted to individual needs (e.g., Scandura, 2007). Not only will this theory identify potential predictors of student success, but it will also help identify possible moderating and mediating roles some variables may play in associations between predictors and success. Ultimately, developing optimal designs for learning support involves answering the question, "What should we do, when, and for who?" It is not necessary to start from scratch in determining this; there is already a considerable number of findings relevant to the factors and behaviors associated with student success in online learning. A model that identifies where these findings do and do not apply would be a useful step towards developing a universally-applicable theory of online learning, one that would both expand understanding and improve student outcomes.

ACKNOWLEDGEMENTS

We would like to thank Elle Wang of Teachers College, Columbia University for helping us work with the BDEMOOC data, Diyi Yang of Carnegie Mellon University for her helpful comments on the CMU studies cited in this paper, and Scott Crossley of Georgia State University for giving us access to the linguistic tools used in his papers.

REFERENCES

1. Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Associations between Sets of Items in Massive Databases. In *Proceedings of the ACM-SIGMOD Int'l Conference on Management of Data* (pp. 207-216).
2. Aleven, V., Sewall, J., Popescu, O., Xhakaj, F., Chand, G., Baker, R., Wang, E., Siemens, G., Rosé, C. P., Gašević, D. (2015). Intelligent tutoring systems and MOOCs: The beginning of a beautiful friendship*?. Proceedings of the 17th International Conference on Artificial Intelligence in Education* (pp. 525-528).
3. Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2014, April). Engaging with massive online courses. In *Proceedings of the 23rd international conference on World wide web* (pp. 687-698). ACM.
4. Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, *12*(4), 439-462.
5. Baker, R. S., Gowda, S. M., & Corbett, A. T. (2011, June). Towards predicting future transfer of learning. In *International Conference on Artificial Intelligence in Education* (pp. 23-30). Springer Berlin Heidelberg.
6. Baker, R., & Siemens, G. (2014) Educational data mining and learning analytics. In Sawyer, K. (Ed.) *Cambridge Handbook of the Learning Sciences: 2nd Edition*, pp. 253-274.
7. Belanger, Y., & Thornton, J. (2013). Bioelectricity: A Quantitative Approach Duke University's First MOOC.

8. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.
9. Brandt, M. J., Ijzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J.A., Perugini, M., Spies, J.R., & van't Veer, A. (2014). The replication recipe: What makes for a convincing replication?. *Journal of Experimental Social Psychology*, *50*, 217-224.
10. Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997, June). Dynamic itemset counting and implication rules for market basket data. In *ACM SIGMOD Record* (Vol. 26, No. 2, pp. 255-264). ACM.
11. Clow, D. 2013. MOOCs and the funnel of participation. In Proceedings of the Third International Conference on Learning Analytics and Knowledge (pp. 185-189). ACM.
12. Comer, D., Baker, R., Wang, Y. (2015) Negativity in Massive Online Open Courses: Impacts on Learning and Teaching. *InSight: A Journal of Scholarly Teaching, 10*.
13. Crossley, S. A., Kyle, K., and McNamara, D. S. (in press). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods.*
14. Crossley, S., McNamara, D. S., Baker, R., Wang, Y., Paquette, L., Barnes, T., & Bergner, Y. (2015). Language to Completion: Success in an Educational Data Mining Massive Open Online Class. *International Educational Data Mining Society*.
15. DeBoer, J., Ho, A., Stump, G.S., Pritchard, D.E., Seaton, D. and Breslow, L., 2013. Bringing student backgrounds online: MOOC user demographics, site usage, and online learning. engineer, 2, pp.0-81.
16. Engle, D., Mankoff, C., & Carbrey, J. (2015). Coursera's introductory human physiology course: Factors that characterize successful completion of a MOOC. *The International Review of Research in Open and Distributed Learning, 16(2).*
17. Friedman-Hill, E. (2002). Jess, the expert system shell for the java platform.*USA: Distributed Computing Systems*.
18. Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational researcher*, *31*(8), 4-14.
19. Glennerster, R., & Takavarasha, K. (2013). *Running randomized evaluations: A practical guide*. Princeton University Press.
20. Hone, K. S., & El Said, G. R. (2016). Exploring the factors affecting MOOC retention: a survey study. *Computers & Education.*
21. Jordan, K. (2014). Initial trends in enrolment and completion of massive open online courses. *The International Review of Research in Open and Distributed Learning*, *15*(1).
22. Kennedy, G., Coffrin, C., de Barba, P., & Corrin, L. (2015, March). Predicting success: how learners' prior knowledge, skills and activities predict MOOC performance. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 136-140). ACM.
23. Koedinger, K. R., & Corbett, A. (2006). Cognitive tutors. *The Cambridge handbook of the learning sciences*, 61-77.
24. Koedinger, K. R., McLaughlin, E. A., & Heffernan, N. T. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational Computing Research*, *43*(4), 489-510.
25. Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly, 49* (4), 757-786.
26. Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial intelligence*, *33*(1), 1-64.

27. Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty replication in the education sciences. *Educational Researcher*, 0013189X14545513.
28. Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research how often do they really occur?. *Perspectives on Psychological Science*, *7*(6), 537-542.
29. McAuley, A., Stewart, B., Siemens, G., & Cormier, D. (2010). The MOOC model for digital practice.
30. Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.
31. Perneger, T.V. (1998). What's wrong with Bonferroni adjustments. British Medical Journal, 316, 1236-1238.
32. Radford, A. W., Coningham, B., & Horn, L. (2015). MOOCs: Not Just for College Students—How Organizations Can Use MOOCs for Professional Development. *Employment Relations Today*, *41*(4), 1-15.
33. Radford, A. W., Robles, J., Cataylo, S., Horn, L., Thornton, J., & Whitfield, K. E. (2014). The employer potential of MOOCs: A mixed-methods study of human resource professionals' thinking on MOOCs. *The International Review of Research in Open and Distributed Learning*, *15*(5).
34. Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H. & Getoor, L. (2013, December). Modeling learner engagement in MOOCs using probabilistic soft logic. In NIPS Workshop on Data Driven Education (Vol. 2, pp. 1-7).
35. Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, *5*(1), 2-14.
36. Scandura, J.M. (2003). US Patent No. 8,750,782 *Method for Building Highly Adaptive Instruction*, June 10, 2014. Continuation in Process.
37. Scandura, J. M. (2007). Knowledge representation in structural learning theory and relationships to adaptive learning and tutoring systems. *Technology, Instruction, Cognition, and Learning (TICL)*, *5*, 169-271.
38. Scandura, J.M. (2014). Adaptive Learning: How It is Learned or What Is Learned?. *Technology, Instruction, Cognition, and Learning (TICL)*, 9, 237–239.
39. Veeramachaneni, K., Dernoncourt, F., Taylor, C., Pardos, Z., & O'Reilly, U. M. (2013, June). Moocdb: Developing data standards for mooc data science. In*AIED 2013 Workshops Proceedings Volume* (p. 17).
40. Wang, Y. 2014. MOOC learner motivation and learning pattern discovery. In the Proceedings of the 7th International Conference on Educational Data Mining (pp. 452-454).
41. Wang, Y., & Baker, R. 2015. Content or platform: Why do students complete MOOCs?. Journal of Online Learning and Teaching, 11(1), 17.
42. Wang, Y., Paquette, L., Baker, R. (2014) A longitudinal study on learner career advancement in MOOCs. *Journal of Learning Analytics, 1 (3),* 203–206.
43. Whitehill, J., Williams, J. J., Lopez, G., Coleman, C. A., & Reich, J. (2015). Beyond prediction: First steps toward automatic intervention in MOOC student stopout. *Available at SSRN 2611750*.
44. Yang, D., Sinha, T., Adamson, D., & Rose, C. P. 2013. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In Proceedings of the 2013 NIPS Data-driven education Workshop (Vol. 11, p. 14).
45. Yang, D., Wen, M., Howley, I., Kraut, R., & Rose, C. 2015. Exploring the effect of confusion in discussion forums of massive open online courses. In Proceedings of the Second (2015) ACM Conference on Learning@ Scale (pp. 121-130). ACM.
46. Yuan, L. and Powell, S. (2013) *MOOCs and Open Education: Implications for Higher Education*. Glasgow: JISC CETIS.