2021

Replication In Massive Open Online Course Research Using The Mooc Replication Framework

Juan Miguel Limjap Andres-Bray
University of Pennsylvania

Replication In Massive Open Online Course Research Using The Mooc Replication Framework

Abstract

The purpose of this dissertation was to develop and use a platform that facilitates Massive Open Online Course (MOOC) replication research. Replication and the verification of previously published findings is an essential step in the scientific process. Unfortunately, a replication crisis has long plagued scientific research, affecting even the field of education. As a result, the validity of more and more published findings is coming into question. Research on MOOCs have not been exempt from this. Due to a number of limiting technical barriers, MOOC literature suffers from such issues as contradictory findings between published works and the unconscious skewing of results caused by overfitting to single datasets. The MOOC Replication Framework (MORF) was developed to allow researchers to bypass these technical barriers. Researchers are able to design their own MOOC analyses and have MORF conduct it for them across its massive store of MOOC data. The first study in this dissertation, which describes the work that went into building the platform that would eventually turn into MORF, conducted a feasibility study that aimed to investigate whether the platform was able to perform the tasks it was built for. This was done through the replication of previously published findings within a single dataset. The second study describes the initial architecture of MORF and sought to demonstrate the platform's scaled feasibility to conduct large-scale replication research. This was done through the execution of a large-scale replication study against data from an entire University's roster of MOOCs. Finally, the third study highlighted how MORF's architecture allows for the execution of more than just replication studies. This was done through the execution of a novel research study that sought to analyze the generalizability of predictive models of completion between the countries present in MORF's expansive dataset—an important issue to address given the massive enrollment numbers of MOOCs from all around the world.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Education

First Advisor

Ryan S. Baker

Keywords

Completion, Cross-cultural, Massive Open Online Courses, MOOC Replication Framework, Predictive Modeling, Replication

Subject Categories

Artificial Intelligence and Robotics | Education | Instructional Media Design

REPLICATION IN MASSIVE OPEN ONLINE COURSE RESEARCH USING THE MOOC REPLICATION FRAMEWORK

Juan Miguel L. Andres-Bray

A DISSERTATION

in

Education

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation:

Ryan S. Baker

Ryan S. Baker, Associate Professor of Education

Graduate Group Chairperson:

Matthew Hartley

Matthew Hartley, Professor of Education and Associate Dean for Academic Affairs

Committee Members:

Susan A. Yoon, Professor of Education

Stephen J. Hutt, Postdoctoral Researcher in Education

George Siemens, Professor of Psychology, University of Texas at Arlington

REPLICATION IN MASSIVE OPEN ONLINE COURSE RESEARCH USING THE

MOOC REPLICATION FRAMEWORK

This dissertation is dedicated to the memory of my dad, Antonio Andres, Jr., who passed away in 2017. He was one of my biggest supporters and would have been extremely proud to see me grow into the man I am today.

I love you, Papa; we miss you every day!

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support, insight, and contributions from a small neighborhood of people to whom I am forever indebted.

First and foremost, to my advisor, Dr. Ryan Baker, without whom none of this would have been possible, your unyielding support and guidance from Day 1 have absolutely made this journey one I will treasure forever. If my career turns out to be as bright and hopeful as it appears, it will have only been because of you and your mentorship. Thank you so much for everything—it has been an honor and a privilege to work with you, and I am ecstatic we get to keep working together!

To my husband, Tyler, my North Star, my best friend, you are always the brightest, kindest, and most loving in everything you do, and I am so lucky I get to have the front seat to all of it. You have saved me from all forms of mental anguish, exhaustion, and imposter syndrome throughout this process, but I continue to stand as tall as I possibly can because of you. You are still and always will be my rock and my biggest cheerleader, and for as long as humanly possible, I will be yours. I love you so much!

To my committee, Drs. Stephen Hutt, Susan Yoon, and George Siemens, thank you for your guidance and feedback throughout the entire process; I would not have been able to put this dissertation together—both in content and efficiency—without your support. I am very honored to have had the chance to work with you.

To the MORF team, Josh Gardner, Dr. Chris Brooks, and Dr. Michael Mogessie, thank you so much for your contributions to MORF and its development—they have revolutionized the platform’s use and its contribution to replication and Open Science research.

To the members of the Baker EDM Lab and the Penn Center for Learning Analytics, Drs. Jaclyn Ocumpaugh, Sweet San Pedro, Mia Almeda, Shimin Kai, Elle Wang, Yang Jiang, and Shamya Karumbaiah, thank you for being a constant oasis of professional, academic, *and* personal support. I can't imagine a better, more caring group to have spent all these years with!

To my family, Alexis, Mama, and Anton, thank you for your constant support and prayers, and for always pushing me to be and do better. I know moving halfway around the world was not easy on our family, and I appreciate all you've done to support me and my dreams regardless. I am the man I am today because of you. I love you all so much!

To the Philly Fam, Dr. Amanda Barany and Stefan Slater, you are the kindest and most supportive best friends anyone could ever ask for. Thank you so much for always being there for us! We love you guys!

To my dogs, Luna and Leo, you guys delayed my progress by a couple of years, but I cannot/refuse to imagine what life would be like without you two, our demon babies. I love you both to bits!

ABSTRACT

REPLICATION IN MASSIVE OPEN ONLINE COURSE RESEARCH USING THE MOOC REPLICATION FRAMEWORK

Juan Miguel L. Andres-Bray

Ryan S. Baker

The purpose of this dissertation was to develop and use a platform that facilitates Massive Open Online Course (MOOC) replication research. Replication and the verification of previously published findings is an essential step in the scientific process. Unfortunately, a replication crisis has long plagued scientific research, affecting even the field of education. As a result, the validity of more and more published findings is coming into question. Research on MOOCs have not been exempt from this. Due to a number of limiting technical barriers, MOOC literature suffers from such issues as contradictory findings between published works and the unconscious skewing of results caused by overfitting to single datasets. The MOOC Replication Framework (MORF) was developed to allow researchers to bypass these technical barriers. Researchers are able design their own MOOC analyses and have MORF conduct it for them across its massive store of MOOC data. The first study in this dissertation, which describes the work that went into building the platform that would eventually turn into MORF, conducted a feasibility study that aimed to investigate whether the platform was able to perform the tasks it was built for. This was done through the replication of previously published findings within a single dataset. The second study describes the initial architecture of MORF and sought to demonstrate the platform's scaled feasibility to conduct large-scale replication research. This was done through the execution of a large-scale replication study against data from an entire University's roster of MOOCs. Finally, the third study highlighted how MORF's architecture allows for the execution of more than just replication studies. This was done through the execution of a novel research study that sought to analyze the generalizability of predictive models of completion between the countries present in

MORF’s expansive dataset—an important issue to address given the massive enrollment numbers of MOOCs from all around the world.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Replication, or the verification of an original study's findings in order to assess their robustness and generalizability (Brandt et al., 2014), is a crucial step in scientific inquiry, enabling researchers to better understand the reliability, validity, and merit of a study's findings. Despite its importance, however, replication studies 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 found that only 0.13% of the studies published involved replication (Makel & Plucker, 2014). There are several reasons for this: in addition to the fact that many educational research studies are difficult to reproduce due to issues of cost, researchers are also faced with access issues in terms of the original studies' design, method, and data. As such, a growing body of research across various fields of science have begun advocating for and implementing open science practices. Open Science (Fecher & Friesike, 2014) is a movement that seeks to increase transparency and access throughout each phase of the scientific process: study design, data collection, data analysis, and publication. One of the most urgent problems this movement seeks to address is the failure to replicate previous findings.

Online learning provides a new source of data that provides the opportunity to bridge the replication gap in the field of education research through the use of open science practices. Perhaps the largest opportunities for replication research in online learning come from Massive Open Online Courses (MOOCs), which afford millions of learners around the world free access to a wide variety of online course topics taught by professors from prestigious universities (Yuan & Powell, 2013). MOOCs' open and online nature afford various stakeholders opportunities to advance the field of education. MOOCs are able to reach massive audiences who would not normally have had access to quality educational materials. The University of Pennsylvania's offering of MOOCs, for example, has reached learners from over 150 countries in the world. Furthermore, MOOCs are known to draw in enrollment numbers in the tens of thousands per session (Jordan, 2014), leading to very rich and diverse datasets. Because of this, MOOCs have given instructors and researchers an unprecedented opportunity to study learner behavior at scale, affording them the opportunity to improve their course designs to better accommodate different cohorts of learners. Despite the size of the data generated by MOOCS, however, the majority of it are subject to strict regulations that seek to protect the privacy of learner records, i.e., the data is not freely accessible. This is a key reason why replication in MOOC research is not prevalent in the field.

The MOOC Replication Framework (MORF), the development of which is documented in this dissertation, seeks to allow researchers to conduct replication research without being hindered by technical barriers through its implementation of open science practices involved in data collection and analysis. It seeks to afford researchers the opportunity to conduct end-to-end replication studies by providing them 1) access to massive and diverse MOOC datasets and the computational power necessary to conduct large-scale analyses, and 2) the ability to archive and fully preserve their entire codebase and runtime environment, for easy review and reuse by external research teams.

In the following subsections, I discuss replication, its necessity in scientific research, and the replication crisis in more detail. I discuss the Open Education Science movement (van der Zee & Reich, 2018) and how it seeks to address this crisis by making scientific research and data more easily accessible. I then discuss how MOOCs and MOOC data can be leveraged in addressing the replication crisis in online learning research. I recount a brief history of MOOCs and discuss the kind of scholarship that formed within the field, the majority of which has been centered around the notion of learner *success* and its different operationalizations. I then discuss the limitations that MOOC research continues to face in terms of generalizability to new and different data and contexts. Next, I explain how the replication crisis manifested in MOOC research, and the technical barriers that have greatly impeded the field's replication efforts. Finally, I discuss the MOOC Replication Framework (the focus of this dissertation) as a solution to these barriers, its open science underpinnings, the goals behind its development, and what its key features and affordances are.

Replication

The repeated verification of an original study's findings is a crucial step in scientific inquiry, enabling researchers to better understand the reliability, validity, and merit of a study's findings. This commonly takes the form of either a reproduction or a replication. A study is deemed *reproducible* if a research team is able to obtain its original results through the execution of its original method and on its original dataset (Goodman, Fanelli, & Ioannidis, 2016). "Reproducibility is a minimum necessary condition for a finding to be believable and informative" (Bollen, Cacioppo, Kaplan, Krosnick, & Olds, 2015, p. 4). On the other hand, a study is *replicable* if a research team is able to employ the original methods as closely as possible on a new dataset in order to evaluate the robustness and generalizability of the original findings (Patil, Peng, &

Leek, 2016). Gardner, Yang, and colleagues (2018) posit that replication requires reproducibility, and because replications are more feasible in fields where data is not easily accessible by researchers outside the original team – a situation still the case in much of education, despite open data initiatives – the studies in this dissertation will focus on replication and replication studies.

Replication is the verification of an original study's findings in order to assess those findings' robustness and generalizability (Brandt et al., 2014), and is essential in scientific research. Replications are used to illustrate that a study's findings can be attained by different researchers in different contexts. "It is the proof that the experiment reflects knowledge that can be separated from the specific circumstances (such as time, place or persons) under which it was gained" (Schmidt, 2009, p. 3). It is the repeated verification of scientific results on new data, and is necessary in order to solidify scientific knowledge, guard against spurious results, discover the potential limitations of findings, and use experimental results to inform theory. Schmidt (2009) defines two notions of replication: direct and conceptual. Direct replications seek to validate an original study's findings through repetition of its original methodology. Conceptual replications, on the other hand, seek to test an original study's hypothesis or validate its findings through the use of different methods.

The Replication Crisis

Despite the importance of replication studies, they remain rare in education research. A 2013 survey of the 100 education journals with the highest 5-year impact factor ratings found that only 0.13% of studies published (221/164,589) involved replication (Makel & Plucker, 2014). Of these 221 studies, only 28.5% were direct replications, or replications where the original study's entire methodology was followed. 69.2% of these were conceptual replications, or replications where different methods were used to analyze the original study's hypotheses, while the remaining 2.3% had characteristics of both. Finally, almost half of the replication studies (48.2%) were conducted by the same research team that conducted the original study, and author overlap was found to relate significantly to the successful replication of a study's original findings (Makel & Plucker, 2014, p. 5). The survey report posits that this may be due to authors of such replication studies benefitting from the experience from having conducted the original study. This may also be due to their easier access to the tools and data sources used. Whatever the case, this difference raises concerns regarding the introduction of and the need to account for potential biases in such replication studies.

Recent evidence has shown that issues with replication are also widespread in the field of big data research, which covers a variety of academic disciplines including machine learning, artificial intelligence, and MOOC research. In a study conducted on 400 previously published works from leading artificial intelligence venues, none of the papers analyzed reported all details necessary to fully replicate their work. In fact, only about 20-30% of the components needed to replicate the original work were reported (Gundersen & Kjensmo, 2017). In a study conducted on 30 previously published works on text mining, for example, only one of the studies provided any technical means of replicating their experiment, i.e., source code or an executable program (Olorisade, Brereton, & Andras, 2017). Lack of access to data, computational capacity, and implementation methods were reported as barriers to replication in the works analyzed. In a survey of 613 published works on computer systems, the published code accompanying the papers failed to run in 20% of cases. In total, 75.1% of studies investigated in the study were not verifiable or replicable using the artifacts provided in the publication (Collberg et al., 2014).

Studies posit that this severe lack of replication studies in education is due to several reasons, such as submission bias (Mackel & Plucker, 2014; Spellman, 2012), where neither successful or unsuccessful replication studies are publishable due to the focus of most publication venues on novel research; funding bias, or the fact that many educational research studies are simply too costly to reproduce; and methodological differences, where replication studies have to contend with differences in study populations, idiosyncrasies between the conditions set in the original study, and current instructional conditions. The lack of replication leads to a surprisingly large proportion of spurious results being widely reported, as reported on by the Open Science Collaboration (OSC; 2015). In their report, the OSC, which is an open collaboration of scientists that seeks to improve scientific values and practices, replicated a hundred studies from three top psychology journals. Their study found that 64% of the replications conducted failed to obtain statistically significant results. These findings highlight the importance of replication research and the need to validate published findings. As such, a growing body of research has begun advocating for and implementing open science practices.

Open Education Science

Open Science is a term used to describe various philosophies and goals regarding the future of knowledge creation and dissemination (Fecher & Friesike, 2014). In the pursuit of improving the quality of published science, proponents of Open Science seek to increase transparency and access across various fields of research through the use of digital technologies and new practices. Open *Education* Science (van der Zee & Reich, 2018) is a movement that seeks to address problems of transparency and access specifically in education research, acknowledging the field's "diverse disciplinary traditions and [its] commitment to impact in policy and practice" (p. 2). This movement seeks to address such issues as publication bias, lack of access to original published research, and the failure to replicate. The practices proposed by Open Education Science fall into four categories, each related to a phase in the education research cycle: 1) open design, 2) open data, 3) open analysis, and 4) open publication.

Open Design relates to practices involved in the creation of a study's design and scope. These practices seek to make such processes more accessible to external readers, affording them an accurate account of the study's hypothesis, method, and analysis plan, and how these evolved over the course of its execution. Such practices can aid in the prevention of gaming the scientific system, where hypotheses are generated *after* the study's significant results are found. To achieve this, researchers from various fields observe a practice called *preregistration* (Gehlbach & Robinson, 2018), a practice in which a study's design is documented and shared publicly before it is conducted.

Open Data relates to practices involved in data collection, storage, and sharing. These practices aim to make data and other research materials freely accessible on public repositories for the purposes of replication, evaluation, and scrutiny by external research teams or the public. Sharing data on a by-request basis has been practiced for decades (e.g., Wollins, 1962), but has been proven ineffective (Wicherts, Borsboom, Kats, & Molenaar, 2006). With the advent of newer technologies, researchers have begun exploring secure online data storage, where data can be freely accessed by interested parties. As will be discussed in more detail in the following subsections, sharing data in its entirety is not always possible due to strict privacy restrictions, which seek to protect learner or subject confidentiality. As such, part of these practices involve researchers making decisions regarding what data can be shared and with whom it can be shared.

Open Analysis relates to practices involved in “the systematic reproduction of analytic methods conducted by other researchers” (van der Zee & Reich, 2018, p. 9). This is commonly practiced in various fields through code sharing (e.g., animal welfare: Wicherts, 2017; biomedicine: Page et al., 2018), where the source code used in the execution of a study’s analysis are uploaded and made publicly available on online repositories like GitHub. Recently, education researchers have also begun using containerization technology, which saves a user’s entire runtime environment into an executable virtual machine, complete with source code, dependencies, and operating system (e.g., Gardner et al., 2018b). This allows for more accurate and seamless execution of the original study’s methodology.

Finally, *open Publication* relates to practices involved in increasing public access to published work that would otherwise be behind a paywall. Several approaches to open publication include the uploading and sharing of whitepapers and publication preprints (e.g., Page et al., 2018)—manuscript drafts that have yet to be peer reviewed—on open platforms like arXiv (McKiernan, 2000); and post-publication peer review (Hunter, 2012), a process by which published works are indexed based on merit and impact.

Massive Open Online Courses

Online learning provides the opportunity to bridge the replication gap in the field of education research through the use of open science practices. Due to their scale and accessibility, Massive Open Online Courses, or MOOCs, present the largest opportunities for replication research in online learning. The term Massive Open Online Course (MOOC) was coined by Dave Cormier and Bryan Alexander in reference to an online course taught by Stephen Downes and George Siemens (Cormier, 2008; Fini, 2009). Downes and Siemens taught an online course on connectivism, a learning theory that highlights the importance of sharing and connecting with peers when learning in online education environments (Siemens, 2005), which attracted over 2,200 registrants (Fini, 2009). It was dubbed the first MOOC, specifically the first *cMOOC*, or *connectivist MOOC*. The cMOOC model places a premium on forming and fostering connections between learners as a form of knowledge building (Morrison, 2013). It builds on existing work on “networked practices… and distributed, many-to-many channels of communication” (Stewart, 2013, p. 230), as opposed to the more traditional teacher-centric classroom. In cMOOCs, instructors typically encourage their students to engage in networking activities, such as discussing course material on social media platforms, posting and responding on blogs, and contributing to community wikis.

The term *MOOC* was next used to describe a set of three courses that were offered by Stanford University in Fall 2011 on artificial intelligence, databases, and machine learning (Cooper & Sahami, 2013). These three experimental MOOCs were offered by the University in an effort to broaden the accessibility of their courses. These three MOOCs attracted over 310,000 registrants from more than 190 countries (Rodriguez, 2013; Jordan, 2014), and had over 43,000 completers (Ng & Widom, 2014), or learners who completed all graded course materials and earned a final grade greater than or equal to the course passing mark (typically 70-75%). Meanwhile, the Massachusetts Institute of Technology (MIT), which had been offering open online course content since 2001 to much smaller audiences (Schroeder, 2012), announced and launched MITx, an initiative that offered a wide range of courses and credentials for those who completed, in the same year (Rodriguez, 2013). These courses followed a model that differed from Downes and Siemens' cMOOC–these were instructor-directed and were modeled on the traditional classroom, as seen in the course materials and teaching methods they used (Morrison, 2013). These kinds of courses, which were later labeled *xMOOCs* or *eXtended MOOCs*, focused more on the "delivery of course content than on the participatory exploration characterized by cMOOCs," (Stewart, 2013, p. 230) and relied on "information transmission, computer-marked assignments, and peer assessment" (Rodriguez, 2013, p. 71). Despite their difference from cMOOCs, these online courses continued to be referred to as MOOCs due to their massive and open nature.

The xMOOC model dominated the MOOC space and MOOC research (Bozkurt, Akgün-Özbek, & Zawacki-Richter, 2017) due in large part to the commercial MOOC platforms that sprung from the Stanford and MIT MOOCs. In the years that followed the launching of these MOOCs, universities around the world began creating their own sets of MOOC offerings. By the end of 2013, over a hundred institutions had already partnered with leading MOOC providers like edX (Finkle & Masters, 2014) and Coursera (Haywood & Macleod, 2014). Since then, the number of institutions offering MOOCs has ballooned to more than 800 around the world, offering a total of more than 9,000 unique courses and pulling in a total of more than 52 million registered users across platforms

(Shah, 2019). The magnitude of this offering and the resulting data gathered by these courses have created new opportunities for various MOOC stakeholders to study learning at scale and improve their own courses to better accommodate their diverse cohorts of learners (Margaryan, Bianco, & Littlejohn, 2015).

The Attrition Problem

In their earlier years, MOOCs were envisioned to revolutionize and cause a “disruptive transformation” in higher education (Reich & Ruipérez-Valiente, 2019, p. 130) due to the opportunity afforded to institutions and instructors to reach a global audience (Lowenthal, Snelson, & Perkins, 2018). Universities began offering MOOCs as alternatives to on-campus for-credit courses (Jaschik, 2013; Sandeen, 2013) that remote learners would otherwise not have access to. In 2016, edX began offering what they termed *MicroMasters* programs, which were series of graduate-level MOOCs that were grouped together in sequence to earn graduate-level credentials or for the purposes of more targeted career advancement (De La Roca et al., 2018). As such, it became accepted among many instructors and researchers that the primary goal in a MOOC should reflect the goal of a traditional college classroom: to gain mastery of the content of the course, traditionally demonstrated by earning a passing mark and completing the course (Breslow et al., 2013).

Because both traditional classroom and MOOC instruction had these similar goals and metrics of success, both needed to provide learners enough support in order to achieve them. The differences between these two contexts, however, are that, unlike traditional classroom instruction, MOOCs do not require physical presence at a lecture, were mostly offered for free (in early years), and are open to participants with varying educational goals and backgrounds, seldom having any prerequisites (Rivard, 2013). MOOC instructors and course teams needed to put in significantly more effort in reaching out to and supporting their thousands of learners (Almatrafi, Johri, & Rangwala, 2018; Chandrasekaran, Ragupathi, Kan, & Tan, 2015), as opposed to teachers needing to reach and support the learners they had in the classroom. Because MOOCs were initially instructor-paced, i.e., content was released on a weekly basis, learners had to engage with the course throughout its duration in order to engage with all graded assessments. As such, learners coming into these MOOCs were expected to engage with the course, but also had complete control of the amount of time they were willing to invest in the endeavor. With little to no follow-up support from the majority of instructors or course teams, learners were bound to fall through the cracks.

Since their emergence, MOOCs have reported low completion rates of around less than 10% (Rivard, 2013; Rai & Chunrao, 2016), regardless of class size. For example, Duke University's MOOC on Bioelectricity in 2012 attracted over 12,000 registrants, but only 313 learners (2.6% of the cohort) completed (Onah, Sinclair, & Boyatt, 2014). Similarly, the University of Toronto's MOOC on Statistics had over 60,000 registrants, but only about 3,000 completers (5% of the cohort) (Gibbs, 2014). In one of the earliest comprehensive analyses on completion in MOOCs, Jordan (2014) looked at the completion numbers of 221 courses, gathered from multiple sources, such as news articles, academic reports, and social media. 78 institutions were present in the report, and the majority of the courses were hosted on either Coursera (54%) or Open2Study (19%). Of the 221 courses investigated, the author found that completion varied from 0.7% to 52.1%, with a median completion rate of 12.6%. The majority of learners were not getting the support they needed to stay engaged—likely due to the learners' varying goals in enrolling in these MOOCs (as will be discussed in later subsections)—and this often resulted in disengagement with the course after just one or two weeks (Jordan, 2014). As such, researchers turned their focus to better understanding what *successful* MOOC learners looked like and how they could best support the rest of their learners.

Successful MOOC Learners

In finding ways to address the attrition problem, which continues in MOOCs and MOOC research today (e.g., Chen, Sonnert, Sadler, & Malan, 2020; Lemay & Doleck, 2020), and bolster learner retention and completion rates, the majority of MOOC research turned to improving learner *success* in MOOCs, having initially operationalized success as earning a course completion certificate. Researchers investigated features related to individual courses, universities, platforms, and learners (Adamopoulos, 2013) as possible explanations of why learners were successful or not. Studies investigated different features relating to the MOOC's context and the MOOC experience and how these related to success. Studies looked at institution features, like the prestige of the offering university (Ospina-Delgado & Zorio-Grima, 2016; Milligan & Littlejohn, 2017); course features, like perceived effectiveness of content (Hone & El Said, 2016); platform features, such as the website on which the course is offered (Tsironis, Katsanos, & Xenos, 2016); and lecture video features, such as video length (Guo, Kim, & Rubin, 2014) and effectiveness of in-video quizzes (Brinton, Buccapatnam, Chiang, & Poor, 2016; Kovacs, 2016).

The majority of MOOC scholarship, however, has been more geared towards studying *learner*-related behaviors, how these related to course completion, and how the *good* behaviors could be supported, and the *bad* behaviors curbed. Studies investigated learner interactions with different sections of the MOOC, like discussion forums, peer assessments, and optional course surveys, and analyzed how these related to their likelihood to complete or drop-out from the course. Behavior in discussion forums, including posting behavior, was of interest to researchers. A study that investigated the amount of time spent interacting with different course resources in an edX MOOC on Electronics found that spending more time in the discussion forums (and less surprisingly, the graded assessments) were significant predictors of higher final scores (DeBoer et al., 2013). In another study conducted on an edX course on Big Data in Education, which investigated different forum posting-related behaviors, the researchers found that posting more frequently and writing longer posts than average were significantly predictive of whether or not a learner completed the course (Crossley et al., 2015). In yet another study, where discussion posts were automatically classified for confusion, a survival analysis was conducted to quantify the effect of confusion on learner dropout (Yang, Wen, Howley, Kraut, & Rosé, 2015). They found that the more confusion a learner expressed or was exposed to, the more likely they were to dropout.

A number of studies also investigated how interactions in peer assessments related to learner completion. A study, for example, that was conducted on two consecutive Coursera MOOCs on Human-Computer Interaction investigated grader reliability, or how closely, on average, a learner grades their peer’s assignments to its true score, by analyzing over 63,000 peer grades (Piech et al., 2013). The authors present peer grading as a solution to address the limitation within MOOCs to evaluate and provide feedback to more complex, open-ended problems. However, they also report that previous studies on the topic had found high numbers of unreliable peer graders who give grades over 10% lower than corresponding grades given by course staff. In their own study, they found what they dubbed *snap graders*, who spent significantly less time grading and were more likely to inflate the marks they gave. Being able to foster a network of reliable peer graders, or in the case of this study, creating an algorithm that is able to correct for peer grader biases and reliabilities, affords instructors a scalable solution to peer grading in MOOCs. The study found that the more reliable a learner was, the more likely they were going to continue engaging with the course.

Finally, a number of studies also looked at survey responses, though these kinds of studies were conducted and published less frequently because surveys were not built-in features in MOOCs and participation was almost always optional. A study that investigated Likert-scale survey responses on learner background and motivations in a Big Data in Education MOOC found that average self-reported self-efficacy, intention to follow instructor pace, and interest in course content as motivation for taking the course were significantly higher among completers than non-completers (Wang & Baker, 2015). Another study, which investigated motivations for enrolling in MOOCs, found that 22% of survey respondents who dropped out initially intended to complete the course, but were ultimately unable to due to academic and personal reasons (Gütl, Rizzardini, Chang, & Morales, 2014). A big majority of these respondents indicated that changes in their job, insufficient time, difficulty with the subject matter and unchallenging activities are some of the reasons for the drop-out. Yet another study, which was conducted on a MOOC on *Learning How to Learn*, surveyed its learners on their attitudes towards some of the course’s instructional design components (Jung, Kim, Yoon, Park, & Oakley, 2019). They found course content and structure to be significant predictors of the learners’ sense of progress towards course completion.

Some took these studies a step further by taking previously published findings and creating and implementing intervention ideas that sought to draw disengaged learners, or learners who had already dropped out, back to the course (Whitehill, Williams, Lopez, Coleman, & Reich, 2015). The researchers created detectors to find learners who were likely to drop out of the course, split them into control and experimental groups, and sent only the latter emails on a weekly basis with questions regarding their intent to continue with the course. Learners in the control condition did not receive emails and were instead used to compute the accuracy of their *stopout* classifier. These emails resulted in a significant difference in *comeback rate*, or the rate at which the learners came back to the course after getting and responding to the emails, between the two conditions.

Beyond Course Completion

MOOCs were designed as a platform wherein knowledge could be created and applied within the span of six to 12-week courses, with the hope that learning would transfer, and could thus be applied beyond it. When research into the improvement of course completion did not help in improving completion rates as hoped, instructors and researchers began investigating other forms of success, both internal and external to a course. Some papers, while also centering their investigations on the learner, instead looked at learner attributes, such as demographics (Dillahunt, Chen, & Teasly, 2014; Zhang et al., 2016), seeking to offer insights into the profiles of learners that take these courses. Dillahunt and colleagues (2014) for example, studied the demographic background of learners who had reported in a survey that they were taking the course due to the inability to afford more formal education. They found that 28% of these learners had less than a 4-year degree, which was significantly different from the 15% of the rest of their survey respondents. Another study, by a team at Penn State University, investigated the demographic breakdown of their learners based on their mode of communication preference with their peers, i.e., synchronous (live chat) vs. asynchronous (blog or forum posts) (Zhang et al., 2016). They found that learners who were more proficient in English preferred asynchronous interactions, male learners significantly preferred synchronous communication than female learners, and that as educational attainment increased, preference for synchronous communication decreased. While research on learner *behaviors* sought to find ways to improve learner success (e.g., promote good behaviors and curb bad behaviors), these kinds of studies instead sought to offer recommendations for how future courses can be designed and improved on to better support the profile of the less successful learners.

MOOCs have more recently been used to augment traditional learning environments through blended course designs, where MOOCs were combined with other forms of instruction. A recent study by Orsini-Jones & Carrascosa (2019), for example, reported on how the FutureLearn MOOC *Becoming a Better Teacher* was combined with English Language Teaching programs as a blended offering to learners. Participating in this study reportedly helped learners feel part of a global community (through asynchronous interactions afforded by the MOOC) and see the value of online collaboration in enhancing their own teaching practice. Another recent study by Wu and colleagues (2019), which sought to support affective development in its learners, combined a nine-week entrepreneurial MOOC and blended curriculum design. Their blended approach involved learners watching video lectures on the MOOC followed by face-to-face group discussions facilitated by the instructor in local classrooms. They concluded that blended MOOC-classroom designs can be effective, but were a time-consuming process.

Beyond the MOOC

Further studies investigated learner activities outside the MOOC platform. Such studies postulated that deeper insights on learning could be drawn when considering data outside of the course. Some studies offered predefined communities within external platforms as additional resources for their learners. For example, researchers from the University of Austin investigated the use of a Facebook group and Twitter feed associated with the course as a means of augmenting what is learned within the MOOC (Liu, McKelroy, Kang, Harron, & Liu, 2016). These additional social media spaces were offered as an optional means for learners to gather and discuss outside the course. Their quantitative and qualitative analyses of the users' feedback and usage found that these social spaces provided a place for their learners to connect with their peers, share additional resources, and provide a space to share personal feelings or reflections in an informal and quick manner.

Other studies also investigated the learners' use of external platforms not offered in or provided by the course. Chen and colleagues (2016) investigated the activity of more than 320,000 learners on various Social Web platforms, such as StackExchange, GitHub, Twitter, and LinkedIn, as a way of supplementing data from the MOOC platform. They tracked the learners' interactions with these platforms during and after the course. They sought to identify sets of traits and user attributes that either drew learners towards specific MOOC topics or were highly relevant to the online learning experience. The findings in their paper are broken down by Social Web platform, each detailing the profile of learners found on these the respective platforms. On Twitter, for example, they found that most of their learners on the platform were in the 20–30-year-old age range and were mostly male (89%). When analyzing their learners on StackExchange, the authors studied the learners' question/answering behavior during and after a MOOC. They found that questions relating to Haskell, the programming language used in the MOOC analyzed, dropped significantly on StackExchange after the course ended, but the answering remained stable, noting that their learners had turned more and more into answerers over time. The ability to track learners over time, beyond their interactions within the course, enabled them to investigate the impact of MOOCs on a learner over a much longer term.

Similarly, a number of studies that have looked into the longitudinal impact of MOOCs on learner success recognize that post-MOOC success can be difficult to measure. These studies posit that the definition of success depends on the learner's own goals and motivations, as many MOOC learners do not consider course completion to be their primary goal (Belanger & Thorton, 2013). Career advancement has also been cited among the primary goals of MOOC learners (Trumbore, 2020; Wang & Baker, 2018). Wang and Baker (2018) conducted a longitudinal study that investigated post-course career advancement. They looked at whether participants in an educational data mining MOOC ended up either joining a scientific community or submitting a paper to publication venues relevant to the course's topic area, dubbed *career advancers*, and analyzed how these types of learners interacted with the course during its run. Trumbore (2020) conducted a large-scale analysis investigating learner motivation and career benefits across 50 Wharton MOOCs. The study looked at self-reported job-related benefits, like receiving a promotion, obtaining their first job, getting a raise, or starting their own business. Both studies found that career advancers earned higher final scores than non-advancers and were more likely to have completed the course. They also found that advancers interacted more frequently with various course resources, like lecture videos, graded assessments, and discussion forums, though they surprisingly posted less often (Wang & Baker, 2018). Finally, Trumbore (2020) found that learners without college degrees were more likely to experience career benefits than those with degrees.

Limitations of MOOC Research

Despite the abundance of research on success and what success looks like in MOOCs, research in the field continues to suffer from limitations in infrastructure and access to data. Most notably, the majority of MOOC research is limited to a small selection of courses, often ones taught by the researchers themselves. This is due in most part to the lack of access to other data, as well as challenges to researchers in working with datasets much larger than those they are used to. While MOOCs do provide a great venue for conducting replication research due to the massive amount of data they generate, the majority of this data are subject to strict regulations that seek to protect the privacy of learner records. This lack of access to other, more diverse datasets can lead to issues of generalizability and replicability. Chapter 3 describes some instances of inconsistency between published works in recent years, where small-scale studies report contradictory findings.

While there has been some interest in data sharing within MOOCs, data-related barriers still persist in the field due to strict privacy regulations. Universities have access to data from hundreds of their own MOOCs, for example, but are unable to make them publicly available, giving full access to only each session's respective course team. EdX introduced the Research Data eXchange (RDX), which sought to make a limited amount of data from multiple universities accessible to researchers at other universities[1]. However, they also restricted the kinds of data available due to concerns of privacy, including key data necessary to replicating many previously published research, like demographic information and discussion forum posts, which commonly contain identifiable information[2].

Other platforms have since been developed, which seek to afford researchers the opportunity to improve replication and thus, validity, in MOOC research. The moocDB database schema was proposed and developed as a means of standardizing the vast amounts of data generated by multiple MOOC platforms (Veeramachaneni, Dernoncourt, Taylor, Pardos, & O'Reilly, 2013). It was mentioned heavily in MOOC scholarship as a solution to data sharing standards (e.g., Baker & Inventado, 2014; Pournaras, 2017; Sun et al., 2019), but was rarely used except in studies involving its developers (Han, Veeramachaneni, & O'Reilly, 2013). Its last published use was in (Han, 2014). MoocRP is an analytics tool that was developed with a goal of supporting replicable research (Pardos & Kao, 2015). It aimed to facilitate the replication of analyses in new MOOCs. However, moocRP did not achieve widespread use and its source code and documentation have not been updated since 2016.

The MOOC Replication Framework, or MORF, was the first platform to offer researchers both the computational power to conduct fully-replicable research and

[1] https://edx.readthedocs.io/projects/devdata/en/latest/

[2] https://edx.readthedocs.io/projects/devdata/en/latest/rdx/rdx_data.html#obfuscated-columns-in-the-auth-user-table

execute-only access to data from hundreds of courses from two universities without compromising any restricted data (Gardner, Brooks, Andres, & Baker, 2018a). Despite this, however, it has not yet been used widely–the majority of the published studies involving MORF were conducted by the same team that developed it (e.g., Andres et al., 2018; Gardner et al., 2018a). Its development team finished upgrades to the platform and its documentation very recently and have since started conducting beta testing in preparation for a wide relaunch.

Another issue MOOCs—and education research in general—have to contend with is the fact that studies in these fields are conducted predominantly on research subjects from Western, educated, industrialized, rich, and democratic (WEIRD) societies—96% based on a 2008 survey of the top psychology journals (Arnett, 2008)—while only accounting for 12% of the world's population (Henrich, Heine, & Norenzayan, 2010). These numbers cast doubt on just how well published findings will generalize to learners from smaller, less represented countries.

Replication in MOOCs

While there has been considerable research on predicting student success in MOOCs, relatively little assessment has been published of whether the models produced generalize across courses, platforms, or student cohorts. The limited number of replication studies on MOOCs has shown that published findings are not guaranteed to replicate. For example, a study that evaluated the generalizability of original findings of a study conducted on dropout predication in MOOCs (Xing, Chen, Stein, & Marcinkowski, 2016) found that only a subset of the findings replicated significantly across a larger sample of over 200 sessions of MOOC data (Gardner, Brooks, & Baker, 2019). The authors specifically investigated two of the original study's core findings: the first regarding which model performed the best, and the second regarding what kinds of learner features performed better on these models (i.e., appended features, or the creation of separate feature sets per week of the course; vs. week-only features, or features from only the current week being investigated). Some findings even replicated in the opposite direction, such as the original study's claim that a stacked ensemble of two classifiers outperformed either of the base classifiers–the replication study found that they actually performed significantly worse in four of the six cases investigated. Further, their study also revealed significant results which were not reported in the original experiment.

Another study, which investigated the generalizability of a dropout prediction model, analyzed the effects of both modeling and experimental design on the replicability of previously published findings (Gardner, Yang, Baker, & Brooks, 2019). In their study, the authors first attempted a direct replication of the original study's (Fei & Yeung, 2015) method. They followed its original design as closely as they were able to without cooperation from the original authors. They found that the model that performed the best in the original study (i.e., Long Short-Term Memory (LSTM) neural network model) was among the worst performers in the replication study. They posit that overfitting may have been the cause of the better performance in the original study.

Finally, a study by researchers from the University of Edinburgh and Monash University conducted a direct replication that investigated the robustness and generalizability of one previously published state-of-the-art classification model using the original study's methods and data set (Farrow, Moore, Gašević, 2019). The original study (Kovanović, et al., 2016) had conducted data rebalancing to account for more and less represented classes in their data. The replication study sought to test the effects of *different* data rebalancing methods, but first calculated a baseline value to see if and how well they could attain the original paper's findings. The study eventually found that, even when following the original method as closely as possible and using the same data, they were unable to achieve similar results; their findings had come out lower than the original study's findings on every outcome metric. They posit that the original study's findings may have been a result of data contamination (i.e., the same data points existing in both training and test sets during the model-building process) leading to overfitting and higher outcome metric scores.

These initial attempts at replication research highlight how the existing body of research on MOOCs may be particularly unreliable, especially given that most MOOC studies have used small samples of data and focused on highly varying subsets of students from the available datasets (e.g., only students who joined in the first ten days of the course, have viewed at least one lecture video, completed the pre-course survey and the first end-of-unit exam, etc.) (Gardner & Brooks, 2018). They show how problematic accepting a study's findings can be without attempting to verify them, especially when these may lead to interventions that alter the way learners learn and interact with a learning system. They highlight the importance of replication in the generalizability of a study's findings.

Barriers to Replication in MOOC Research

Though replication has been rare in MOOC research, many publications in the field end by stating the need to replicate their findings across different, more diverse data. Due to a number of barriers that exist, researchers are rarely able to follow-up on this claimed intention. This section identifies key barriers that contribute to this lack of general replicability in educational big data research, and, in particular, the lack of replication within the field of MOOC research.

Experimental challenges with replication relate to the difficulty encountered when reproducing the exact experimental environment (technical or otherwise) used in the original study (Gundersen & Kjensmo, 2018). In MOOC research, this commonly entails the proper and sufficient sharing of tools and algorithms used in the development of machine-learned models. Over the years, researchers have asked for the open sharing of source code as a minimum solution to address these issues with replication (Stodden & Miguez, 2013). However, even when a study's code is made available for others to use and build upon, other technical issues may still prevent replication in computational research workflows (Donoho, 2017; Kitzes, Turek, & Deniz, 2017). Even when source code can run error-free and is publicly shared, issues that are not resolved by code-sharing include 1) code rot, in which source code becomes outdated or nonfunctioning as the syntaxes and libraries used by the code change over time (e.g., the revision of the implementation of an algorithm which changes the way it computes its results); and 2) *dependency hell*, in which configuring the software necessary to install or run source code prevents successful implementation (Boettiger, 2015). As such, researchers have advocated for the sharing of complete software environments, as opposed to simply sharing the source code used in the study, citing this as a necessary condition for reproducing computational results (Buckheit & Donoho, 1995). However, such open sharing of complete software environments remains rare in MOOC research (and in computer science research more broadly).

Methodological challenges to replication reflect challenges related to the methods of the study, such as its procedure for model tuning or statistical evaluation. Much of existing work on replication focuses on technical challenges, but methodological issues are just as crucial to address. These include the use of biased model evaluation procedures (Cawley & Talbot, 2010; Varma & Simon, 2006). A common manifestation of such issues within prediction modeling research is seen in massive unreported searches during the model tuning process, in which researchers systematically test all possible model parameters in order to achieve better apparent performance (Henderson et al., 2018).

Finally, *data challenges* to replication relate to the availability of data. As stated previously, the majority of educational data are subject to strict regulations that seek to protect the privacy of learner data. As a result, researchers and instructors are often barred from making their data publicly accessible. Some have attempted to address this barrier. The Pittsburgh Science of Learning Center DataShop (Koedinger et al., 2010) and the HarvardX MOOC datasets (Hardvard-MITx, 2014), for example, have attempted to address this problem in educational research by only releasing limited non-reidentifiable data, but many analyses require the original, unprocessed data for a full replication. As previously discussed, restricted data sharing is one of the main factors hindering replication analysis in MOOC research, as investigators are generally limited to only small samples of data, and models generated on them are often overfit to the data available.

The MOOC Replication Framework

The MOOC Replication Framework (MORF) was designed to address technical, methodological, and data-related barriers to replication research in MOOCs (Gardner, Brooks, Andres, & Baker, 2018a) through its implementation of Open Science practices. In their manifesto for reproducible science, Munafò, Nosek, and colleagues (2017) propose a set of measures that directly target threads to reproducible science. They posit that the adoption, evaluation, and improvement of these measures will contribute to more robust scientific research. Through its implementation of Open Science practices, MORF is able to either support or more directly address a number of these measures. Specifically, MORF is able to support their proposal for collaboration and team science, which involve such initiative as multi-site studies and distributed data collection. These initiatives seek to facilitate "high-powered designs and [provide] greater potential for testing generalizability across the settings and populations sampled" (Munafò et al., 2017, p. 2). MORF is able to indirectly support this through both its Open Data and Open Analysis practices.

As discussed previously, Open Analysis practices seek to increase access and transparency to a study's methodology. These practices help prevent gaming of the scientific process—for example, through p-hacking, where researchers choose only to publish significant results they find interesting or pleasing. They also allow external research teams to fully replicate original studies. MORF contributes to Open Analysis in a couple of ways. Its main feature is its Platform-as-a-Service (PaaS) infrastructure, which consists of a running instance of its back-end infrastructure coupled with computational resources. Its design allows researchers to design, conduct, and share end-to-end replication of experiments through its use and open sharing of Docker containers (Boettiger, 2015). The containers, which are sent to MORF for feature extraction, are executable virtual machines that contain both 1) the end-user's source code, and 2) the runtime environment necessary for the code to run. Containers submitted to MORF are automatically shared on the Docker Hub[3], Docker's public registry of container images. Through this process, other researchers can simply access these containers to conduct their own research, possibly on their own data or in their own research context.

Open Data practices seek to increase access to a study's data. Most importantly, these practices aid external research teams in gaining access to actionable data where data collection efforts would otherwise be too costly or onerous to conduct. These practices also aid in the execution of replication research. MORF contributes to Open Data by allowing end-users controlled, execute-only access to its massive data store. This means that while MORF's entire dataset is available for learners to run analyses on, they are unable to see the actual dataset. Instead, end-users are given access to a sample dataset and documentation, allowing them the ability to write scripts that will ultimately work with MORF's dataset. Doing this allows end-users access to a massive MOOC dataset while still ensuring its compliance with data privacy regulations that seek to protect learner confidentiality. The end-users' scripts are submitted (as part of the Docker containers), which MORF then uses to perform extraction, training, testing, and model evaluation on the cloud. Intermediate outputs between these steps are stored securely on private Amazon Web Services[4] buckets. End-users then received a controlled set of outputs sent to their email, reporting their model's performance metrics.

[3] https://hub.docker.com/
[4] https://aws.amazon.com/

MORF more directly aids in addressing another category of Manufò et al.'s (2017) manifesto, which seeks to address study replication by encouraging transparency and open science. The initiatives under this category specifically cite Open Data and Open Analysis practices as a means of producing transparent and accessible evidence and scientific claims.

An in-depth discussion of the platform's initial goals and architecture can be found in Chapter 3. A discussion on its current architecture, which allows for direct replications, and a description of the available data can be found in Chapter 4.

Purpose of the Study

This dissertation focuses on the use of the MOOC Replication Framework (MORF) as a solution to the current technical barriers that exist to the conducting of replication studies. Specifically, this dissertation looks at how well previously published findings on completion in MOOCs replicate to new and different contexts. After the development of the platform, a feasibility study (Study 1) was first conducted to test whether it was able to execute the kind of research it was intended for. A usability study (Study 2) was then conducted to demonstrate its capability of running large-scale replications against multiple datasets. The final study of this dissertation (Study 3) sought to demonstrate what else MORF can be used for: the execution of novel, generalizable MOOC research.

Study 1: Development of MORF

This study sought to demonstrate the feasibility of the MOOC Replication Framework, outlining the work that went into the development of its initial architecture and the method it used in conducting replication. The chapter discusses the implications of the lack of replication in online learning and describes why MOOCs are the optimal platform for beginning to address this gap. In its first iteration, MORF was only able to conduct conceptual replication research through its implementation of an expert system comprised of multiple simple if-then production rules. Researchers interested in analyzing the replicability of their own work could transform their findings into simple if-then formulations for MORF to ingest. In turn, MORF would return the significance of the replications, i.e., how significantly the relationships held-up in a new dataset. In order to create the initial list of findings to replicate, a literature review was conducted. Specifically, we were interested in published works sought to investigate the relationships between learner-related behaviors or attributes and course completion. Findings from published papers were transformed into production rules and tested against a new dataset different from the various datasets used in the original studies. Finally, a feasibility study was conducted, where the replicability of 21 previously published findings were analyzed on a MOOC on Big Data in Education.

Study 2: Conduct Large-Scale Replication Using MORF

This study sought to demonstrate MORF's scaled feasibility through the execution of a large-scale replication using the platform. The study outlines MORF's goals and architecture in more detail. The replicability of 15 previously published findings were analyzed on data from the University of Edinburgh's entire MOOC offering on Coursera until 2015—a total of 29 sessions of 17 MOOCs, which attracted a total of 514,656 registrants. A meta-analysis was then conducted in order to combine results per production rule in order to obtain a single statistical significance score across all MOOC sessions.

Study 3: Conduct New Research Using Replicated Findings

This study sought to conduct research using the replicated findings from Studies 1 and 2 as a means of demonstrating MORF's capability of conducting novel research. Replicated findings were used as features in the development of completion prediction models, the generalizability of which were then tested between countries present in the dataset. Here, we utilized MORF's new prediction modeling module, which allows for richer, more direct forms of replication research. A dataset involving 81 countries was obtained. Completion prediction models were developed per country. Their models were then tested on every other country in the dataset. Within-country (i.e., baseline) model performances and cross-country performances were then used to compute distance, a metric used to quantify the models' cross-country generalizability. Finally, correlation mining and regression analyses were conducted to investigate the relationship between these model distances and different country-level measures of culture, happiness, wealth, and size. These analyses sought to take a close look at how significantly completion models built using entire MOOC datasets apply to learners from different geographic and cultural backgrounds.

Overview of Chapters

This dissertation proposal is organized into four chapters. Chapter 1 establishes the status of MOOC scholarship and the challenges that contribute to their lack of replication. A summary of the purpose of the study, a brief review of the existing MOOC and replication literature, the significance and implications of the work, and the study designs of each of the articles included in this dissertation are provided.

Chapter 2 presents Study 1.

Chapter 3 presents Study 2.

Chapter 4 presents Study 3.

Chapter 5 provides a conclusion to this dissertation, summarizing key findings across the three studies. The discussion highlights how each study contributed to the dissertation's overall goals. Finally, the chapter outlines MORF's current production roadmap and proposes some replication and novel research that the platform's new features enable.

CHAPTER 2: 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

Andres, J.M.L., Baker, R.S., Siemens, G., Gašević, D., & Spann, C.A. (2017).
Replicating 21 Findings on Student Success in Online Learning.
Technology, Instruction, Cognition, & Learning Journal, 10(1), 313-333.

Abstract

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.

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 in the previous 5 years 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 found that only 0.13% of the studies were those of 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. Many educational studies from the 1980s could no longer be easily replicated today, even if the desire to do so were present.

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. We picked this problem 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, it 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. Thus, research concerning MOOC completion is an active area for researchers as well as practitioners and one in need of a replication study.

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.

Method

Initial Data Set and 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 (Siemens & Baker, 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.

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.

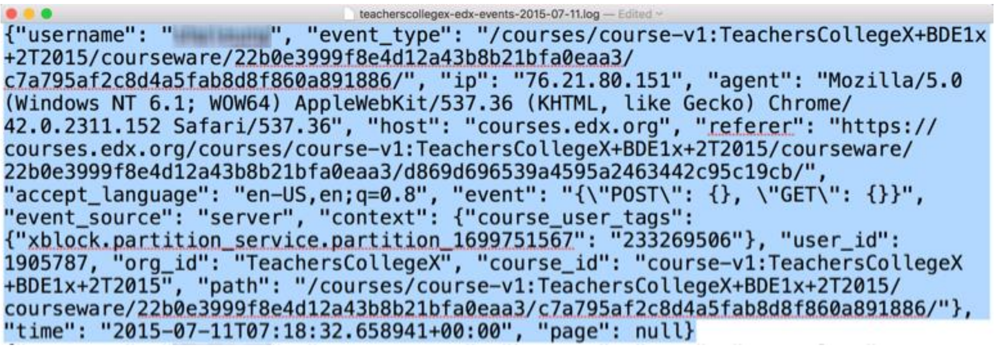
The current study also included five papers that investigated different student features and behaviors within the discussion forums. These papers found that writing longer posts (Crossley et al., 2015; Yang et al., 2013), writing more often (Crossley et al., 2015; Yang, Wen, Howley, Kraut, & Rosé, 2015), 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*.

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.

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.

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).

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 12 and 13, Rule 12 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 13 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).

Findings and Discussion

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.

Table 1. Production rule analysis results.

#	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	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
8	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
9	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
10	In forums: Starts thread	Likely to earn certificate	$\chi^2(1, N=516) = 0.004$, $p = 0.951$	Yang et al., 2013
11	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
12	In forums: Has respondents on thread	Likely to earn certificate	$\chi^2(1, N=516) = 2.067$, $p = 0.150$	Ramesh et al., 2013
13	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***
14	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
15	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
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

Note. 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 ***.

Fourteen production rules examined discussion forum behaviors and content features. Only the 518 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 assessment were included. The 21 production rules can be found in Table 1. 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).

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.

Nine production rules replicated significantly within the current data. Rule 6 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. It was drawn from a study that analyzed survey and log data from the previous iteration of BDEMOOC (Wang & Baker, 2015). Rules 8, 9, and 13 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. Posting and responding frequently on the forums implies an understanding of the topics being discussed or, at the very least, an interest to learn.

Rules 14 and 15 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. 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 than average, more trigrams than average, or more sophisticated words than average in their posts, they are more likely to complete. 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 11 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 11), and that students who make more posts are more likely to obtain a certificate (Rule 8, 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 8 and Rule 11) 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.

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 further stresses the importance of conducting replication studies in order to validate a study's results.

Conclusion and 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 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 BDE MOOC data, Diyi Yang of Carnegie Mellon University for her helpful comments, and Scott Crossley of Georgia State University for giving us access to the linguistic tools used in his papers.

CHAPTER 3: STUDYING MOOC COMPLETION AT SCALE USING THE MOOC REPLICATION FRAMEWORK

Juan Miguel L. Andres, University of Pennsylvania
Ryan S. Baker, University of Pennsylvania
Dragan Gašević, Monash University
George Siemens, University of Texas at Arlington
Scott Crossley, Georgia State University
Srećko Joksimović, University of South Australia

Andres, J. M. L., Baker, R. S., Gašević, D., Siemens, G., Crossley, S. A., & Joksimović, S. (2018). Studying MOOC completion at scale using the MOOC replication framework. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 71-78).

Abstract

Research on learner behaviors and course completion within Massive Open Online Courses (MOOCs) has been mostly confined to single courses, making the findings difficult to generalize across different data sets and to assess which contexts and types of courses these findings apply to. This paper reports on the development of the MOOC Replication Framework (MORF), a framework that facilitates the replication of previously published findings across multiple data sets and the seamless integration of new findings as new research is conducted or new hypotheses are generated. In the proof of concept presented here, we use MORF to attempt to replicate 15 previously published findings across 29 iterations of 17 MOOCs. The findings indicate that 12 of the 15 findings replicated significantly across the data sets, and that two findings replicated significantly in the opposite direction. MORF enables larger-scale analysis of MOOC research questions than previously feasible, and enables researchers around the world to conduct analyses on huge multi-MOOC data sets without having to negotiate access to data.

Introduction

Massive Open Online Courses (MOOCs) have created new opportunities to study how learning occurs across contexts, with millions of users registered, thousands of courses offered, and billions of student-platform interactions (Jordan, 2014). Both the popularity of MOOCs among students (Adamopoulos, 2013) and their benefits to those who complete them (Zhenghao et al., 2015) suggest that MOOCs present a new, easily scalable, and easily accessible opportunity for learning. A major criticism of MOOC platforms, however, is their frequently high attrition rates (Clow, 2013), with only 10% or fewer learners completing many popular MOOC courses (Jordan, 2015; Yang, Sinha, Adamson, & Rosé, 2013). As such, a majority of research on MOOCs in the past 3 years has been geared towards understanding and increasing student completion. Researchers have investigated features of individual courses, universities, platforms, and students (Adamopoulos, 2013) as possible explanations of why students complete or fail to complete.

A majority of MOOC research has been limited to single courses, often taught by the researchers themselves, which is due in most part to the lack of access to other data, as well as challenges to researchers in working with data sets much larger than those they are used to. While understandable, the practice of conducting analyses on small samples often leads to inconsistent findings and questions about the generalizability and replicability of what is learned. In the context of MOOCs, for example, one study investigated the possibility of predicting course completion based on forum posting behavior in a 3D graphics course (Andersson, Arvemo, & Gellerstedt, 2016). They found that starting threads more frequently than average was predictive of completion. Another study investigating this relationship in two courses on Algebra and

Microeconomics found the opposite to be true; participants that started threads more frequently were less likely to complete (Yang, Wen, Howley, Kraut, & Rosé, 2015). Research in single courses has the risk of producing contradictory findings which are difficult to resolve. Running analyses on single-course data sets limits the generalizability of findings, and leads to inconsistency between published reports (Łukasz, Sharma, Shirvani Boroujeni, & Dillenbourg, 2016).

In another example of this problem, one study investigating the relationship between students' motivations in taking the course and course completion across three open online learning environments found that students who were taking a course for credit were more likely to complete (Clow, 2013). An attempt to replicate this finding in a different MOOC found that this feature was not a statistically significant predictor of completion (Andres, Baker, Siemens, Gasević, & Spann, 2017).

The current limited scope of much of the current research within MOOCs has led to several contradictory findings of this nature, duplicating the "crisis of replication" seen in the social psychology community (Makel & Plucker, 2014). The ability to determine which findings generalize across MOOCs, which findings don't, and in what contexts less universal findings are relevant, will lead to trustworthy and ultimately more actionable knowledge about learning and engagement in MOOCs.

While there has been some initial interest in data sharing within MOOCs, prior efforts have not yet changed this state of affairs. Individual universities store data on dozens of MOOCs, but have mostly not yet made this data available to researchers in a fashion that enables large-scale analysis (although individual examples of multi-MOOC analyses exist (cf. Kim et al., 2014; Whitehill, Williams, Lopez, Coleman, & Reich, 2015). The edX RDX data exchange has made limited data from multiple universities accessible to researchers at other universities (edX, 2017), but has also restricted the data available due to concerns about privacy, restricting key data necessary to replicating many published analyses. The moocDB data format and moocRP analytics tools were developed with a goal of supporting research in this area (Pardos & Kao, 2015). Their tool allows for the implementation of several analytic models, with the goal of facilitating the re-use and replication of an analysis in a new MOOC. However, the use of moocRP has not yet scaled beyond analyses of single MOOCs, making it uncertain how useful it will be for the types of broad, cross-contextual research that are needed to get MOOC research past its own replication crisis.

In this paper, we present a solution that seeks to address this problem of replicability in the context of MOOCs. We do this by investigating the replicability of findings previously published in articles that leveraged learning analytics methods and data through the use of the MOOC Replication Framework.

MORF: Goals and Architecture

One of the common approaches to resolving the uncertainty caused by contradictory findings is to conduct meta-analyses (Schmidt & Hunter, 2014), where the results of several previous findings are integrated together to produce a more general answer to a research question. The meta-analysis research community has developed powerful statistical techniques for synthesizing many studies together despite incomplete information. By definition, however, a meta-analysis must wait on the completion of analyses by multiple research groups.

An alternate approach is to collect large and diverse data sets to then test published findings in. Such an approach has historically been infeasible in learning contexts, where data sources were, up until relatively recently, disparate, incompatible, and small. Even though large amounts of data have become available for individual intelligent tutoring systems over the last decade (Koedinger et al., 2010), the differences in the design of different tutoring systems and the semantics of data fields—even when the data field has the same name across different systems, or the systems share a common data format as in the Pittsburgh Science of Learning Center DataShop (Koedinger et al., 2010)—has made statistical analyses across multiple platforms relatively rare. However, analysis across large ranges of courses becomes more feasible for MOOCs, where a small number of providers generate huge amounts of data on courses with very different content, but relatively similar high-level design.

Table 2. Courses and iteration counts.

Course Title	Number of Iterations
Artificial Intelligence Planning	2
Animal Behavior and Welfare	1
Astrobiology	2
AstroTech: The Science and Technology Behind Astronomical Discovery	2
Clinical Psychology	1
Code Yourself! An Introduction to Programming	1
E-Learning and Digital Cultures	3
EDIVET: Do you have what it takes to be a veterinarian?	2
Equine Nutrition	2
General Elections 2015	1
Introduction to Philosophy	4
Mental Health: A Global Priority	1
Fundamentals of Music Theory	1
Nudge-It	1
Philosophy and the Sciences	2
Introduction to Sustainability	1
The Life and Work of Andy Warhol	2

To leverage this opportunity, we have developed MORF, the MOOC Replication Framework, a framework for investigating research questions in MOOCs within data from multiple MOOC data sets. Our goal is to determine which relationships (particularly, previously published findings) hold across different courses and iterations of those courses, and which findings are unique to specific kinds of courses and/or kinds of participants. In our first report on MORF (Andres et al., 2017), we discussed the MORF architecture and attempted to replicate 21 published findings in the context of a single MOOC. In this paper, we report the first large-scale use of MORF, attempting to replicate 15 published findings in 29 iterations of 17 MOOCs, listed in Table 2.

In its current version, MORF represents findings as production rules, a simple formalism previously used in work to develop human-understandable computational theory in psychology and education (Anderson, Matessa, & Lebiere, 1997; Laird, Newell, & Rosenbloom, 1987). This approach allows findings to be represented 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 used in MORF was built using Jess, an expert system programming language (Friedman-Hill, 2002). All findings were converted into if-else production rules following the format, "If a student who is <attribute> does <operator>, then <outcome>." Attributes are pieces of information about a student, such as whether a student reports a certain goal on a pre-course questionnaire. Operators are actions a student does within the MOOC. Outcomes can represent a number of indicators of student success or failure including watching a majority of videos (e.g., Kim et al., 2014; Sinha, Jermann, & Dillenbourg, 2014) or publishing a scientific paper after participating in the MOOC (e.g., Wang & Baker, 2015). In the current study, we focus on the most commonly-studied research question, whether or not the student in question completed the MOOC. Not all production rules need to have both attributes and operators. For example, production rules that look at time spent in specific course pages may have only operators (e.g., spending more time in the forums than the average student) and outcomes (i.e., whether or not the participant completed the MOOC) (e.g., DeBoer et al., 2013).

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

A chi-square test of independence can then be calculated comparing each confidence to each conviction. The chi-square test can 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 data set. Odds ratio effect sizes per production rule are also calculated. In this study, we tested MORF on 29 data sets obtained from the University of Edinburgh's large MOOC program. In integrating across MOOCs, we choose the conservative and straightforward method of using Stouffer's (1949) Z-score method to combine the results per finding across the multiple MOOC data sets, to obtain a single statistical significance result across all MOOCs. We also report mean and median odds ratios across data sets.

Scope of Analysis

In a first report on MORF's infrastructure, we attempted to replicate a set of 21 previously published findings in a single MOOC on Big Data in Education (Andres et al., 2017). Six findings analyzed in this first report required questionnaire data that was not available for the broader set of MOOCs investigated in the current study. As such, the current study analyzes the remaining 15 of these findings on MOOC completion across 29 iterations of 17 different MOOCs offered through Coursera by the University of Edinburgh. There was a total of 514,656 registrants and 86,535,662 user events across these 29 MOOC data sets.

Within the context of these MOOCs, we investigate previously published findings from five papers demonstrating that discussion forum behaviors were associated with successful course completion. This category of findings was studied for two reasons. First, it has importance to the design of effective MOOCs. Understanding the role that discussion forum participation plays in course completion is important to designing discussion forums that create a positive social environment that enhances learner success (Wen, Yang, & Rosé, 2014). Second, it represents a type of finding that has been difficult to investigate at scale with existing data sets, since there has been limited sharing of the type of discussion forum data necessary for this type of research, due to the difficulty of deidentifying this type of data. Prominent findings on MOOC completion involving time spent within the forums, as compared to other activities, were also considered.

Table 3. Production rules included in the study.

#	If	Then	Source
1	Participant spends more time in forums than average	Likely to complete	(DeBoer et al., 2013)
2	Participant spends more time on assignments than average	Likely to complete	(DeBoer et al., 2013)
3	Participant's average length of posts is longer than the course average	Likely to complete	(Yang et al., 2013; Crossley et al., 2015)
4	Participant posts on the forums more frequently than average	Likely to complete	(Yang et al., 2015; Crossley et al., 2015)
5	Participant responds more frequently to other participants' posts than average	Likely to complete	(Yang et al., 2013)
6	Participant starts a thread	Likely to complete	(Yang et al., 2013)
7	Participant starts threads more frequently than average	Not likely to complete	(Łukasz et al., 2016)
8	Participant has respondents on threads they started	Likely to complete	(Ramesh et al., 2013)
9	Participant has respondents on threads they started greater than average	Likely to complete	(Ramesh et al., 2013)
10	Participant uses more concrete words than average	Likely to complete	(Crossley et al., 2015)
11	Participant uses more bigrams than average	Likely to complete	(Crossley et al., 2015)
12	Participant uses more trigrams than average	Likely to complete	(Crossley et al., 2015)
13	Participants uses less meaningful words than average	Likely to complete	(Crossley et al., 2015)
14	Participant uses more sophisticated words than average	Likely to complete	(Crossley et al., 2015)
15	Participant uses a wider variety of words than average	Likely to complete	(Crossley et al., 2015)

Note. Previous findings are presented as production rules. The articles from which the findings were drawn from are also reported.

These five past papers found that writing longer posts (Yang et al., 2013; Crossley et al., 2015), writing posts more often (Łukasz et al., 2016; Crossley et al., 2015), starting a thread, receiving replies on one's thread, and replying to others' threads (Yang et al., 2013; Łukasz et al., 2016; Ramesh et al., 2013), and just generally spending more time in the forums and on quizzes (DeBoer et al., 2013) were significantly associated with course completion. The original papers on these findings involved one edX MOOC on Electronics (DeBoer et al., 2013), and Coursera MOOCs on Surviving Disruptive Technology (Ramesh et al., 2013), Algebra (Yang et al., 2013; Yang et al., 2015), Microeconomics (Yang et al., 2013; Yang et al., 2015), and Big Data in Education (Crossley et al., 2015). The full list of findings investigated is given in Table 3.

One area of particular interest for many MOOC researchers is learners' failure to complete MOOC courses, due the problem's importance and potential actionability. Completion is important even beyond the context of a single MOOC. Though not all MOOC learners have the goal of completion (Wilkowski, Deutsch, & Russell, 2014), completion is one of the best predictors of eventual participation in the community of practice associated with the MOOC (Wang & Baker, 2015). As such, understanding why learners fail to complete MOOCs may enable the design of interventions that increase the proportion of students who succeed in MOOCs. The studies included in this paper's set of analyses sought to understand which student behaviors were significantly related to course completion, as a step towards designing interventions.

In the first of these five articles, DeBoer and colleagues (2013) explored the impact of resource use on achievement within edX's first MOOC, Circuits and Electronics, offered in Spring 2012. The class reportedly drew students from nearly every country in the world. The study correlated course completion to the amount of time spent on different online course resources, and found that time spent on the forums and time spent on assignments were predictive of higher overall final scores (required for course completion with a certificate), even when controlling for prior ability and country of origin. These results show that time allocation is an important predictor of student success in MOOCs.

Two studies by Yang and her colleagues (2013; 2015) explored dropout rates, confusion, and forum posting behaviors within two Coursera MOOCs, one on Algebra and the other on Microeconomics. Their first study developed a survival model that measured the influence of student behavior and social positioning within the discussion forum on student dropout rates on a week-to-week basis. The second study attempted to quantify the effect of behaviors indicative of confusion on participation through the development of another survival model. They found that the more a participant engaged in behaviors they believed indicative of confusion (i.e., starting threads more frequently than the average student), the lower their probability of retention in the course. The findings of these two studies on the relationship of posting behavior (i.e., starting threads, writing frequent and lengthy posts, and responding to others' posts) to course completion are crucial to the design of MOOCs because they suggest that social factors are associated with a student's propensity to drop out during their progression through a MOOC.

Crossley and colleagues (2015) conducted a similar investigation on the relationship between discussion forum posting behaviors and MOOC completion in a MOOC on Big Data in Education. In their study, they also found that a range of linguistic features, computed through natural language processing, were associated with successful MOOC completion, including the use of concrete, meaningful, and sophisticated words, and the use of bigrams and trigrams. Concreteness is assessed based on how closely a word is connected to specific objects. "If one can describe a word by simply pointing to the object it signifies, such as the word apple, a word can be said to be concrete, while if a word can be explained only using other words, such as infinity or impossible, it can be considered more abstract" (Kyle & Crossley, 2015, p. 762). Meaningfulness is assessed based on how related a word is to other words. According to the definition in (Kyle & Crossley, 2015), words like "animal," for example, are likely to be more meaningful than field-specific terms like "equine". Lexical sophistication involves the “depth and breadth of lexical knowledge” (Kyle & Crossley, 2015). It is usually assessed using word frequency indices, which look at the frequency by which words from multiple large-scale corpora appear in a body of text (Kyle & Crossley, 2015). More concrete or more sophisticated words were found to be associated with a greater probability of course completion, while more meaningful words were found to be associated with a lower probability of course completion. The findings of their study have important implications for how individual differences among students that go beyond observed behaviors (e.g., language skills and usage choices) can predict success.

As mentioned, the current study attempts to replicate 15 previously published findings relating to participant behaviors and MOOC completion. These findings are presented in Table 3 as if-then production rules; the previous articles the findings were drawn from are also included. The findings are divided into three categories: findings involving data drawn from clickstream logs concerning time spent on specific activities within the MOOC (Rules 1-2), findings involving data drawn from the discussion forum that look at the participants’ posting behavior (Rules 3-9), and findings involving data from the forum posts that look at linguistic features of the participants’ contributions (Rules 10-15). The Tool for the Automated Analysis of Lexical Sophistication 1.4, or TAALES (Kyle & Crossley, 2015), and the Tool for the Automatic Analysis of Cohesion 1.0, or TAACO (Crossley, Kyle, & McNamara, 2016), were used to generate the linguistic variables used in the analyses.

In TAALES, sophistication is derived from word occurrence across multiple large-scale corpora and are computed using five frequency indices: the Thorndike-Lorge (1944) index based on Lorge’s 4.5 million-word corpus on magazine articles, the Brown

(1984) index based on the 1 million-word London-Lund Corpus of English Conservation (Svartvik & Quirk, 1980), the Kucera-Francis (1967) index based on the Brown corpus, which consists of about 1 million words published in the US, the British National Corpus (BNC; 2007) index based on about 100 million word of written and spoken English in Great Britain, and the SUBTLEXus index based on a corpus of subtitles from about 8000 films and television series in the US (Brysbaert & New, 2009). TAALES returns a sophistication score per corpus. The more words from these five corpuses are used, the higher the respective sophistication score is. For more information on these corpora, see (Kyle & Crossley, 2015). Bigram and trigram frequency are two other metrics of lexical sophistication (Kyle & Crossley, 2015), i.e., the more bigrams and trigrams used, the more sophisticated a body of text is.

One production rule studied in this paper is a re-parameterized version of an original finding that was carried over into the current study from the first use of MORF in a single MOOC (Andres et al., 2017). Rule 8 was the original finding, i.e., participants having respondents on their threads in the discussion forum. Within (Andres et al., 2017), we created a variant of this rule, Rule 9, participants having more respondents on their threads than average, due to the relatively low numbers of threads with zero respondents in some MOOCs.

Using MORF

The production rule analysis of MORF makes use of two different kinds of data: 1) clickstream events used to analyze the rules relating to the amount of time spent in the forums and on the assignments, and 2) relational database forum data used to analyze the rules relating to forum behavior and linguistic features. MORF utilizes

Amazon Web Services (AWS) for data storage of the clickstream events, which are stored in Amazon S3 buckets, and database access for the forum-related data via Amazon’s Relational Database Service (RDS).

When MORF is run, it connects securely and remotely to AWS to access all necessary data. The user simply needs to state which courses and course iterations they intend to run the production rule analysis on, and once the analysis is complete, the user is presented with the results of the analysis. This consists of the list of MOOCs currently in MORF’s data storage, whether or not each production rule replicated significantly within each course iteration, and the significance level and effect size for each analysis, as well as the overall analysis.

Utilizing such an architecture protects data ownership by enabling users to run analyses without getting direct access to any of the raw data, a crucial feature in conducting research with data privacy limitations. Users are also able to either contribute their own data sets to MORF, or conduct their own analyses against MORF’s data set, which is currently comprised of 131 iterations of 61 MOOCs.

Results

The results of the 15 analyses across MOOCs can be found in Table 4, where each row represents the result of testing each previously published finding across the full set of MOOCs. The table reports each finding, again presented as an if-then production rule, the respective Z-scores and p-values for the analysis across MOOCs, as well as the number of MOOCs in which the finding significantly replicated, the number of MOOCs that had the counterfactual replicate, and the number of MOOCs where the finding failed to replicate in in either direction. Counterfactuals that are statistically significant overall, across MOOCs, are marked by shaded bands. Findings that failed to replicate in either direction are italicized. Table 4 also reports the mean and median odds ratio effect sizes of each production rule across the 29 data sets.

Table 4. Meta-Analysis results per production rule.

Production Rules	Z	p	+	-	null	Odds Ratio Mean	Odds Ratio Median
More time in forums	26.93	< 0.001	29	0	0	27.235	12.060
More time on assignments	26.93	< 0.001	29	0	0	251.979	121.349
Longer posts than average	11.76	< 0.001	15	1	13	1.362	1.238
Posts more frequently than average	26.04	< 0.001	27	0	2	4.667	3.406
Responds more frequently than average	23.84	< 0.001	25	0	4	2.959	2.569
Starts a thread	12.34	< 0.001	15	0	14	1.874	1.676
Starts threads more frequently than average[a]*	26.39	< 0.001	0	27	2	4.601	3.571
Has respondents	22.29	< 0.001	26	0	3	2.321	1.997
Has respondents greater than average	22.72	< 0.001	24	0	5	2.544	2.250
Uses more concrete words[b]	*1.51*	*0.131*	*3*	*5*	*21*	*1.036*	*1.076*
Uses more bigrams	12.68	< 0.001	15	1	13	1.376	1.292
Uses more trigrams	12.84	< 0.001	16	1	12	1.390	1.281
Uses less meaningful words	10.18	< 0.001	16	0	13	0.799	0.782
Uses more sophisticated words	17.54	< 0.001	20	0	9	1.623	1.472
Uses wider variety of words[a]	-4.11	< 0.001	2	13	14	0.987	0.875

[a] Shaded bands indicate that our replication found the reverse of the published finding.

[b] Italics represent null results.

* All outcomes are “likely to complete,” except for the rule suffixed by an asterisk, where the outcome is “not likely to complete.”

As shown in Table 4, two of the 15 previous findings had their counterfactuals come out statistically significant, i.e., they had the opposite result from the result previously reported. Whereas Yang and colleagues (2015) found that students who start threads on the forums more frequently than the average student are less likely to complete, we found that in 27 cases out of 29 (with 0 positive replications and 2 null effects) that students who start threads less frequently are less likely to complete. Also, whereas Crossley and colleagues (2015) found that students who used a wider variety of words in their forum posts than the average student were more likely to complete, we found in 13 cases out of 29 (with 2 positive replications, and 14 null effects) that students who used a narrower variety of words were more likely to complete. Finally, one finding, which originally stated that students who used more concrete words in their forum posts than the average student were more likely to complete, failed to replicate overall in either direction (with 3 positive replications, and 5 negative replications). The remaining 12 of the 15 previous findings replicated significantly across the 29 data sets.

Implications

Twelve of the fifteen production rules investigated significantly replicated across the data sets. The previously published findings related to time spent in the forums and on assignments – stating that more time spent on these activities is associated with completion – replicated significantly across all 29 data sets. These findings 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 within the MOOC and are also more likely to complete. Spending more time with the material may also increase the chance 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 that encourage participation.

Beyond this, most rules on posting behaviors replicated significantly across the 29 data sets as well. These rules found that writing longer posts, writing posts more frequently, responding more frequently to other students' posts, and having others respond more frequently to one's own posts are all significant predictors of completion. Interactions among and between students and course staff, and certainly, the behavior of posting and responding frequently on the forums implies, at the very least, an interest to learn. This greater effort spent in participation in many cases is probably also associated with learning from one's peers, an important aspect of MOOCs.

One rule in this area, however, replicated significantly in the opposite direction. The finding originally stated that students who start threads more frequently are less likely to complete (Andres et al., 2017). Its counterfactual, however, which states that students who start threads less frequently than the average student are less likely to complete, replicated significantly across 27 of the 29 MOOCs. Yang and colleagues interpreted starting a thread as indicating confusion, and indeed, this may motivate some students to start threads. It is likely, however, that students start threads for many reasons beyond confusion, including to share ideas (Sharif & Magrill, 2015), make personal contact with other students (Sharif & Magrill, 2015; Milligan, Littlejohn, & Margaryan, 2013), and even to insult their instructor (Comer, Baker, & Wang, 2015). It may be valuable in future work to more thoroughly study the content of discussion threads in order to see if different posts have different associations to student outcomes.

In terms of the linguistic features of the participants' forum posts, the analysis found that students more likely to complete the MOOCs produced more sophisticated language and used more bigrams and trigrams, but used less meaningful words, replicating the findings of Crossley and his colleagues (2015). However, Crossley et al.'s previous findings on concreteness failed to replicate (but did not replicate in reverse either).

One of the findings that did replicate was the negative relationship between using meaningful words and course completion. Within TAALES, meaningful words are words with greater numbers of associations to other words, regardless of domain (DeBoer et al., 2013). In other words, the finding seen here—replicating (Crossley et al., 2015)—may be because words interpreted as linguistically meaningful by TAALES may be less relevant to course content than other words. Using fewer meaningful words could thus mean that participants were using field-specific terms in their discussion posts. Conversing using field-specific terms could imply better understanding of the content being taught in the course. By contrast, lexical sophistication involves the "depth and breadth of lexical knowledge" (Kyle & Crossley, 2015). Word sophistication, bigram use, and trigram use are all measures of lexical sophistication within TAALES. The findings positively linking lexical sophistication to course completion, thus, imply that more sophisticated posts are associated with remaining engaged in the course. More sophisticated language may also be associated with positive understanding of the course content.

One production rule turned out to be significant in the reverse direction from what was reported in its original article. The finding was part of a set of linguistic features that were correlated with course completion (Crossley et al., 2015). The rule originally states that participants who post on the forums using a wider variety of words than the average student were more likely to complete. This analysis, however, found that using a narrower variety of words was significantly related to course completion. One possibility is that students who use a considerable variety of words are not focusing on words of specific importance for their current course, but are instead rambling on a range of other (often unrelated) topics (cf. Comer et al., 2015; Wang, Yang, Wen, Koedinger, & Rosé, 2015).

Overall, these findings suggest that there is considerable commonality in which behaviors are associated with success in MOOCs, across MOOCs on a heterogeneous range of topics, creating the possibility that interventions that encourage specific behaviors from the set studied here may have positive incomes on student success, even in entirely new courses.

Conclusion and Future Work

In this paper, we investigate the degree to which previously published findings on MOOC course completion replicate across multiple new and different data sets. This was achieved through the development of the MOOC Replication Framework, or MORF, a framework that was used to attempt the replication of 15 previously published findings on MOOC completion on 29 MOOC data sets, drawn from 17 distinct courses on a range of topics. These 15 findings, represented as productions rules, were drawn from 5 studies that sought to understand the high attrition rate in MOOCs. Of these 15 findings, 12 successfully replicated across the 29 data sets, while 2 were statistically significant in the opposite direction. Through the development of MORF and the resulting analyses conducted, this study presents a larger-scale analysis of MOOC research questions than previously feasible.

Our next steps include extending our work published here in several ways. First, we plan to expand the current set of variables being modeled in MORF, both in terms of predictor (independent) variables and outcome (dependent) variables. This will enable us to replicate a broader range of published findings. 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 move our framework beyond simply capturing findings that can be expressed and production rules, and also analyze findings that can only be expressed as more complex predictive models, in partnership with researchers at the University of Michigan. While we view production rules as a highly interpretable and reasonably flexible framework, more complex prediction models are already in use to determine which students are at risk of failing to complete a course (Kim et al., 2014; Whitehill et al., 2015; Wen et al., 2014). Being able to test these more complex models for replication as well will broaden the applicability of the MORF framework.

Third, we plan to expand to an even greater range of data. Initially, we plan to apply the production rules to data from other MOOC courses. This should be a straightforward process as MORF is able to ingest raw edX and Coursera data seamlessly. At the time of this writing, we are nearing completion of the ingestion of edX and Coursera data from two other universities. Eventually, we hope to add data from other platforms as well.

Fourth, we intend to add to MORF a characterization of the features of the MOOCs themselves, towards studying whether some findings fail to replicate in specific MOOCs due to the differences in design, domain, or audience between MOOCs. Although 13 findings replicated overall, not all findings replicated in all MOOCs. Understanding how the features of the MOOC itself can explain differences in which results replicate may help us to explain some of the contradictory findings previously reported in single-MOOC research. With the large pool of courses MORF currently has access to, we intend to go beyond simple replication to study how factors like course design, target and actual population, domain, and instructor pedagogy influence the applicability of these findings. In turn, this will help us to understand which findings apply in which contexts, towards understanding how the different design of different MOOCs drive differences in the factors associated with student success.

Fifth, and perhaps most importantly, we are currently working with colleagues at the University of Michigan to create an infrastructure which will enable us to share access to MORF – while not sharing the data sets themselves – to a broader audience. This will enable a broader range of researchers to access and utilize large-scale MOOC data to conduct generalizable research on learning in this context. By broadening the base of access to large-scale learning data, we can incorporate a wider variety of ideas and a greater amount of energy and researcher time, with the hope of eventual speeding progress in this emerging scientific area.

Acknowledgements. Thanks to the Bill and Melinda Gates Foundation, the University of Pennsylvania, the University of Edinburgh, and the University of Texas at Arlington, for their support for this research.

CHAPTER 4: EXPLORING CROSS-COUNTRY PREDICTION MODEL GENERALIZABILITY IN MOOCS

Juan Miguel L. Andres-Bray, University of Pennsylvania
Stephen J. Hutt, University of Pennsylvania
Ryan S. Baker, University of Pennsylvania

Andres-Bray, J. M. L., Hutt, S. J., & Baker, R. S. (Under review). Exploring Cross-Country Prediction Model Generalizability in MOOCs.
To *the 11th International Conference on Learning Analytics and Knowledge*.

Abstract

Massive Open Online Courses (MOOCs) have increased the accessibility of quality educational content to a wider audience across a global network. They provide access for students to material that would be difficult to obtain locally, and an abundance of data for educational researchers. Despite the international reach of MOOCs, however, the majority of MOOC research does not account for demographic differences relating to the learners' country of origin or cultural background, which have been shown to have implications on the robustness of predictive models and interventions. This paper presents an exploration into the role of nation-level measures of culture, happiness, wealth, and size on the generalizability of completion prediction models across countries. The findings indicate that various dimensions of culture are predictive of cross-country model generalizability. Specifically, learners from indulgent, collectivist, uncertainty-accepting, or short-term oriented countries produce more generalizable predictive models of learner completion.

Introduction

Massive Open Online Courses (MOOCs) are a recent innovation within e-learning and distance education and have increased the accessibility of quality educational content to a wider audience across a global network (Adamopoulos, 2013). They have opened multiple opportunities for learning across different contexts, and for millions of users across the thousands of available courses (Shah, 2019). However, MOOCs have suffered from steep attrition rates since their inception (Jordan, 2014). In seeking to address this issue, researchers have investigated various learner-related features (Adamopoulos, 2013) and how these relate to the learners' likelihood to complete. To this day, MOOC scholarship continues its attempt to find ways to support learner retention (e.g., Moore & Wang, 2021; Pereira, 2021), expressing a continued need for accurate prediction of learner outcomes and the resulting development of automated interventions.

Despite MOOCs having a worldwide audience, however, the majority of MOOC research has not been able to account for the large differences in their learners' country of origin or cultural background. Studies have found that learners from Western, educated, industrialized, rich, and democratic (WEIRD) societies account for the majority of research subjects in psychology—96% based on a 2008 survey of the top psychology journals (Arnett, 2008)—while only accounting for 12% of the world's population (Henrich et al., 2010). However, because this phenomenon is often unavoidable—in MOOCs, for example, where the majority of courses are hosted in and offered by institutions within these WEIRD societies—studies thus need to turn their attention towards investigating how well their published findings generalize across country borders. A recent study by Li and colleagues (2021), for example, sought to investigate the generalizability of models trained on data from the United States (a WEIRD country) on data gathered from learners from other countries. They found that US-trained models could predict achievement in data from other developed countries with high accuracy but dropped linearly with the other country's degree of economic development. Investigating the cross-country generalizability of published findings can be a step towards better understanding and supporting the needs of learners from less represented countries.

This study explores the role of demographic differences in country-level measures of culture, size, wealth, and national happiness within a dataset of almost 2 million learners enrolled in Penn's 2012-2015 selection of Coursera MOOCs. Using log data, this study examined learners based on the country they interacted with the MOOC from. Learners from the United States are the largest group of learners in the dataset (33%). To better contextualize this, the next most represented country, India, accounts for just 8% of the dataset. I examine the impact of country-level demographics on the generalizability of completion prediction models across diverse learner populations from 81 different countries, as well as to identify which features and differences relate to the degree of generalizability seen. To our knowledge, this paper presents the broadest exploration yet into the role of country-level demographics on model generalizability and application across countries.

Related Literature

Cross-Country Generalizability in e-Learning Research

Investigations into the cross-country generalizability of published findings have been rare across e-learning fields. Some studies have reported promising results, like a study by San Pedro and colleagues (2011), which reported on a successful generalization of carelessness models between learners in the US and in the Philippines. However, other studies suggest that transferring models across learner populations can lead to poor model performance, relative to the training country's own baseline model performance. A study that investigated help-seeking behaviors in intelligent tutoring systems found that help-seeking models transferred to some degree between learners from the US and the Philippines, but not to Costa Rica (Ogan et al., 2015). They explained these findings to be a result of differing classroom practices between country sites, e.g., positing that the greater collaboration observed in Costa Rica resulted in help-seeking behaviors occurring *outside* the technology studied. They concluded by cautioning against the assumption that the models underlying educational systems will generalize across cultures and contexts.

Need for Cross-Country Generalizability in MOOC Research

MOOC scholarship has yet to investigate the issue of cross-country generalizability, likely due to a lack of access to the data or computational power necessary to handle such massive investigations. This is a critical avenue of research given findings that country of origin is significantly related to how learners engage with MOOCs (Liu et al., 2016; Guo & Reinecke, 2014; Kizilcec, Piech, & Schneider, 2013), implying that learners from different countries behave differently when interacting with educational systems. A study by Liu and colleagues (2016), which was conducted on a course on Big Data and Education, found significant differences in learner interactions between learners from the countries present in their dataset. Their study identified learner profiles based on how they participated in the MOOC (e.g., those who predominantly only took quizzes, only watched videos, etc.), clustered the countries in their dataset based on Hofstede’s (1986) cultural dimensions, and found significantly different learner profile compositions per cultural cluster. They posit that these differences may be due to differing educational traditions observed across cultures. Guo and Reinecke (2014) found that learners were more or less likely to interact with the course in a non-linear manner (i.e., by navigating backwards to a previous module instead of continuing on the sequence) based on their country of origin. Specifically, learners from countries with lower student-teacher ratios (e.g., the US and European countries) were significantly more likely to interact in a non-linear manner than those from higher student-teacher ratios (e.g., Kenya, India).

Ultimately, in order to better support all learners towards success, published findings in MOOC research need to generalize across different learner populations. This leads to this study’s main research question: what country-level measures lead to better or worse generalizability in cross-country predictive modelling? A recent review article in the inaugural issue of the journal *Computer-Based Learning in Context* (Baker, Ogan, Madaio, & Walker, 2019) notes that despite a small number of examples (such as the ones given above), this question has not been systematically investigated by the field, and researchers still do not have a clear idea of what factors to look at. It may be possible to select factors for consideration based on studies that investigate the effectiveness of findings across different groups of students, such as socio-economic status (Buolamwini & Gebru, 2018), national wealth (Kulik & Fletcher, 2016), and whether the student comes from a collectivist or individualist cultural background (Kizilcec & Cohen, 2017). Some studies have suggested that cultural and contextual factors and pedagogical outcomes not only matter but interrelate (see review in Baker et al., 2019), and their combination may dictate what content and methods are most appropriate for given samples and demographic groups. As such, identifying which measures relate or contribute to better (or worse) generalization of models across countries can help us ensure that the models we use for intervention are accurate and appropriate for the full variety of learners being impacted around the world.

The MOOC Replication Framework (MORF)

This study was conducted using the MOOC Replication Framework (MORF) (Gardner, Brooks, Andres, & Baker, 2018), a research platform that has been developed with the goal of reducing technical, data, and methodological barriers to conducting replication studies on MOOCs. For reasons of security, privacy, and data ownership, the data available in MORF is not available for export or download, but instead is available for analysis through a secure platform governed by a data use agreement (Gardner et al., 2018).

This study was conducted using learner data from the University of Pennsylvania. Only courses that were taught primarily in English were used in this study, as other courses tended to have learners from a smaller set of countries. In MOOCs during the time period studied, a course typically ran for a set number of weeks in which learners could enroll, engage in, and earn a completion certificate. Due to demand, some courses were offered multiple times. Each offering or instance of a course is a session. That is, each course could have had multiple sessions, depending on how many offerings were made over the period of time covered in the dataset. This dataset had a total of 45 courses. 27 of these courses had multiple sessions, resulting in a total of 98 sessions. For reasons of security, privacy, and data ownership, the data available in MORF is not available for export or download, but instead is available for analysis through a secure platform which is governed by a data use agreement (Gardner et al., 2018).

The volume of data within the framework allows for the investigation of research questions within data from multiple MOOC datasets, with the goal of determining what findings hold across different courses and iterations of those courses, and which findings are unique to specific kinds of courses and/or kinds of participants (Andres et al., 2018). MORF functionality supports both predictive modelling and production rule analyses (Gardner et al., 2018).

Architecture and Job Submission

MORF allows users to conduct studies by submitting them as *jobs*. To do this, a user must first create and submit a configuration file, either using an HTTP request or MORF's API. This configuration file contains job metadata and includes pointers to 1) an executable containerized image which encapsulates all software dependencies needed to run the experiment, and 2) a Python controller script that specifies the study's high-level workflow, such as how model training and testing should occur and whether cross-validation should be used. The use of controller provides a single script to fully replicate a study, and is human-readable, providing researchers an intuitive overview of the study.

Another core feature of MORF's architecture is its use of executable containerized images as a means of overcoming technological barriers related to computational power and method replication. Researchers may not have access to the computational capacity necessary to conduct large-scale analyses; sending a containerized image of their study to MORF allows the platform to run the analysis on its own servers for them. Doing so also preserves their study's full methodology, ensuring that the same data extraction and model creation is conducted across all of MORF's datasets. These images are lightweight virtual machines that contain software dependencies and the execution environment of an end-to-end study in a single file. After generating this image, the user must then upload it to a public repository. The image's URL is included in the configuration file submitted to MORF, which the platform then fetches and executes according to the workflow specified in the controller script. This combination of using controller scripts and containerization allows users to work with whatever programming language they are most comfortable with. A more technical and in-depth description of the platform can be found in (Gardner et al., 2018).

Data

Learners from countries not present in either the Hofstede or Happiness databases (described below) were dropped from all analyses in the study, resulting in a dataset of over 1.9 million learners across a total of 81 countries (listed in Appendix A).

Measures of National Culture

Culture significantly impacts the way people feel, think, and in the context of education, the way people teach or learn and the support they need (Hofstede, 1986). Hofstede (2005) defines culture as a collective phenomenon that differs across various groups, e.g., across countries, organizations and occupations, genders, generations, etc. This study investigates how cross-country cultural differences relate to the generalizability of completion prediction models.

This study considers two different types of measures in quantifying national culture: Hofstede's cultural dimensions framework (Hofstede, Hofstede, & Minkov, 2010) and overall national happiness, as measured by the World Happiness Report (Helliwell, Richard, & Sachs, 2015). The former is among the more commonly-used cultural frameworks in investigating cultural differences in computer-based learning systems (Baker et al., 2019). The latter, on the other hand, has never been used to directly investigate learning. Instead, it has been used extensively to measure psychological well-being (Lan, Ma, & Radin, 2019), the perceptions of which vary by country and are reported to significantly affect different facets of a person's life, such as their nutrition and education, as well as the conditions that support a person's continued drive to learn (Helliwell, Layard, & Sachs, 2012).

Hofstede's Cultural Dimensions

Hofstede's cultural dimensions are used in this study to more closely examine cross-cultural variations within the learner sample. This initial cultural framework of four dimensions was developed from the survey responses of over 100,000 participants across 70 countries (Hofstede et al. 2010) gathered in 1967 and 1973 on personal values and related sentiments (Hofstede, 2011). Additional data was gathered in the 1980s and 2000s, which led to the calculation and addition of the fifth and sixth dimensions, respectively. Dimension scores, which range from 0 to 120, are currently available online for 107 countries or regions[5]. This dataset was last updated in 2015, which lines up with the final year the MOOCs investigated in this study were active in. This framework has six cultural dimensions:

Power Distance Index (PDI). This dimension describes the distribution of power within organizations and institutions (including the family structure). The inequality of power distribution is viewed from the perspective of those with less authority within

[5] https://geerthofstede.com/research-and-vsm/dimension-data-matrix/

hierarchical systems and these perceptions guide the social perceptions around dependence, organization, and structure within a society (Soares, Farhangmehr, & Shoham, 2007; Hofstede, 2011). High-scoring cultures in this dimension denote a large power distance, where people tend to be deferential to figures of authority and accepting of an unequal distribution of power. Teachers and students in a large power distance classroom acknowledge a power dynamic between them, where the quality of education relies solely on the excellence of the teacher. On the other hand, teachers and students see each other as equals in a small power distance classroom, where the quality of learning depends on the excellence of both the teacher *and* students. People from low power distance cultures readily question authority and expect to participate in decision making (Hofstede et al., 2010).

Individualism vs. Collectivism (IDV). This dimension looks at the different focus of individual relationships within differing cultures. Within this framework, high-scoring cultures are individualistic. People from individualistic cultures are characterized by a tendency to focus on their own needs and those of their immediate family. As a result, social ties to extended family and other individuals are relatively loose. Individuals within a collectivist culture (i.e., low-scoring cultures in this dimension), on the other hand, more often associate with larger social groups and conversely focus on the needs of the group rather than on their own. Collectivist cultures value loyalty and harmony which emerge from strong and cohesive in-groups (Hofstede, 2011). Hofstede (2011) contends that the purpose of learning in individualistic cultures is to learn, whereas the purpose of learning in collectivist cultures is to do.

Gendered Role Index (GRI). This dimension was originally called “Masculinity vs. Femininity” in Hofstede’s (2011) cultural framework. However, I have chosen instead to rename this dimension “Gendered Role Index” for two reasons. First, this dimension focused more on a culture’s adherence to strict gender roles, describing the strict attribution of certain values between genders in a society and how these influence social dynamics. High-scoring cultures in this dimension, where gender roles were more clearly defined (i.e., *gendered*), were also found to be driven by achievement, success, competition, and assertiveness—values this dimension inadvertently attributed to being more *masculine*. On the other hand, low-scoring cultures in this dimension, where gender roles were more likely to overlap, were found to be more caring, modest, and focused on improving society’s quality of life—values this dimension originally attributed to being more *feminine* (Soares et al., 2007; Hofstede, 2011; Hofstede, 1998). These attributions of values to masculinity and femininity are the second reason I chose to rename the dimension. It is a complex issue outside the scope of this study, but much has changed since these attributions were first labeled in this cultural framework in the way society and academic research perceive, treat, and investigate these terms (e.g., Johnson, 2020; Kostas, 2021).

Uncertainty Avoidance Index (UAI). This dimension refers to the social tolerance for ambiguity and uncertainty and the degree to which individuals from this culture would avoid such situations. This dimension involves the degree to which a society has developed rules for prescribed social behaviors as well as the level of comfort or discomfort individuals experience in unstructured situations. High-scoring cultures are uncertainty avoidant, and people in these cultures believe that uncertainty is a “continuous threat that must be fought” (Hofstede, 2011, p. 10). Avoidant cultures tend to minimize such situations through comprehensive behavioral and social codes and an adherence to a common truth (Soares et al., 2007; Hofstede, 2011; Hofstede, 1998).

Low-scoring cultures in this dimension, on the other hand, are uncertainty accepting, more tolerant of others' opinions, and believe that while uncertainty is inherent in life, it is instead treated as a curiosity. As a result, accepting cultures have more relaxed rules and regulations. Teachers in avoidant classrooms are expected to know all the answers, while teachers in accepting classrooms are allowed to say, "I don't know" (Hofstede, 2011, p.10).

Long-Term Orientation vs. Short-Term Normative Orientation (LTO). This dimension describes the inclination of a given culture to focus on future rewards with regard to values such as perseverance and thrift where relationships were ordered by social status and a sense of shame (Hofstede, 2011). High-scoring, long-term oriented cultures are focused on the future and willing to delay short-term success. They place importance on values like thrift, perseverance, adaptability (Hofstede & Minkov, 2010). Lower-scoring, short-term oriented cultures, on the other hand, often give importance to the past and present, valuing reciprocity in social obligations, respect for tradition, personal steadiness, and the fulfillment of social obligations. Cultures with long-term orientation, on the other hand, are focused on the future and willing to delay short-term success. These cultures value thrift, perseverance, and adaptability (Hofstede, 2011).

Indulgence vs. Restraint (IND). The final dimension, which was added to the framework in 2010, refers to the social perceptions around human desires and gratification in comparison to regulation and strict social norms. Hofstede (2011) reports that this dimension focuses on aspects not covered by the previous five, and "known from literature related to *happiness research*" (p. 15). Individuals from more indulgent cultures score higher on this scale and are described as having a strong sense of personal control, valuing leisure, and more lenient sexual norms. On the other hand, individuals from more restrained cultures tend to value policing, strict social norms, and etiquette (Hofstede, 2011).

These dimensions have been widely cited across multiple disciplines, including psychology, sociology, education, and marketing (Soares et al., 2007; Søndergaard, 1994; Steenkamp, 2001). They have been used to analyze and explain differences in various behaviors in educational technology (Kizilcec & Cohen, 2017; Ogan et al., 2015). In their study on help-seeking model transfer between countries, Ogan and colleagues (2015) hypothesized that their mixed results were due to differences among the three countries in Hofstede's cultural dimension on adherence to gender roles. Specifically, they speculated that the poor model transfer may have been due to the US and the Philippines both scoring high and Costa Rica scoring low. In this dimension, high-scoring nations are driven by competition, achievement, and success, while low-scoring nations are more concerned with care for others and quality of life.

Kizilcec and Cohen (2017) investigated the efficacy of a self-regulation strategy between countries on opposite ends of the Hofstede's *individualism* dimension. In this dimension, high-scoring nations tend to be more collectivist by nature, placing importance on the goals and well-being of the group. Low-scoring nations, on the other hand, were more individualistic, placing a premium on the importance of personal goals. The study noted that this strategy was developed in Western countries, appealing to more individualist tendencies. Their study found this strategy to significantly improve completion rates among learners from individualist countries (like the US, Australia, and France), but had no effect on learners from collectivist countries (like India, China, and Mexico). The findings of their study highlight how even highly efficacious interventions may be culturally bounded in their effects.

More recently, a study by Muthukrishna and colleagues (2020) sought to measure and investigate cultural distances between societies, noting the dominance of WEIRD subjects in psychological data—particularly from the United States. Their study devices a statistic to compute cultural distance using individual-level data drawn from responses to the World Values Survey of cultural beliefs (Inglehart et al., 2014), focusing primarily on cultural distance from the US. In relating their findings to Hofstede's dimensions, they found that distance from the US was most strongly correlated with Hofstede's *individualism* scale, reporting that collectivist countries were further away from the US on their scale. They found that countries with a larger power distance and more restrictive (vs. indulgent) countries were further away from the US as well.

Gross National Happiness

Another country-level metric considered in this study is Gross National Happiness (GNH) or overall societal happiness, as reported in the World Happiness Report (Helliwell et al., 2015), an annual publication of the United Nations Sustainable Development Solutions Network. This report contains an index of national happiness based on respondent ratings of their lives within a given country. The happiness index measures self-reported satisfaction across a range of dimensions including their country's gross domestic product (GDP), social support, health and life expectancy, freedom to make life choices, generosity, and perception of corruption. Country-level economic variables such as unemployment and inequality are excluded from the calculation of the GNH since these values are not readily available across all countries (Helliwell et al., 2015). The GNH surveys make use of the Cantril ladder, a conceptualization of one's life across the length of a ladder, with the best possible life (scored with a 10) at the top of the ladder, and the worst (scored at 0), at the bottom. The rankings are collected from nationally representative samples. The World Happiness Report publishes the estimated extent to which each of the six factors contribute to societal happiness. For this study, the GNH values used were from 2015 to match both the final year the MOOCs in this study were active, and the other country-level measures pulled, as described in the following sections.

Additional Country-Level Measures

In addition to pulling each country's set of Hofstede's cultural dimension indices and happiness index, the measures outlined in Table 5 were also pulled per country. The latter three measures were pulled from publicly available data from the World Bank[6] for the year 2015 to match the final year in which the courses in the dataset were active.

Research Design

In this study, I considered course completion as the metric of learner success. The study recognizes that not all learners enroll in MOOCs with goal of completing. For example, some seek to gain just enough knowledge to publish in their field (Wang & Baker, 2018) or attain various job-related benefits (Trumbore, 2020), while others form connections in order to join a professional society (Wang & Baker, 2018). Course completion, however, continues to be the most researched and widely used metric of success in MOOCs.

[6] https://data.worldbank.org/indicator/

Table 5. Country-level measures investigated.

Measure	Definition
Enrollment Size	The current study made use of enrollment sizes by country, taken from the number of unique users from each country across the entire dataset (derived from data in MORF).
National Population	The population of a country is defined as the number of people living within its borders, as measured by national level censuses (Eurostat, 2020).
Gross Domestic Product (GDP)	The monetary value of all final goods and services produced within a country within a specified amount of time, most often a year. This value considers all production within a national economic zone, including goods and services produced for both market and nonmarket based products (e.g., defence and education). National GDP is often calculated by national statistical agencies, upheld by standards compiled by the World Bank (Callen, 2020).
Per Capita GDP	GDP, by itself, is highly correlated to population. GDP per capita approximates the standard of living of citizens of a given country.

This study was divided into three phases. The first phase establishes best-performing completion models per country. The second phase considers the *distance* between every country pair (i.e., a training country and a testing country), by comparing the cross-country model performance with the training country's own within-country, baseline model performance. Finally, the third phase seeks to explore the relationship between the cross-country distances and several country-level measures.

Phase 1: Within Country Models and Baseline Performances

Methodology

Data Cleaning and Feature Engineering

Features were extracted by querying the MORF database. First, completion and country of origin were pulled per learner. Completion was assessed based on the learner’s achievement type. There were three possible achievement types in the dataset: “none”, “normal”, and “distinction.” Learners with no achievement—“none” in the database—were non-completers, i.e., learners who either dropped out or failed to attain a passing mark. “Normal” learners attained at least a passing mark in a course, while “distinction” learners attained a high enough mark to earn a distinction (typically a final grade of 85% or higher). Both “normal” and “distinction” learners were treated as completers.

The learners’ IP addresses were used to geolocate their country, labelled using MaxMind’s GeoIP2 Precision Country Service API[7]. Learner IP addresses were pulled from clickstream data. In the cases where multiple IP addresses were used by a learner, the IP address that was used the most was the one attached to the learner. GeoIP2 labels the country of an IP address based on GeoNames[8] geographical data. Dependencies, such as overseas territories (e.g., Bonaire, Sint Eustatius, and Saba/the Caribbean Netherlands), and constituent countries (e.g., Curaçao, constituency of the Netherlands); and Areas of Special Sovereignty or autonomous territories (e.g., Puerto Rico, territory of the US; Saint Barthélemy, territory of France) are labelled by GeoNames separately from their governing countries. As such, all analyses treated dependencies as separate from their governing countries.

Official start and end dates were pulled per session and were used to compute the total number of days each session was active. In order to conduct analyses across sessions, session lengths were divided into eight equal increments, ranging from 3.5 days (i.e., three days and 12 hours) to 11.375 days (i.e., 11 days and nine hours), with a

[7] https://www.maxmind.com/en/geoip2-precision-country-service
[8] http://www.geonames.org/

median of 6.125 days and a standard deviation of 2.26. Its distribution can be found in Figure 2. The start and end dates and times of these increments were used in conducting feature engineering.

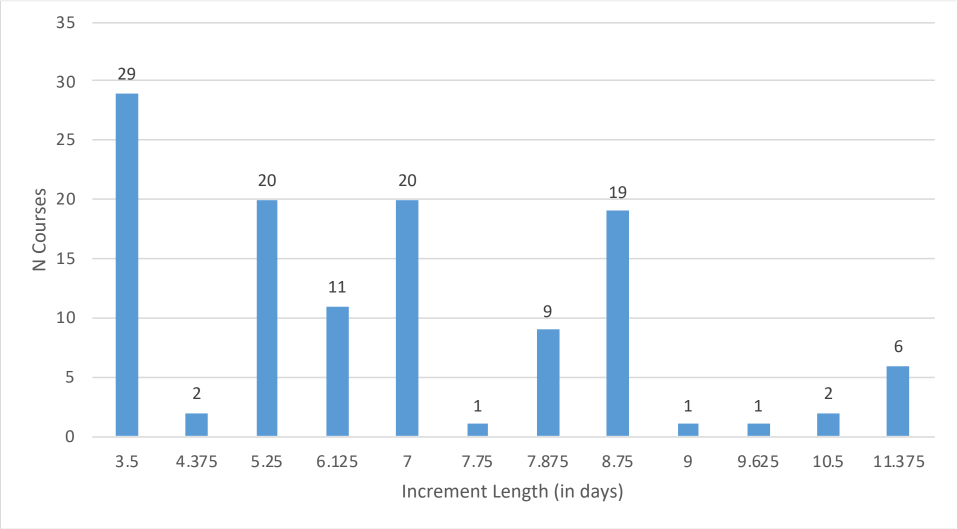


Figure 2. Distribution of courses by increment length in days.

In each course, learners used several resources, e.g., the discussion forums, quizzes, peer assessments, and lecture videos. Each interaction was tracked in the course’s clickstream log, which also contained date, time, and URL information. The features investigated in this phase of the study are listed below. These features are pulled per learner and broken down by increment (Table 6).

Each of the features were normalized through z-score transformations to account for the variability of session durations, likely resulting in larger raw incremental pageview counts for learners in longer-running courses. Doing so also allowed for both the aggregation of learners by country and the comparison of learner features across course sessions. This was done using each feature’s respective session mean and standard deviation values. For example, if learners A, B, and C were the only learners in a course session, and A visited the forums 36 times, B visited 20 times, and C visited 41 times, then the session's mean forum visit count would be 32.333 and its standard deviation would be 10.970. After normalizing, Learner A's forum visit z-score would have a value of 0.334, B would be -1.124, and C would be 0.790.

Table 6. Incremental features used in building completion prediction models.

Feature	Definition
Forum Views	Total number of clicks related to any forum activity (e.g., viewing, posting, commenting)
Quiz Views	Total number of clicks related to any quiz activity (e.g., viewing, answering, submitting)
Peer Assessment Views	Total number of clicks to any peer-assessment-related activity
Lecture Video Views	Total number of clicks related to any video lecture activity (e.g., playing, pausing, increasing video speed, etc.)
Days Active	Total number of days active
Forum Threads Started	Total number of forum threads started
Responses	Total number of responses to others' forum posts
Respondents	Total number of others' responses on one's own forum posts
Time Spent	Time spent (in seconds) in the forums, quizzes, peer assessment, and video lectures; actions with a computed duration of over one hour were treated as disengagement and excluded from the sum

Predictive Modeling

In order to determine the best-fitting model per country, three different classifiers were used to build completion prediction models using the scikit-learn and xgboost libraries in Python: CART, Random Forest (RF), and XGBoost (XGB). CART (Classification and Regression Trees) is scikit-learn's implementation of both decision trees and regression trees. Since the models predicted a categorical label (i.e., completion), only the former was used. Random Forest is an ensemble classifier that generates multiple decision trees while training a model. Its output is the class selected by the majority of its decision trees. It also produces a list of feature importances, which can be graphed in order to visualize how important each feature is to the model's classification (Zhang et al., 2019). Finally, the Extreme Gradient Boosting (XGBoost) classifier (Chen & Guestrin, 2016) uses an ensemble technique in which an initial, weak decision tree is trained, and its prediction errors are calculated. Subsequent decision trees are then trained iteratively to predict the error of the decision tree before them. The final prediction is the sum of the predictions of all the trees in the set (Chen & Guestrin, 2016).

Informal hyperparameter tuning was conducted on the RF and XGB classifiers in order to determine which value for n_estimators was optimal for the dataset. Hyperparameters are parameters that are set before and are used to control the classifier's learning process. N_estimators, for example, is a hyperparameter used to determine how many trees will be used in the process of training the model. Hyperparameter tuning was conducted on data from three representative small (Mauritius, MU, *N*=1008), medium (Egypt, EG, *N*=20368), and large (GB, United Kingdom, *N*=70260) countries. Five values for n_estimators were tested per classifier: 100 (default), 300, 500, 700, and 900. The process revealed the following to be optimal across all three countries, feature sets, and increments: n_estimators=700 for RF and n_estimators=100 for XGB. These values were then set for whenever either classifier was used to train and test predictive models.

In training and testing the models, the classifiers iterated through the eight increments of the course in predicting learner completion, beginning with features from only the first increment of the course, then moving on to features from the second increment, and so on. At each iteration, two feature sets were tested: 1) increment-only: features from only the current increment ($N_{features}$=13) and 2) appended: features from the current and all previous increments ($N_{features}$=13 * increment number). Per classifier-feature set-increment combination, 10-fold cross-validation was conducted, repeatedly building the model on some learners' data and testing it on other learners' data. Stratified sampling was used in determining the folds in order to preserve completion rates. Fold-level models were pickled (i.e., saved to file) for the ensuing cross-country analysis. A total of 480 models were trained and tested per country, ten (one per fold) for each combination of classifier, feature set, and increment.

The performance of each model was assessed using the Area Under the Receiver Operating Characteristic Curve (AUC ROC or A'), which is the probability that given 1 instance of 'completed' and 1 instance of 'not completed', the model is able to tell which instance is which. An AUC ROC of 0.5 indicates chance level of performance, while a value of 1 means perfect accuracy. AUC ROC scores were averaged across each classifier-feature set-increment combination's respective ten folds. In order to determine each country's best performing model, averaged AUC ROC scores were compared across increments in each classifier-feature set combination using the statistical testing procedure from (Fogarty, Baker, & Hudson, 2005), which utilizes the equivalence of the AUC ROC scores and the Wilcoxon statistic to generate a test statistic Z to evaluate a null hypothesis of equivalent performance between two predictive models, as seen in Equation (1).

$$Z = \frac{AUC\ ROC_1 - AUC\ ROC_2}{\sqrt{SE(AUC\ ROC_1)^2 - SE(UC\ ROC_2)^2}} \qquad (1)$$

This was performed by iteratively comparing the AUC ROC of an increment with the AUC ROC of all future increments. If any comparison came out significant after conducting a Bonferroni correction (Dunn, 1961), then that increment was not the best performing model. Otherwise, if no comparisons came out significant, that increment provided the best performing model. Models requiring data from Increment 8 (i.e., the final increment) were dropped from consideration for two reasons:

1. Having to wait for the final increment of a course was counterintuitive to the goal of predicting learner completion, and
2. Models that used the appended feature set in Increment 8 outperformed all other incremental models 100% of the time due to their use of the data of the *entire* course run.

The comparisons resulted in a final selection of six AUC ROC scores per country, one for each classifier and feature set combination. From here, the best performing completion prediction model was chosen per country, and its AUC ROC was treated in the subsequent analyses as the country's baseline model performance.

Relationship Mining

Nonparametric correlations were conducted between the country's baseline model performances and the set of country-level measures. The Benjamini-Hochberg (1995) post-hoc correction was conducted to account for the number of correlations conducted.

Linear regression was then conducted to determine whether each country's country-level measures were predictive of their baseline model performances. Two linear models were fit, the first using only the countries' six cultural dimension indices, and the second using the remaining measures (i.e., happiness index, enrollment size, population size, GDP, and per capita GDP). Due to the high correlations between the country-level measures (Table 12), stepwise backward selection was conducted to account for collinearities and to remove suppression effects in both linear models using the step function in R’s stats library. This function searches for the best possible regression model by iteratively selecting and dropping variables to arrive at a model with the lowest possible AIC (Akaike Information Criteria; Bozdogan, 1987), an estimator of a model’s quality relative to other models built on the same dataset.

Results

Baseline AUC ROC scores across the 81 countries ranged from 0.874 (Iraq) to 0.992 (China), with a median of 0.979. The summary of a descriptive analysis of the predictive models can be found in Table 7. The large difference between the XGB-appended combination and all other combinations warranted further investigation. The descriptive results of the breakdown of countries using the XGB-appended combination by best-performing increment can be seen in Table 8. As a reminder, increments span an eighth (i.e., 12.5%) of each course, where Increment 1 is the first eighth, Increment 2 is the second eighth, and so on. Interestingly, despite the majority of models in this category performing their best using data until Increment 4 (i.e., until halfway through the course), countries with larger enrolment sizes required more data, as evidenced by the substantial leap in the mean enrolment size of countries needing data from either Increments 5 or 6. These numbers show that the majority of the countries’ models were able to predict learner completion using data until just Increment 4 (halfway through the course).

Table 7. Descriptive results of the parameters used in the best performing models.

	Increment Only	Appended	Total
Random Forest	5	18	23 (28%)
XGBoost	5	53	58 (72%)
Total	10 (12%)	71 (88%)	81

Note. Parameters presented across the different combinations of classifiers (rows) and feature sets (columns). Each combination reports the number of countries whose best performing model used the respective combination.

Table 8. Descriptive results of the increments used in the best performing XGB-appended models.

Increment	*N* Countries
1	1 (2%)
2	3 (6%)
3	4 (8%)
4	31 (58%)
5	6 (11%)
6	8 (15%)

Note. Each row reports the number of countries whose best performing XGB-appended model uses the respective increment, *N*=53.

Table 9. Correlation results between baseline AUC ROC scores and the country-level measures.

Measure	Correlation, *rho*
Enrolment Size	0.880 *
Gross Domestic Product	0.765 *
Long-Term/Short-Term	0.480 *
Per capita GDP	0.466 *
Individualist/Collectivist	0.423 *
Happiness	0.354 *
Population	0.353 *
Completion Rate	0.246 *
Gendered Role Index	0.221
Power Distance	-0.219
Indulgence/Restraint	0.120
Uncertainty Avoidance	-0.093

* *p*<.001 and significant after Benjamini-Hochberg (1995) correction

Correlation Analysis

Nonparametric correlations were conducted between the country's baseline performances and the set of country-level measures. The Benjamini-Hochberg (1995) post-hoc correction was conducted to account for the number of correlations conducted. Results show that enrollment size was significantly positively related with baseline model performance (*rho*=.880, *p*<.001), suggesting that as enrollment size increased, so did baseline model performance. Measures of country wealth were also among the most strongly correlated with baseline model performance (GDP: *rho*=.765, *p*<.001; per capita GDP: *rho*=.480, *p*<.001), suggesting that better-performing predictive models are obtained by wealthier countries. Happiness (*rho*=.354, *p*=.001) and cultural dimensions that look at individualism/collectivism (*rho*=.423, *p*<.001) and long-term/short-term orientation (*rho*=.480, *p*<.001) also had significant positive relationships with model performance, suggesting that better-performing models were obtained for happier, more individualistic, and more long-term oriented countries. The full results can be found in Table 9.

Regression Analysis

In order to further investigate this relationship, linear regression was conducted to determine how predictive a country's country-level measures were of their baseline model performance. Two linear models were fit on the country-level dataset (*N*=81) to estimate the effect of the country-level measures on each country's baseline model performance.

The first model regressed AUC ROC scores on Hofstede's dimension indices. Feature selection revealed that long-term/short-term orientation (LTO; *F*(1, 78)=13.114, *p*<.01) and individualism/collectivism (IDV; *F*(1, 78)=4.806, *p*=.031) were most relevant to model performance. A model that regressed AUC ROC scores on indices from just these two dimensions revealed that only long-term/short-term orientation significantly predicted baseline model performance (*β*=.39, *p*<.001).

Feature selection on the country-level measures of happiness, wealth, and size, and revealed that a country's self-reported happiness (*F*(1, 78)=20.123, *p*<.001) and population (*F*(1, 78)=9.796, *p*=.002) were most relevant to model performance. A second model, which was regressed on just these two measures, revealed that both were predictive of model performance within the full model (happiness: *β*=.51, *p*<.001; population: *β*=.31, *p*=.002). The results of fitting both linear models can be found in Table 10.

Table 10. Within-country model performance regression results.

Predictors	*β*	*p*	*β*	*p*
(Intercept)	-0.00	<0.001	-0.00	<0.001
LTO	**0.39**	<0.001		
IDV	0.13	0.239		
Happiness			**0.51**	<0.001
Population			**0.31**	0.002

Phase 2: Cross-Country Model Distances

Methodology

The study next considered how models trained in Phase 1 performed when classifying instances from data other than the training country. First, a list of all possible training and testing country pairs was compiled, resulting in a total of 6480 pairs (81 training countries x 80 testing countries). Prediction modeling in this phase iterated over all train-test country pairs. In each iteration, the details of the training country's predictive model were pulled (i.e., feature set and increment) and applied to the testing country's dataset. Each of the training country's 10 fold-level models were then loaded from file and tested on the test country data. This resulted in ten AUC ROC scores, which were averaged to determine the models' cross-country performance. Finally, *distances* between country pairs were computed by subtracting the cross-country AUC ROC score from the training country's baseline performance, as seen in Equation (2). Distances track how well the training country's predictive model generalized to the testing country's data. A negative distance implies that the model performed better cross-country, while a positive difference implies worse performance.

$$distance = AUC\ ROC_{baseline} - AUC\ ROC_{cross} \qquad (2)$$

Correlation Mining

Nonparametric correlations were conducted between the cross-country AUC ROC scores (raw AUC scores, not differences) and the training country's country-level measures. The Benjamini-Hochberg (1995) post-hoc correction was conducted to weed out any spurious findings that could have emerged as a result of the number of correlations conducted.

Results

Cross-country AUC ROC scores ranged from 0.747 (Iraq→Mauritius) to 0.993 (Brazil→Luxembourg), with a median of 0.973 across the 6480 country pairs.

Distances, on the other hand, ranged from -0.042 (Lebanon→Ethiopia) to 0.217 (Netherlands→Mauritius), with a median distance of 0.005 across the 6480 country pairs. Negative distances represent cases wherein the cross-country performance outperformed the training country's baseline model performance. For example, the performance of Lebanon's model on Ethiopia's data (*AUC ROC*=0.976) outperformed Lebanon's own baseline model performance (*AUC ROC*=0.935), resulting in a distance of -0.042. The distribution of distances can be found in Figure 3.

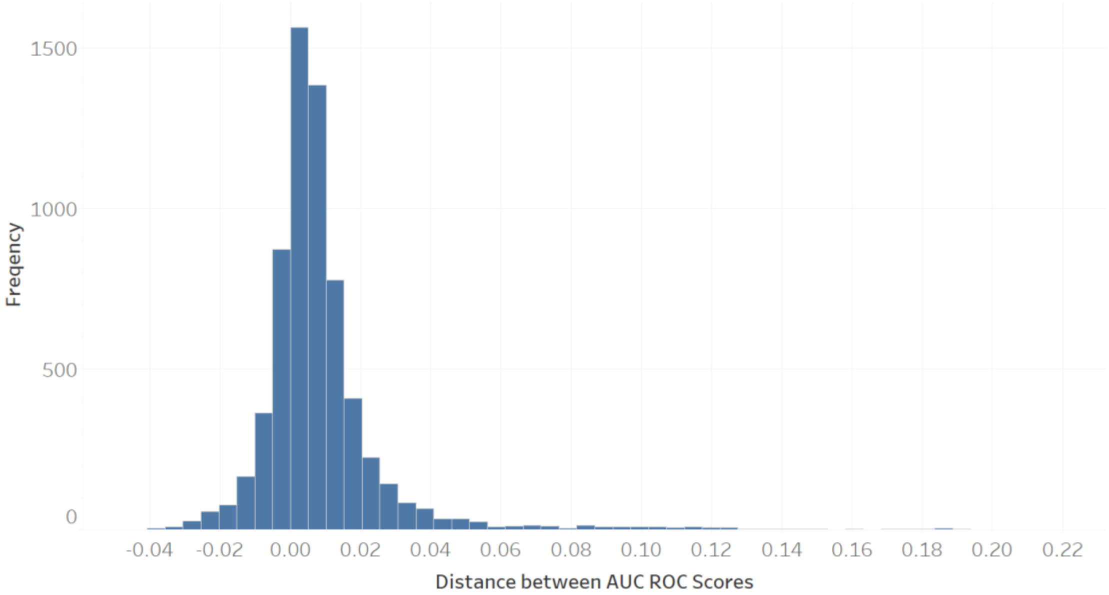


Figure 3. Distribution of cross-country model distances.

Correlation Mining

Nonparametric correlations were conducted between each training country's mean cross-country AUC ROC score and the training country's country-level measures (Table 11). The Benjamini-Hochberg (1995) post-hoc correction was conducted to weed out any spurious findings that could have emerged as a result of the number of correlations conducted. The training country's enrollment size (i.e., its number of training data points) was the most strongly correlated with mean cross-country model performance (*rho*=.846, *p*<.001), suggesting that, despite our hypothesis that differences in demographic and cultural factors lead to degraded model performance, models trained on countries with a large enrollment size are able to perform well on data from other countries. Measures of country wealth also strongly related to mean cross-country performance (GDP: *rho*=.732, *p*<.001; per capita: *rho*=.311, *p*=.005), suggesting that wealthier countries are also able to produce more generalizable models.

Table 11. Correlation results between cross-country model performance and training country country-level measures.

Measure	Correlation, *rho*
Enrolment Size	0.846 **
Power Distance	-0.019
Individualist/Collectivist	0.265 *
Gendered Role Index	0.320 **
Uncertainty Avoidance	0.035
Long-Term/Short-Term	0.304 **
Indulgence/Restraint	0.189
Gross Domestic Product	0.732 **
Happiness	0.201
Population	0.430 **
Per Capita	0.311 **

* *p*<.05, ** *p*<.001 and significant after Benjamini-Hochberg (1995) correction.

Phase 3: Understanding Model Distances

In this third phase of the study, we explore the relationship between the cross-country distances and the differences in the country-level measures in order to analyze how each measure relates to model generalizability.

Methodology

Correlation Mining

Correlation mining (Baker, 2020) was conducted to investigate relationships that exist among the country-level measures. Nonparametric correlations that came out significant after conducting the Benjamini-Hochberg (1995) correction were used to describe the profile of countries that exist in the dataset.

Correlation mining was then conducted again to investigate whether relationships exist between any of the country-level measures and the cross-country distances. A data frame was compiled listing all train-test country pairs, and their respective country-level measures. Per row, the differences of each measure were computed (e.g., difference in the IDV cultural dimension, difference in GDP, difference in population size, etc.). Nonparametric correlations were then conducted between each of the differences and the cross-country distances using SPSS.

Regression Analysis

The goal of this analysis was to further investigate the relationship between the feature differences and the cross-country distances. Linear mixed-effects models were fit on the cross-country dataset (*N*=6480) to determine whether the set of feature differences was predictive of cross-country distances. Two mixed-effect models were fit on the data, the first using only the differences across the countries' six cultural dimension indices, and the second using the differences of the remaining measures. Training country was set as a random factor. As was done previously, feature selection was conducted due to the high correlations between the cross-country measure differences (Table 13). Backward elimination was conducted to eliminate non-significant effects in both linear mixed-effects models using the step function in the lmerTest R library. This algorithm starts with the full model and eliminates variables iteratively, first across the random-effect features, then across the fixed-effect features.

Results

Country Profiles

Nonparametric correlations were conducted on SPSS across all nation-level measures in order to investigate their relationships with one another. The Benjamini-Hochberg (1995) post-hoc correction was used to weed out findings that were likely to be spurious due to the number of tests conducted. Correlation results can be found in Table 8.

MOOC Presence. A country's enrollment size (i.e., its number of data points) was significantly positively related to its GDP (*rho*=.816, *p*<.001) and population (*rho*=.582, *p*<.001), as well as two of Hofstede's cultural dimensions: long-term vs. short-term orientation (LTO; *rho*=.331, *p*=.003) and gender role adherence (GRI; *rho*=.284, *p*=.01). Further, GDP was significantly positively related to population (*rho*=.638, *p*<.001) and both cultural dimensions (LTO: *rho*=.287, *p*=.009; GRI: *rho*=.335, *p*=.002). Together, these findings suggests that larger and wealthier countries, which are more long-term oriented and more strictly adhered to gender roles, are more likely to have a presence in MOOC platforms. Countries like the US, China, Japan, Germany, and the United Kingdom fit this profile of having high MOOC presence. On the opposite end of the spectrum, countries like Guatemala, Uruguay, Costa Rica, and Mauritius fit the profile of having lower MOOC presence.

Wealthy Countries. Two other cultural dimensions were found to relate significantly with GDP: individualism vs. collectivism (IDV; *rho*=.33, *p*=.003) and indulgence vs. restraint (IVR; *rho*=.263, *p*=.018), as well as self-reported national happiness (*rho*=.28, *p*=.011). These findings suggest that wealthier countries are:

1. more populous,
2. happier,
3. more strictly adherent to distinct gender roles,
4. long-term oriented, valuing thrift, perseverance, and preparing for the future,
5. more individualistic, interested primarily in their own welfare and the welfare of their immediate family, and
6. more indulgent, valuing leisure and the gratification of human needs.

In addition to the countries with high MOOC presence listed above, which also fit the profile of wealthy countries, France and India are also on this list. Countries like Tanzania, Lebanon, Jordan, and Uganda fit the profile of less wealthy countries on the opposite end of the same spectrum.

Happy Countries. Interestingly, however, national happiness was significantly *negatively* correlated with population size (*rho*=-.383, *p*<.001), suggesting that larger countries tend to self-report lower levels of happiness. Happiness also had significant relationships with various cultural dimensions. The findings suggest that happier countries were likely more individualistic (*rho*=.48, *p*<.001), more indulgent (*rho*=.336, *p*=.002), and valued a more distributed form of power—lower Power Distance Index or PDI—where its less powerful members of society expect to participate in decision making (*rho*=-.575, *p*<.001). Countries like Finland, Denmark, Norway, the Netherlands, and Switzerland best fit this profile of happy countries. Tanzania, Uganda, Croatia, Lebanon, and Bulgaria are countries on the opposite end of the spectrum, the less happy countries.

Completers. Finally, a country's completion rate (i.e., the number of completers divided by its enrollment size) was significantly related to its population size (*rho*=-.473, *p*<.001), its per capita GDP (*rho*=.618, *p*<.001), and its self-reported national happiness (*rho*=.563, *p*<.001), suggesting that learners from smaller, happier countries with higher average income are more likely to complete. Completion rate was also significantly correlated with a number of cultural indices: PDI (*rho*=-.401, *p*<.001), IDV (*rho*=.55, *p*<.001), and LTO (*rho*=.286, *p*=.01). These findings suggest that learners from countries that are more individualistic and long-term oriented and had more distributed views on hierarchy and power were likely to complete a MOOC. Countries like Luxembourg, Spain, the Netherlands, and Sweden best fit this profile of countries with higher completion rates. Bangladesh, Egypt, Morocco, Ethiopia, and Iran best fit this profile of countries with lower completion rates.

Correlation Mining

Nonparametric correlations were conducted between the cross-country distances and differences in the country-level measures, and the Benjamini-Hochberg (1995) post-hoc correction was used to account for the number of comparisons conducted. Correlations were also conducted between distance and the *absolute* country-level measure differences in order to assess whether simply the presence of a difference mattered, or the direction of a difference mattered. The results of this analysis can be found in Table 14.

Table 12. Correlations between factors used to study model generalization.

	PDI	IDV	GRI	UAI	LTO	IVR	GDP (B)	Happy	Pop	Per Capita
Enroll Size	-0.045	**0.231***	**0.284***	-0.076	**0.331****	0.132	**0.816****	0.173	**0.582****	**0.251***
PDI		**-0.554****	0.071	0.146	-0.205	**-0.300****	-0.153	**-0.575****	**0.248***	**-0.500****
IDV			0.165	-0.184	**0.320****	0.180	**0.330****	**0.480****	-0.119	**0.608****
GRI				-0.137	-0.039	0.151	**0.335****	-0.060	**0.310****	0.064
UAI					-0.022	-0.191	-0.112	0.053	-0.102	-0.056
LTO						-0.209	**0.287****	0.175	-0.005	**0.405****
IVR							**0.263***	**0.336****	0.033	0.210
GDP (B)								**0.280***	**0.638****	**0.352****
Happy									**-0.383****	**0.769****
Pop										**-0.423****

* *p*<.05, ** *p*<.001 and significant after Benjamini-Hochberg (1995) correction.

Table 13. Correlations between differences in the factors used to study model generalization.

	PDI	IDV	GRI	UAI	LTO	IVR	GDP (B)	Happy	Pop	Per Capita
Enroll Size	0.018	**0.256****	**0.258****	**-0.086****	**0.281****	**0.126****	**0.805****	**0.151****	**0.567****	**0.191****
PDI		**-0.548****	**0.167****	**0.221****	**-0.116****	**-0.258****	**-0.111****	**-0.518****	**0.260****	**-0.446****
IDV			**0.146****	**-0.132****	**0.304****	**0.216****	**0.361****	**0.535****	**-0.121****	**0.601****
GRI				0.005	**0.033****	**0.123****	**0.295****	**-0.078****	**0.282****	**0.052****
UAI					0.006	**-0.173****	**-0.067****	**0.072****	**-0.122****	**-0.108****
LTO						**-0.184****	**0.328****	**0.178****	**0.044****	**0.358****
IVR							**0.218****	**0.288****	-0.003	**0.209****
GDP (B)								**0.243****	**0.593****	**0.294****
Happy									**-0.358****	**0.741****
Pop										**-0.365****

** *p*<.001 and significant after Benjamini-Hochberg (1995) correction.

Table 14. Correlation results between cross-country model performance and training country country-level measures.

Difference In	Correlation with Difference	Correlation with Absolute Difference
Enrolment Size	0.016	0.050*
Power Distance	-0.249*	-0.010
Individualist/Collectivist	0.306*	-0.014
Gendered Role Index	-0.046*	-0.008
Uncertainty Avoidance	-0.057*	-0.019
Long-Term/Short-Term	0.288*	-0.006
Indulgence/Restraint	-0.128*	-0.009
Gross Domestic Product	0.033*	0.006
Happiness	0.208*	0.055*
Population	-0.142*	0.049*
Per Capita GDP	0.296*	0.009

* *p*<.001 and significant after Benjamini-Hochberg (1995) correction.

If the direction of a difference didn't matter—if just the presence of a difference mattered—then the absolute difference analysis would have resulted in a stronger correlation than the difference analysis. However, these results suggest that the direction of difference is more important than the absolute difference in these variables between countries (e.g., Figure 4), except for differences in enrollment size.

Differences in power distance, adherence to gender roles, uncertainty avoidance, and indulgence were significantly negatively correlated with cross-country model distances. These findings suggest that models trained on data from countries scoring high in these dimensions are likely to generalize (i.e., have a lower distance) on data from countries scoring low in the respective dimension, but not the other way around (e.g., indulgent country to restrictive country). Differences in happiness, individuality, and long-term orientation, on the other hand, were significantly positively correlated with model distance, suggesting that the *lower* in these dimensions the training country scored compared to a testing country, the more generalizable their models (e.g., less happy country to happier country).

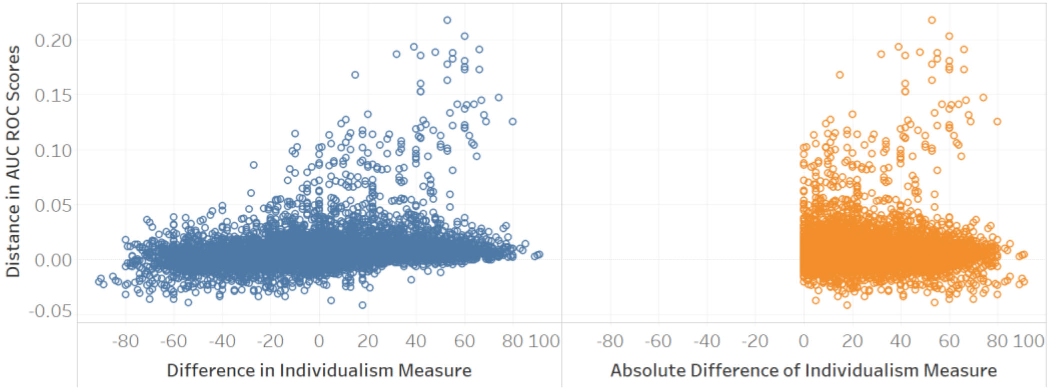


Figure 4. Graphical representation of distance x difference and absolute difference in measures of individuality.

Regression Analysis

In order to further investigate the relationship between the feature differences and the cross-country distances, regression analyses were conducted to measure the effects of each country-level measure difference on cross-country distance. Two linear mixed-effects models were fit on the country-pair dataset (*N*=6480) to estimate the effect of the cross-country measure differences on each pair's distance, with the pair's training country as the model's random factor. The results can be found in Table 15. The first model was regressed on differences related to Hofstede's six cultural indices. After backward elimination, only the Gendered Role Index was dropped from the model.

Table 15. Cross-country distance regression results.

Predictors	*β*	*p*	*β*	*p*
(Intercept)	**-0.00**	<0.001	-0.00	<0.001
Power Distance	**-0.18**	<0.001		
Individualist/Collectivist	**0.08**	<0.001		
Uncertainty Avoidance	**0.07**	<0.001		
Long-Term/Short-Term	**0.18**	<0.001		
Indulgence/Restraint	**-0.07**	<0.001		
Enroll Size			**-0.14**	<0.001
Happiness			**0.08**	<0.001
GDP ($B)			**0.14**	<0.001
Population			**-0.03**	0.040
Per Capita			**0.18**	<0.001

In order to understand the relationships implied by the coefficients, Table 12 contains worked examples of four cases:

1. When the feature difference is positive and the coefficient is negative, the resulting effect on the predicted distance is a negative value, decreasing the distance, thus implying a more generalizable model from train to test country.

2. When the feature difference is negative and the coefficient is negative, the resulting effect on the predicted distance is a positive value, increasing the distance, thus implying a less generalizable model from train to test country.
3. When the feature difference is positive and the coefficient is positive, the resulting effect on the predicted distance is a positive value, increasing the distance, thus implying a less generalizable model from train to test country.
4. When the feature difference is negative and the coefficient is positive, the resulting effect on the predicted distance is a negative value, decreasing the distance, thus implying a more generalizable model from train to test country.

Differences in views on power distance and indulgence/restraint had significant negative effects on the cross-country distances, as in Table 16(1). This implies that as the training country ranked higher in either dimension (i.e., $\text{index}_{\text{train}} > \text{index}_{\text{test}}$) and the country pairs' views of that dimension diverged (i.e., greater difference), the more generalizable the models were (i.e., the lesser the distance). In other words, these findings imply that data trained on learners from more indulgent countries or countries where a hierarchy of power is more accepted are likely to generalize on data gathered from their neighbors on the opposite end of the respective dimension (i.e., the more restrictive countries or countries that are more accepting of distributed power).

The opposite was true for the other three dimensions (e.g., Table 16(3)): as the training country ranked higher in either dimension and the country pair's views in that dimension diverged, the less generalizable the models were (i.e., the greater the distance). This finding implies that data gathered on learners from more collectivist, uncertainty-accepting, and short-term oriented are more likely to generalize to their respective counterparts, but not the other way around. Despite the statistical significance of the effects of these cultural index differences, however, they only explain a very small percentage of the variance in the cross-country distances, R^2=.101.

Table 16. Worked examples for negative and positive cross-country distance regression coefficients.

	Train Val	Test Val	Diff Val	Coefficient	Effect on Predicted Distance
(1)	80	49	31	-0.18	-5.58
(2)	55	77	-22	-0.18	3.96
(3)	33	25	8	0.14	1.12
(4)	40	83	-43	0.14	-6.02

The second model was regressed on the other cross-country measure differences—differences in enrolment size, GDP, self-reported national happiness index, population, and per capita GDP. Despite the high collinearity between features (Table 13), all differences were included in the final model. Differences in enrolment size and population had significant negative effects on the cross-country distance. This implies that the more populous the training country was, or the more learners from the training country were enrolled compared to the test country, the more generalizable the models were. Conversely, the happier or wealthier the training country was compared to the test country, the less generalizable the models were. As in the Hofstede model, despite the statistical significance of the effects of these country-level measure differences, they only explain a very small percentage of the variance in the cross-country distances, R^2=.067.

Discussion

In this study, we examined how country or cultural factors affect the cross-country generalizability of predictive models in MOOC research. We did this by first determining each country's earliest best performing completion prediction models. This was conducted to establish baseline model performance per country without using data from the MOOCs' entire runtimes. Next, we determined cross-country model generalizability by applying each country's completion prediction model on data from every other country in the dataset. This was conducted to analyze how well models generalized across countries in the dataset and comparing the results to baseline model performances. We then computed cross-country model distances as a metric of cross-country model generalizability using the baseline and cross-country AUC ROC scores. The results of this analysis showed low median degradation of models when tested on other countries, suggesting that predictive models built on a single country's MOOC data will tend to generalize when tested on MOOC learners from other countries. This could be due to the well-documented selection bias in those who take MOOCs (Ferrer-Mico, 2016; Tovar et al., 2015), where the typical learner profile is that of a Western, educated (with at least a Bachelor's degree), and employed male, irrespective of native language.

Finally, distances were used to investigate the relationship between model generalizability and differences in various country-level metrics. In these analyses, we found that cross-country generalization of completion models generally performed on par with their baseline model performances, only degrading by half a percentage point on average. Results suggest that the degree to which models generalized across countries was significantly related to the differences in country-level measures of culture, happiness, wealth, and size.

Hofstede's cultural dimensions were found to relate significantly to both the performance and generalizability of the completion prediction models. The study found that more individualistic (IDV) or more long-term oriented (LTO) countries were more likely to have better-performing within-country (baseline) models. It is worth noting that both indices were significantly positively correlated to the country's GDP and enrollment size, suggesting that individualistic or long-term oriented countries were also likely to be wealthier and have a larger MOOC presence (i.e., larger training dataset).

Further, differences in cultural views relating to power distribution (PDI), indulgence (IVR), individualism, and long-term orientation were significantly related to model generalizability. In the case of IVR, for example, models trained on a more indulgent country (like Mexico or Sweden) will generalize better on a more restrictive country, but caution should be placed when generalizing models trained on a more restrictive country. Ultimately, the findings suggest that training models on countries scoring higher in the PDI (e.g., China, the Philippines) or IVR dimensions, or lower in the IDV (e.g., Guatemala, Panama) or LTO (e.g., Ghana, Nigeria) dimensions, were more likely to produce generalizable models. Countries that fit this profile, scoring high across all four dimensions include Venezuela, Mexico, Ghana, and Nigeria, all of which have mid-range enrollment (*mean*=12627) and GDP (*mean*=$5.1B). On the other end are countries like Estonia, Lithuania, Latvia, and Hungary, which have both low enrollment (*mean*=4262) and GDP (*mean*=$0.5B). These numbers are consistent with the significant positive correlations between differences in IDV and LTO and differences in GDP and per capita GDP, which imply that as either of these scores go down (and contribute to making a model more generalizable), the less wealthy the country is. Both groups of countries have similar average baseline model performances, *AUC ROC*=0.97.

Gross National Happiness, or self-reported nation-level happiness, as measured by the World Happiness Report (Helliwell et al., 2015), was also found to relate significantly to both model performance and generalizability. Interestingly, while happiness was found to have a positive effect on model performance within and cross-country, the difference in happiness between countries had an inverse relationship with model generalizability. That is, the happier a training country is compared to a testing country, the less generalizable the models. The relationship suggests that models produced using data from low-happiness countries were more likely to generalize compared to models produced using data from their happier neighbors.

Finally, measures of wealth and size were also found to relate significantly to both model performance and generalizability. GDP, per capita, population, and enrollment size were all significantly related to within-country model performance, suggesting that larger, wealthier countries with a larger MOOC presence were likely to produce better-performing models. This finding is intuitive—larger and wealthier countries are likely to have more learners enrolled in MOOCs (as evidenced by significant correlations between these measures), and a standard principle in machine learning states that having a larger training data set ensures better model performance. Likewise, differences in these features all had significant effects on model generalizability. The relationship with differences in size metrics—population and enrollment size—suggests that the larger the training country is compared to the testing country, the more generalizable the training country's model is. The findings related to differences in wealth, on the other hand, suggest that the wealthier the training country is compared to a testing country, in either GDP or per capita GDP, the less likely its model will generalize (i.e., higher positive difference in GDP or per capita suggests higher distance score).

However, despite the statistical significance of the effects of these country-level measure differences, they only explain a very small percentage of the variance in the cross-country distances. A likely explanation is that a number of other country-level factors are at play, ones not considered in this study. Perhaps Hofstede's (2010) cultural dimension framework is not sufficient in fully describing cultural differences across countries, or even within countries (as explained in the Limitations section below). Perhaps other access or socioeconomic differences not accounted for in this study are also contributing to the model distances.

Limitations

As noted above, the study was limited by the type of success metric investigated in the training and testing of predictive models. MOOC scholarship has evolved from investigating course completion as the sole metric of learner success—learners have been found to come into these courses with varied goals and motivations. An early paper by Kizilcec and colleagues (2013) found subpopulations of learners to emerge based on the way they interacted with the MOOC: some just watched lecture videos, some just interacted with the graded assessments, while others did a combination of both or neither, revealing a "plurality of [learner] trajectories" (p. 7). A study by Wang and Baker (2018) and Trumbore (2020) described other post-course success metrics, like publishing or joining a professional organization in the same field, or attaining various job-related benefits. Because such outcome variables are typically more difficult to gather—and at the scale MORF operates on, especially—our study was limited to investigating course completion, which is automatically tracked in all MOOC platforms, and continues to be the most researched and widely used metric of success in MOOCs.

Our study was also limited by the metrics used to quantify culture. A review by Baker and colleagues (2019) differentiates between macro- and micro-theories of culture. Macro-theories of culture attempt to “categorize all *groups* in the world according to some number of cultural dimensions” (p. 2). Hofstede’s cultural dimension framework falls into this category of cultural theories, in addition to other widely-cited frameworks: the Model of National Cultural Differences (Trompenaars & Hampden-Turner, 2011) and the nine dimensions presented in the GLOBE study (House, Hanges, Javidan, Dorfman, & Gupta, 2014). Micro-theories on culture, on the other hand, seek to contextualize culture down to the individual-level. In these theories, culture is “embedded in particular actors’ specific practices and activities that take place in particular contexts” (p. 6). They place an emphasis on a subject’s own cultural identity. However, because micro-theoretical approaches to culture are limited in their generalizability (Baker et al., 2019), and because this granularity of data would again be difficult to gather at the scale MORF operates on, our study was limited to macro-views of culture—specifically Hofstede’s cultural dimensions.

Conclusion and Next Steps

The findings in this study serve as a preliminary attempt to examine relationships and patterns across countries more closely and introduce several new and interesting questions that can aid in further investigating the cross-cultural generalizability of predictive models. Because this study sought to answer the *what*—where we found that differences in several country-level measures were linearly related with model generalizability. The wealth of data gathered and generated by this study allows for deeper investigations into *why* and *how* these country-level measures affect model generalizability.

In order to do this, I plan to investigate similarities or differences in predictive models themselves (i.e., feature importances on learner interaction) and how they differ between country pairs across each country-level measure using different difference thresholds (i.e., large positive difference, small positive difference, small negative difference, large negative difference in each of the country-level measures). For example, looking at how country-pair models are similar or different in their feature importances where the train country was substantially more individualistic than the test country, cases where the train country was only marginally more individualistic than test country, then cases where the train country was marginally more collectivist than the test country, then finally cases where the train country was substantially more collectivist than the test country. Being able to understand how different or similar these models are will better contextualize our findings and give us a clearer picture of why models tend to generalize in the directions reported in this study.

Second, I plan to conduct multidimensional scaling (MDS), which provides a visual representation of the distances among a set of objects—in the case of this study, cross-country model distances. This analysis will visualize how the combination of all these discovered relationships affects overall model generalizability (i.e., see which countries are closest to or farthest from each other given all these complex relationships) by assigning each country two dimension values (i.e., an x- and y-value), which can then be used to plot each country on a plane. Results from this analysis can then be used to form clusters among neighboring countries, which can be used to study how different or similar predictive models are across countries plotted closer to or farther from one other.

The methods used in this study provide a novel approach to examining cross-country prediction model generalization. Understanding what, why, and how factors lead to generalization of predictive models between countries will not only lead to better informed culturally-sensitive pedagogy for learners around the world, it will also lead to a new and deeper understanding of how culture influences learner-computer interaction. In the meantime, the implications from the findings of this paper are clear: researchers developing and studying predictive models in MOOCs need to start accounting for differences in learner nationality.

CHAPTER 5: CONCLUSION

Replication is a crucial step in the scientific process, as it enables researchers to better understand the reliability, validity, and merit of a study's findings. However, despite their importance, replication studies make up only 0.13% of published education research (Makel & Plucker, 2014). This is due to the fact that replication research is often faced with access issues when it comes to the original study's design, data, and methods. As such, Open Science practices have recently been getting increased attention across various fields of research. These practices aim to increase transparency and access to every facet of published research for the purposes of evaluation, reanalysis, and scrutiny.

Massive Open Online Courses (MOOCs) are known to draw in enrollment numbers in the tens of thousands per session (Jordan, 2014). In 2020, for example, more than 950 universities around the world offered their own selection of MOOCs, reaching over 180 million learners (Shah, 2019). Due to the massive enrollment numbers MOOCs continue to attract, these courses have become a source of rich and diverse data, which provides an opportunity to bridge the replication gap in online learning research. However, in order to protect the privacy of learner records, most of this data is subject to strict access regulations. As a result, researchers commonly have access to data from only the MOOCs they teach and are often barred from making the data publicly accessible for use by others. Over the years, education researchers have attempted to find a solution to this issue, developing different Open Science tools and platforms to overcome the costs and technical barriers linked to conducting replication studies (e.g., Pardos & Kao, 2015; Veeramachaneni et al., 2013), though none have achieved widespread use.

Through the studies in this dissertation, I developed, upgraded, and leveraged the MOOC Replication Framework (MORF), a platform that facilitates conducting and replicating MOOC research. The platform is both a big data repository of over 13 million users enrolled in over 100 MOOC sessions, and a tool for conducting end-to-end predictive modeling research. MORF was used to facilitate both replication studies (Studies 1 and 2) and a novel, fully replicable study on learner completion prediction models (Study 3).

MORF was first developed as a production system (Study 1) that allowed for conceptual replications, i.e., replications that validate a study's findings through the use of different methods. The previously published findings we tested were first transformed into if-then formulations, and then analyzed for replicability across MORF's database of MOOC sessions. Our feasibility study, which analyzed the replicability of 21 previously published findings in a MOOC on Big Data and Education, found that only nine findings (42.9%) replicated. The remaining production rules either failed to replicate (47.6%) or had a significant finding in the opposite direction of the original research (9.5%). Further, most of the production rules that failed to replicate were based on findings from studies that investigated the previous session of the *same course*. This lack of replication across such similar courses highlights the importance of conducting replication studies within MOOC research.

In order to both assess MORF's capacity to conduct large-scale replication research and investigate the replicability of these findings across a larger sample of data, a scaled-up version of the feasibility study was conducted (Study 2). This study analyzed the replicability of the previously published findings across the University of Edinburgh's then-entire Coursera MOOC line-up. Of the 15 findings investigated, 80%

replicated significantly in the new, larger dataset. Two findings (13.3%) had their opposite come out significant, while only one production rule (6.67%) failed to replicate. Overall, these findings suggest that there is considerable commonality in which behaviors are associated with success in MOOCs, even across courses on a heterogeneous range of topics.

In order to support more direct forms of replication, MORF was updated with a predictive modeling module. This module allows users to leverage an original study's methods in order to validate its findings. This version of MORF utilizes containerization technology, which affords users the ability to fully dictate how data extraction is conducted. These upgrades to MORF allow users to execute their own code at scale in the runtime environment necessary for the code to run. This predictive modeling module was used in a novel study, which sought to investigate how factors relating to a learner's country of origin or culture affected the cross-country generalizability of completion prediction models in MOOCs (Study 3). This study leveraged the previously published findings analyzed in the two previous studies in engineering the learner features that were used to train each country's completion prediction model. Models were trained to determine the best performing one per country. The performance of these models were treated as baseline model performances per country. Models were then tested against every other country in the dataset to establish distances between baseline and cross-country model performance. Finally, relationship mining was conducted between these distances and various country-level measures of culture, wealth, size, and happiness to assess which measures contributed to the cross-country generalizability of models. The study found that differences in these measures were linearly related to model generalizability. That is, how much higher a country scores in a given country-level measure (e.g., wealth or cultural views) compared to another country contributes to a model's generalizability.

The studies in this dissertation have proven MORF to be a useful tool in addressing technical and access difficulties surrounding the execution of replication research through its use of Open Data and Open Analysis practices. MORF facilitates novel and replication research by 1) giving users access to more massive amounts of rich and diverse learner data unlike what has been made accessible in the past, and 2) storing analysis or job submission artifacts (like source code and dockerfiles, which are used to build a user's runtime environment) in public repositories for purposes of replication and evaluation.

Through the studies presented in this dissertation, I hope to encourage future replication by providing an accessible and robust platform to people interested in conducting research in MOOCs, which can hopefully contribute to addressing the replication crisis, as well as open the door to answering a wide array of new and interesting questions in MOOC research.

Next Steps

The successful execution of this dissertation's third study is evidence that MORF's new infrastructure allows users to conduct both replication and novel research. Upgrades to MORF, which seek to extend its capabilities, are currently in development. In addition to giving users the ability to run their own feature extraction on the available data, this new version of MORF now gives users the ability to dictate the methodology involved in training and testing the predictive models as well. These modifications to the framework give users the freedom to design their own methods surrounding cross-validation, feature selection, and model performance output. It also allows the controlled output of additional information or visualization (in addition to model performance), as long as no personally identifiable information is shared, using a pre-selected but extensive set of output functions. These modifications allow for the facilitation of a range of new replication and novel research studies. For example, as was done in Study 3, data can now be aggregated based on grouping variables other than courses and sessions. MORF can be used to investigate the replicability of findings across subjects, content areas, or course designs. Such research can contribute to better understanding and supporting the needs of learners across these various categories.

While beta testing is ongoing, we are also working to achieve some high-priority milestones in our MORF roadmap. First, we are working on the ingestion of new MOOC datasets into MORF. The addition of over 130 sessions of Coursera data and over 110 sessions of edX data will increase MORF's already rich dataset by over 4.86 million learners. Because users have the freedom to program how features are extracted from the available raw data, users will simply need to know what each platform's data schema looks like in order to properly query MORF. To this end, we are currently drafting documentation and putting together sample datasets we can disseminate to help users extract the features they need across platforms. Because most MOOC providers have changed the way MOOCs are offered—charging steeper fees to those who are interested in completing—interactions with and within MOOCs have also likely changed compared to the way MOOCs were used when they were first offered. The ingestion of this new data will allow users to conduct research using more recent MOOC datasets, and thus provide more meaningful insight into learner success in these courses.

We are also working on MORF's interoperability with data gathered from other online learning platforms. This modification is the platform's first step in incorporating non-MOOC data, which will aid in making more e-learning data readily available and facilitating a wider array of research studies and use cases. Specifically, we are working on integrating ASSISTments (Heffernan & Heffernan, 2014) data into our data repository. ASSISTments is a web-based intelligent tutoring platform that implements a range of diverse student supports. The ingestion of external data will allow users to conduct cross-platform feature extraction in courses where quizzes or assignments are done in ASSISTments rather than natively in either Coursera or edX, which can then contribute to richer predictive models. Doing so also allows users the ability to leverage ASSISTments' capability of providing adaptive feedback on quizzes or assignments.

In addition to adding the data into our repository, we are also building a pipeline that can automatically create user mapping tables to track user IDs between ASSISTments and whichever MOOC platform it was used with. Our ultimate goal in this endeavor is to establish whether or not external partnerships like this are feasible. If they are, we want to eventually partner with more online learning platforms to both make their data more publicly accessible (and thus be more rigorously studied).

Finally, our roadmap also includes the addition of new modules beyond production rule mining and predictive modeling. We aim to support the execution of more descriptive analyses and the use of unsupervised learning algorithms, as well as more qualitative research methods, like those necessary in natural language processing and epistemic network analyses. These modifications will allow for an even wider range of MOOC research, such as the use of clustering and association rule mining analyses, or the use of linguistic features in the prediction of learner success—projects users had sought to conduct in MORF in the past but were unable to due to the framework's prior lack of support for such use cases.

Conclusion

The studies in this dissertation have proven the MOOC Replication Framework to be a useful tool in addressing the technical difficulties surrounding the execution of replication research through its use of Open Science practices. MORF facilitates novel and replication research by giving users access to more massive amounts of rich and diverse learner data unlike what has ever been made accessible before. Through the studies presented in this dissertation, I hope to encourage future replication by providing an accessible and robust platform to researchers interested in conducting research across large and representative MOOC datasets. This can contribute to addressing the replication crisis and open the door to answering a wide array of new and interesting questions in MOOC research.

APPENDIX

Appendix A

Table 17. Countries included in the Chapter 4 study and their enrollment sizes.

Country	Enrollment Size	Country	Enrollment Size
United Arab Emirates	9,849	Kuwait	1,452
Argentina	7,878	Lebanon	1,738
Austria	4,946	Lithuania	4,997
Australia	35,532	Luxembourg	1,084
Bangladesh	4,186	Latvia	3,072
Belgium	7,833	Morocco	4,710
Bulgaria	7,945	Mauritius	1,008
Brazil	60,892	Mexico	29,309
Canada	68,345	Malaysia	11,595
Switzerland	10,254	Nigeria	11,146
Chile	6,428	Netherlands	19,483
China	109,727	Norway	4,219
Colombia	18,520	New Zealand	5,651
Costa Rica	2,926	Panama	1,409
Czech Republic	7,283	Peru	10,381
Germany	37,713	Philippines	20,245
Denmark	7,001	Pakistan	15,650
Estonia	2,237	Poland	17,241
Egypt	20,368	Portugal	13,589
Spain	47,138	Romania	14,471
Ethiopia	1,305	Serbia	6,024
Finland	4,418	Russian Federation	55,165
France	28,793	Saudi Arabia	12,016
United Kingdom	70,260	Sudan	1,237
Ghana	5,199	Sweden	7,762
Greece	19,122	Singapore	27,600
Guatemala	2,540	Slovenia	2,908
Hong Kong	16,995	Slovak Republic	3,326
Croatia	4,942	El Salvador	1,349
Hungary	6,742	Thailand	11,068
Indonesia	10,083	Turkey	15,298
Ireland	8,200	Trinidad and Tobago	2,207
Israel	9,701	Taiwan	15,291
India	168,947	Tanzania	968
Iraq	749	Uganda	1,229
Iran	6,504	United States	635,531
Italy	21,550	Uruguay	1,340
Jordan	1,637	Venezuela RB	4,857
Japan	15,034	Vietnam	15,812
Kenya	3,417	South Africa	11,104
South Korea	12,196		

BIBLIOGRAPHY

Adamopoulos, P. (2013). What makes a great MOOC? An interdisciplinary analysis of student retention in online courses.

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).

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).

Almatrafi, O., Johri, A., & Rangwala, H. (2018). Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums. Computers & Education, 118, 1-9.

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2014). Engaging with massive online courses. In Proceedings of the 23rd international conference on World wide web (pp. 687-698). ACM.

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.

Andersson, U., Arvemo, T., & Gellerstedt, M. (2016). How well can completion of online courses be predicted using binary logistic regression?. In IRIS39-The 39th Information Systems Research Conference in Scandinavia, Ljungskile, Sweden, 7-10 August 2016.

Andres, J. M. L., Baker, R. S., Gašević, D., Siemens, G., Crossley, S. A., & Joksimović, S. (2018). Studying MOOC completion at scale using the MOOC replication framework. In Proceedings of the 8th International Conference on Learning Analytics and Knowledge (pp. 71-78).

Andres, J.M.L., Baker, R.S., Siemens, G., Gašević, D., & Spann, C.A. (2017). Replicating 21 Findings on Student Success in Online Learning. Technology, Instruction, Cognition, & Learning.

Arnett, J. (2008). The weirdest people in the world. American Psychologist, 63(7), 602-14.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In Learning analytics (pp. 61-75). Springer, New York, NY.

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.

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.

Baker, R.S., Ogan, A.E., Madaio, M., Walker, E. (2020). Culture in Computer-Based Learning Systems: Challenges and Opportunities. Computer-Based Learning in Context, 1(1), 1-13.

Belanger, Y., & Thornton, J. (2013). Bioelectricity: A Quantitative Approach to Duke University’s First MOOC.

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), 57(1), 289-300.

BNC Consortium. (2007), British National Corpus, version 3 (BNC XML ed.). Retrieved from www.natcorp.ox.ac.uk

Boettiger, C. (2015). An introduction to Docker for reproducible research. ACM SIGOPS Operating Systems Review, 49(1), 71-79.

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., Olds, J. L., & Dean, H. (2015). Social, behavioral, and economic sciences perspectives on robust and reliable science. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences, 3(4).

Bozkurt, A., Akgün-Özbek, E., & Zawacki-Richter, O. (2017). Trends and patterns in massive open online courses: Review and content analysis of research on MOOCs (2008-2015). International Review of Research in Open and Distributed Learning: IRRODL, 18(5), 118-147.

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.

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

Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In ACM SIGMOD Record (Vol. 26, No. 2, pp. 255-264). ACM.

Brinton, C. G., Buccapatnam, S., Chiang, M., & Poor, H. V. (2016). Mining MOOC clickstreams: Video-watching behavior vs. in-video quiz performance. IEEE Transactions on Signal Processing, 64(14), 3677-3692.

Brown, G. D. A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. Behavior Research Methods, Instrumentation & Computers, 16, 502–532. doi:10.3758/BF03200836

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41, 977–990. doi:10.3758/BRM.41.4.977

Buckheit, J. B., & Donoho, D. L. (1995). Wavelab and reproducible research. In Wavelets and statistics (pp. 55-81). Springer, New York, NY.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency (pp. 77-91).

Callen, T. (2020). Gross domestic product: An economy's all. International Monetary Fund. https://www.imf.org/external/pubs/ft/fandd/basics/gdp.htm

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research, 11, 2079-2107.

Chandrasekaran, M., Ragupathi, K., Kan, M. Y., & Tan, B. (2015). Towards feasible instructor intervention in MOOC discussion forums.

Chen, C., Sonnert, G., Sadler, P. M., & Malan, D. J. (2020). Computational thinking and assignment resubmission predict persistence in a computer science MOOC. Journal of Computer Assisted Learning.

Chen, G., Davis, D., Lin, J., Hauff, C., & Houben, G. J. (2016). Beyond the MOOC platform: gaining insights about learners from the social web. In Proceedings of the 8th ACM Conference on Web Science (pp. 15-24).

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

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

Collberg, C., Proebsting, T., Moraila, G., Shankaran, A., Shi, Z., & Warren, A. M. (2014). Measuring reproducibility in computer systems research. Department of Computer Science, University of Arizona, Tech. Rep, 37.

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.

Cooper, S., & Sahami, M. (2013). Reflections on stanford's moocs. Communications of the ACM, 56(2), 28-30.

Cormier, D. (2008). The CCK08 MOOC–Connectivism course, 1/4 way. Retrieved from http://davecormier.com/edblog/2008/10/02/the-cck08-mooc-connectivism-course-14-way/

Crossley, S. A., Kyle, K., and McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. Behavior Research Methods.

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.

De La Roca, M., Morales, M., Teixeira, A. M., Sagastume, F., Rizzardini, R. H., & Barchino, R. (2018). MOOCs as a disruptive innovation to develop digital competence teaching: A micromasters program edX experience. European Journal of Open, Distance and E-learning, 21(2).

DeBoer, J., Stump, G. S., Seaton, D., Ho, A., Pritchard, D. E., & Breslow, L. (2013). Bringing student backgrounds online: MOOC user demographics, site usage, and online learning. In Educational data mining 2013.

Dillahunt, T., Chen, B., & Teasley, S. (2014). Model thinking: demographics and performance of MOOC students unable to afford a formal education. In Proceedings of the first ACM conference on Learning@ scale conference (pp. 145-146).

Donoho, D. (2017). 50 years of data science. Journal of Computational and Graphical Statistics, 26(4), 745-766.

Dunn, O. J. (1961). Multiple comparisons among means. Journal of the American statistical association, 56(293), 52-64.

edX. (2017). edX Research Guide. Retrieved from the edX website: http://edx.readthedocs.io/projects/devdata/en/latest/

Eurostat. (2020). Beginners: Population. Eurostat Statistics. Retrieved from: https://ec.europa.eu/eurostat/statistics-explained/index.php/Beginners:Population#How_is_the_population_in_a_country_or_given_area_calculated.3F

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

Fecher, B., & Friesike, S. (2014). Open science: one term, five schools of thought. Opening science, 17-47.

Fei, M., & Yeung, D. Y. (2015). Temporal models for predicting student dropout in massive open online courses. In 2015 IEEE International Conference on Data Mining Workshop (ICDMW) (pp. 256-263). IEEE.

Ferrer-Mico, M. T. (2016). Community of Inquiry (COI) and Self-Directed Learning (SDL) in Online Environments: An Exploratory, Correlational and Critical Analysis of MOOCs. Introduction to Cybersecurity MOOC Case Study (Doctoral dissertation, Universitat Ramon Llull).

Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. Educational researcher, 31(8), 4-14.

Fini, A. (2009). The technological dimension of a massive open online course: The case of the CCK08 course tools. International Review of Research in Open and Distributed Learning, 10(5).

Finkle, T. A., & Masters, E. (2014). Do MOOCs Pose a Threat to Higher Education?. Research in Higher Education Journal, 26.

Fogarty, J., Baker, R. S., & Hudson, S. E. (2005). Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. In Proceedings of Graphics Interface 2005 (pp. 129-136).

Friedman-Hill, E. (2002). Jess, the expert system shell for the java platform.USA: Distributed Computing Systems.

Gardner, J., & Brooks, C. (2018). Dropout model evaluation in MOOCs. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

Gardner, J., Brooks, C., Andres, J. M., & Baker, R. S. (2018a). MORF: A framework for predictive modeling and replication at scale with Privacy-Restricted MOOC data. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 3235-3244). IEEE.

Gardner, J., Brooks, C., Andres, J. M., & Baker, R. S. (2018b). Replicating MOOC predictive models at scale. In Proceedings of the Fifth Annual ACM Conference on Learning at Scale (pp. 1-10).

Gardner, J., Brooks, C., Baker, R. (2019) Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. Proceedings of the 9th International Learning Analytics and Knowledge Conference, 225-234.

Gardner, J., Yang, Y., Baker, R. S., & Brooks, C. (2019). Modeling and Experimental Design for MOOC Dropout Prediction: A Replication Perspective. International Educational Data Mining Society.

Gardner, J., Yang, Y., Baker, R., & Brooks, C. (2018). Enabling End-To-End machine learning replicability: A case study in educational data mining. arXiv preprint arXiv:1806.05208.

Gehlbach, H., & Robinson, C. D. (2018). Mitigating illusory results through preregistration in education. Journal of Research on Educational Effectiveness, 11, 296–315. doi:10.1080/1934574 7.2017.1387950

Gibbs, A. L. (2014). Experiences teaching an introductory statistics MOOC. In Proceedings of the ninth international conference on teaching statistics (ICOTS9).

Glennerster, R., & Takavarasha, K. (2013). Running randomized evaluations: A practical guide. Princeton University Press.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean?. Science translational medicine, 8(341), 341ps12-341ps12.

Gundersen, O. E., & Kjensmo, S. (2018). State of the art: Reproducibility in artificial intelligence. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

Guo, P. J., & Reinecke, K. (2014). Demographic differences in how students navigate through MOOCs. In Proceedings of the first ACM conference on Learning@ scale conference (pp. 21-30).

Gütl, C., Rizzardini, R. H., Chang, V., & Morales, M. (2014). Attrition in MOOC: Lessons learned from drop-out students. In International workshop on learning technology for education in cloud (pp. 37-48). Springer, Cham.

Han, F. (2014). Modeling Problem Solving in Massive Open Online Courses (Doctoral dissertation, Massachusetts Institute of Technology).

Han, F., Veeramachaneni, K., & O'Reilly, U. M. (2013). Analyzing millions of submissions to help MOOC instructors understand problem solving. In NIPS Workshop on Data Driven Education (pp. 1-5).

Harvard-MITx. HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0. Harvard Dataverse (2014).

Haywood, J., & Macleod, H. (2014). To MOOC or not to MOOC? University decision-making and agile governance for educational innovation. In Massive Open Online Courses (pp. 56-70). Routledge.

Helliwell, J. F., Layard, R., & Sachs, J. (2012). World happiness report 2012. World Happiness Report 2015. New York: Sustainable Development Solutions Network. Retrieved from: https://worldhappiness.report/ed/2012/

Helliwell, J. F., Richard L., & Sachs, J. (2015). World Happiness Report 2015. New York: Sustainable Development Solutions Network. Retrieved from: https://worldhappiness.report/ed/2015/

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. Nature, 466(7302), 29. https://doi.org/10.1038/466029a

Hofstede, G. (1986). Cultural differences in teaching and learning. International Journal of intercultural relations, 10(3), 301-320.

Hofstede, G. (1998). Masculinity and femininity: The taboo dimension of national cultures (Vol. 3). Sage Publications.

Hofstede, G. (2005) Cultures and organizations: software of the mind. ISBN 0-07-143959-5

Hofstede, G. (2011). Dimensionalizing cultures: The Hofstede model in context. Online readings in psychology and culture, 2(1), 2307-0919.

Hofstede, G., & Minkov, M. (2010). Long-versus short-term orientation: new perspectives. Asia Pacific business review, 16(4), 493-504.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). Cultures and Organizations, Software of the mind. Intercultural Cooperation and Its Importance for survival.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (Eds.). (2004). Culture, leadership, and organizations: The GLOBE study of 62 societies. Sage publications.

Hunter, J. (2012). Post-publication peer review: opening up scientific conversation. Frontiers in computational neuroscience, 6, 63.

Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., ... & Puranen, B. (2014). World values survey: Round six-country-pooled datafile version. Madrid: JD Systems Institute, 12.

Jaschik, S. (2013). MOOCs for credit. Inside Higher Ed, 23.

Johnson, C. (2020). Gender, emotion and political discourse: masculinity, femininity and populism. In The Rhetoric of Political Leadership. Edward Elgar Publishing.

Jordan, K. (2014). Initial trends in enrolment and completion of massive open online courses. International Review of Research in Open and Distributed Learning, 15(1), 133-160.

Jung, E., Kim, D., Yoon, M., Park, S., & Oakley, B. (2019). The influence of instructional design on learner control, sense of achievement, and perceived effectiveness in a supersize MOOC course. Computers & Education, 128, 377-388.

Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, K. Z., & Miller, R. C. (2014). Understanding in-video dropouts and interaction peaks in online lecture videos.

In Proceedings of the first ACM conference on Learning@ scale conference (pp. 31-40). ACM.

Kitzes, J., Turek, D., & Deniz, F. (2017). The practice of reproducible research: case studies and lessons from the data-intensive sciences. Univ of California Press.

Kizilcec, R. F. & Cohen, G. L. (2017). Eight-minute self-regulation intervention improves educational attainment at scale in individualist but not collectivist cultures. Proceedings of the National Academy of Sciences (PNAS), 114(17), 4348–4353.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In Proceedings of the third international conference on learning analytics and knowledge (pp. 170-179).

Koedinger, K. R., & Corbett, A. (2006). Cognitive tutors. The Cambridge handbook of the learning sciences, 61-77.

Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. Handbook of educational data mining, 43, 43-56.

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.

Kostas, M. (2021). Discursive construction of hegemonic masculinity and emphasised femininity in the textbooks of primary education: children's discursive agency and polysemy of the narratives. Gender and Education, 33(1), 50-67.

Kovacs, G. (2016). Effects of in-video quizzes on MOOC lecture viewing. In Proceedings of the third (2016) ACM conference on Learning@ Scale (pp. 31-40).

Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., & Siemens, G. (2016). Towards automated content analysis of discussion transcripts: A cognitive presence case. In Proceedings of the sixth international conference on learning analytics & knowledge (pp. 15-24).)

Kucera, H., & Francis, W. N. (1967). Computational analysis of present-day American English. Providence, RI: Brown University Press.

Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. TESOL Quarterly, 49 (4), 757-786.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. Artificial intelligence, 33(1), 1-64.

Lan, X., Ma, C., & Radin, R. (2019). Parental autonomy support and psychological well-being in Tibetan and Han emerging adults: A serial multiple mediation model. Frontiers in psychology, 10, 621.

Li, X., Song, D., Han, M., Zhang, Y., & Kizilcec, R. F. (2021). On the limits of algorithmic prediction across the globe. arXiv preprint arXiv:2103.15212.

Liu, M., McKelroy, E., Kang, J., Harron, J., & Liu, S. (2016). Examining the use of Facebook and Twitter as an additional social space in a MOOC. American Journal of Distance Education, 30(1), 14-26.

Liu, Z., Brown, R., Lynch, C., Barnes, T., Baker, R., Bergner, Y., McNamara, D. (2016) MOOC Learner Behaviors by Country and Culture: an Exploratory Analysis. Proceedings of the 9th International Conference on Educational Data Mining, 127-134.

Lowenthal, P., Snelson, C., & Perkins, R. (2018). Teaching massive, open, online, courses (MOOCs): Tales from the front line. International Review of Research in Open and Distributed Learning, 19(3).

Łukasz, K., Sharma, K., Shirvani Boroujeni, M., & Dillenbourg, P. (2016). On generalizability of MOOC models. In Proceedings of the 9th International Conference on Educational Data Mining (No. EPFL-CONF-223613, pp. 406-411).

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. Educational Researcher, 43(6), 304-316.

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.

Margaryan, A., Bianco, M., & Littlejohn, A. (2015). Instructional quality of massive open online courses (MOOCs). Computers & Education, 80, 77-83.

McAuley, A., Stewart, B., Siemens, G., & Cormier, D. (2010). The MOOC model for digital practice.

McKiernan, G. (2000). arXiv. org: the Los Alamos National Laboratory e-print server. International Journal on Grey Literature.

Milligan, C., Littlejohn, A., & Margaryan, A. (2013). Patterns of engagement in connectivist MOOCs. Journal of Online Learning and Teaching, 9(2), 149.

Moore, R. L., & Wang, C. (2021). Influence of learner motivational dispositions on MOOC completion. Journal of Computing in Higher Education, 33(1), 121-134.

Morrison, D. (2013). How NOT to design a MOOC: The disaster at Coursera and how to fix it [Blog post]. Online Learning Insights. Retrieved from http://

onlinelearninginsights.wordpress.com/2013/02/01/how-not-to-design-a-mooc-the-disasterat-coursera-and-how-to-fix-it/

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E., Ware., J. J., & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature human behaviour*, *1*(1), 1-9.

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. Psychological science, 31(6), 678-701.

Ng, A., & Widom, J. (2014). Origins of the modern MOOC (xMOOC). Hrsg. Fiona M. Hollands, Devayani Tirthali: MOOCs: Expectations and Reality: Full Report, 34-47.

Ogan, A., Walker, E., Baker, R., Rodrigo, M.M.T., Soriano, J.C., Castro, M.J. (2015) Towards Understanding How to Assess Help-Seeking Behavior Across Cultures. International Journal of Artificial Intelligence in Education, 25 (2), 229-248.

Olorisade, B. K., Brereton, P., & Andras, P. (2017). Reproducibility in machine Learning-Based studies: An example of text mining.

Onah, D. F., Sinclair, J. E., & Boyatt, R. (2014). Exploring the use of MOOC discussion forums. In Proceedings of London International Conference on Education (pp. 1-4).

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.

Orsini-Jones, M., & Carrascosa, A. C. (2019). BMELTET--Blending MOOCs into English Language Teacher Education with Telecollaboration. New educational landscapes: innovative perspectives in language learning and technology, 47.

Ospina-Delgado, J. E., Zorio-Grima, A., & García-Benau, M. A. (2016). Massive open online courses in higher education: A data analysis of the MOOC supply. Intangible Capital, 12(5), 1401-1450.

Page, M. J., Altman, D. G., Shamseer, L., McKenzie, J. E., Ahmadzai, N., Wolfe, D., Yazdi, F., Catala-Lopez, F., Tricco, A. C., & Moher, D. (2018). Reproducible research practices are underused in systematic reviews of biomedical interventions. Journal of clinical epidemiology, 94, 8-18.

Pardos, Z. A., & Kao, K. (2015). moocRP: An open-source analytics platform. In Proceedings of the Second (2015) ACM conference on learning@ scale (pp. 103-110).

Patil, P., Peng, R. D., & Leek, J. T. (2016). A statistical definition for reproducibility and replicability. BioRxiv, 066803.

Pereira, F. D. (2021). MOOC Next Week Dropout Prediction: Weekly Assessing Time and Learning Patterns. In Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings (p. 119). Springer Nature.

Perneger, T.V. (1998). What's wrong with Bonferroni adjustments. British Medical Journal, 316, 1236-1238.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). Tuned models of peer assessment in MOOCs. arXiv preprint arXiv:1307.2579.

Pournaras, E. (2017). Cross-disciplinary higher education of data science–beyond the computer science student. Data Science, 1(1-2), 101-117.

Rai, L., & Chunrao, D. (2016). Influencing factors of success and failure in MOOC and general analysis of learner behavior. International Journal of Information and Education Technology, 6(4), 262.

Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H. & Getoor, L. (2013). Modeling learner engagement in MOOCs using probabilistic soft logic. In NIPS Workshop on Data Driven Education (Vol. 2, pp. 1-7).

Reich, J., & Ruipérez-Valiente, J. A. (2019). The MOOC pivot. Science, 363(6423), 130-131.

Rivard, R. (2013). Measuring the MOOC dropout rate. Inside Higher Ed, 8, 2013.

Rodriguez, O. (2013). The concept of openness behind c and x-MOOCs (Massive Open Online Courses). Open Praxis, 5(1), 67-73.

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. Personality and Social Psychology Review, 5(1), 2-14.

San Pedro, M.O.C., Baker, R. & Rodrigo, M.M. (2011). Detecting carelessness through contextual estimation of slip probabilities among students using an intelligent tutor for mathematics. Proceedings of 15th International Conference on Artificial Intelligence in Education, 304-311.

Sandeen, C. (2013). Integrating MOOCs into traditional higher education: The emerging “MOOC 3.0” era. Change: The magazine of higher learning, 45(6), 34-39.

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.

Scandura, J.M. (2014). Adaptive Learning: How It is Learned or What Is Learned?. Tecnology., Instruction, Cognition, and Learning (TICL), 9, 237–239.

Schmidt, F. L., & Hunter, J. E. (2014). Methods of meta-analysis: Correcting error and bias in research findings. Sage publications.

Schroeder, R. (2012). Emerging open online distance education environment. Continuing higher education review, 76, 90-99.

Shah, D. (2019). Online degrees slowdown: A review of MOOC stats and trends in 2019. Class Central. Retrieved on 6 Aug 2020, at https://www.edsurge.com/news/2019-12-18-online-degrees-slowdown-a-review-of-mooc-stats-and-trends-in-2019

Sharif, A., & Magrill, B. (2015). Discussion forums in MOOCs. International Journal of Learning, Teaching and Educational Research, 12(1).

Siemens, G. (2005). Connectivism: A learning theory for the digital age. International Journal of Instructional Technology and Distance Learning. 2(1). Retrieved from http://www.itdl.org/Journal/Jan_05/article01.htm

Sinha, T., Jermann, P., Li, N., & Dillenbourg, P. (2014). Your click decides your fate: Inferring information processing and attrition behavior from MOOC video clickstream interactions. arXiv preprint arXiv:1407.7131.

Soares, A. M., Farhangmehr, M., & Shoham, A. (2007). Hofstede's dimensions of culture in international marketing studies. Journal of business research, 60(3), 277-284.

Søndergaard, M. (1994). Research note: Hofstede's consequences: a study of reviews, citations and replications. Organization studies, 15(3), 447-456.

Spellman, B. A. (2012). Introduction to the special section: Data, data, everywhere... especially in my file drawer. Perspectives on Psychological Science, 7(1), 58.

Steenkamp, J. B. E. (2001). The role of national culture in international marketing research. International Marketing Review.

Stewart, B. (2013). Massiveness+openness=new literacies of participation. Journal of Online Learning and Teaching, 9(2), 228-238.

Stodden, V., & Miguez, S. (2013). Best practices for computational science: Software infrastructure and environments for reproducible and extensible research. Available at SSRN 2322276.

Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A. & Williams, R.M. Jr. (1949). The American Soldier, Vol. 1: Adjustment during Army Life. Princeton University Press, Princeton.

Sun, D., Mao, Y., Du, J., Xu, P., Zheng, Q., & Sun, H. (2019). Deep learning for dropout prediction in MOOCs. In 2019 Eighth International Conference on Educational Innovation through Technology (EITT) (pp. 87-90). IEEE.

Svartvik, J., & Quirk, R. (1980). A corpus of English conversation. Lund, Sweden: Gleerup.

Thorndike, E. L., & Lorge, I. (1944). The teacher's word book of 30,000 words. New York, NY: Teachers College, Columbia University.

Tovar, E., Cabedo, R., Kalz, M., Walhout, J., Kreijns, K., & Niellisen, G. (2015). Who is taking European MOOCs and why? A large-scale, cross provider data collection about participants of European Open Online Courses. *Proceedings of Open Education Global*.

Trompenaars, F., & Hampden-Turner, C. (2011). Riding the waves of culture: Understanding diversity in global business. Nicholas Brealey International.

Trumbore, A. M. (2020). Learner Behavior and Career Benefits in Massive Open Online Courses (Doctoral dissertation, University of Pennsylvania).

Tsironis, A., Katsanos, C., & Xenos, M. (2016). Comparative usability evaluation of three popular MOOC platforms. In 2016 IEEE Global Engineering Education Conference (EDUCON) (pp. 608-612). IEEE.

van der Zee, T., & Reich, J. (2018). Open Education Science. AERA Open, 4(3).

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC bioinformatics, 7(1), 91.

Veeramachaneni, K., Dernoncourt, F., Taylor, C., Pardos, Z., & O'Reilly, U. M. (2013). Moocdb: Developing data standards for mooc data science. InAIED 2013 Workshops Proceedings Volume (p. 17).

Veeramachaneni, K., Dernoncourt, F., Taylor, C., Pardos, Z., & O'Reilly, U. M. (2013). Moocdb: Developing data standards for mooc data science. In AIED 2013 workshops proceedings volume (Vol. 17).

Wang, X., Yang, D., Wen, M., Koedinger, K., & Rosé, C. P. (2015). Investigating How Student's Cognitive Behavior in MOOC Discussion Forums Affect Learning Gains. International Educational Data Mining Society.

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).

Wang, Y., & Baker, R. (2015). Content or platform: Why do students complete MOOCs. MERLOT Journal of Online Learning and Teaching, 11(1), 17-30.

Wang, Y., & Baker, R. (2018). Grit and intention: Why do learners complete MOOCs?. The International Review of Research in Open and Distributed Learning, 19(3).

Wen, M., Yang, D., & Rose, C. (2014). Sentiment Analysis in MOOC Discussion Forums: What does it tell us?. In Educational data mining 2014.

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.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. American psychologist, 61(7), 726.

Wilkowski, J., Deutsch, A., & Russell, D. M. (2014). Student skill and goal achievement in the mapping with google MOOC. In Proceedings of the first ACM conference on Learning@ scale conference (pp. 3-10). ACM.

Wolins, L. (1962). Responsibility for raw data. American Psychologist, 17, 657–658.

Wu, W. H., Kao, H. Y., Wu, S. H., & Wei, C. W. (2019). Development and Evaluation of Affective Domain Using Student's Feedback in Entrepreneurial MOOC Courses. Frontiers in psychology, 10, 1109.

Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. Computers in human behavior, 58, 119-129.

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)

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.

Yuan, L., & Powell, S. (2013). MOOCs and disruptive innovation: Implications for higher education. eLearning Papers, In-depth, 33(2), 1-7.

Zhang, C., Huang, Y., Wang, J., Lu, D., Fang, W., Stamper, J., Fancsali, S., Holstein, K., & Aleven, V. (2019). Early Detection of Wheel Spinning: Comparison across Tutors, Models, Features, and Operationalizations. International Educational Data Mining Society.

Zhang, Q., Peck, K. L., Hristova, A., Jablokow, K. W., Hoffman, V., Park, E., & Bayeck, R. Y. (2016). Exploring the communication preferences of MOOC learners and the value of preference-based groups: Is grouping enough?. Educational Technology Research and Development, 64(4), 809-837.

Zhenghao, C., Alcorn, B., Christensen, G., Eriksson, N., Koller, D., & Emanuel, E. (2015). Who’s Benefiting from MOOCs, and Why. Harvard Business Review.