The Dissertation Committee for Jason Don Ferrell Certifies that this is the approved version of the following dissertation:

The Effectiveness of Synchronous Massive Online Courses at The University of Texas at Austin

Committee:

James W. Pennebaker, Supervisor

Samuel D. Gosling

Michael P. Domjan

Tal Yarkoni

The Effectiveness of Synchronous Massive Online Courses at The University of Texas at Austin

by

Jason Don Ferrell

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2017

Dedication

This dissertation is dedicated to the memory of Jennifer Malin, who played an important role in the early implementation of the SMOC and believed the system would help bring high-quality education to disadvantaged people who could not attend traditional university.

Acknowledgements

I thank James Pennebaker (chairperson), Samuel Gosling, Michael Domjan, and Tal Yarkoni for serving on this dissertation committee and for providing feedback on this project. Dr. Pennebaker, thank you for being an inspirational role model as an academic, teaching me the importance of the big picture, and for not giving up on me. Dr. Gosling, thank you for teaching me the nuance of masterful teaching, and for being "critical, skeptical, not easily impressed" in the best Blockian way. Dr. Domjan, thank you for teaching me the role of learning principles in daily life, and for being a gentle guide in rough seas. Dr. Yarkoni, thank you for teaching me what rigorous science is and how to do it.

I thank Elliot Tucker-Drob for helping with earlier versions of this dissertation, and Joshua Blank for helping extract data for this project. I thank Kimberly Terry for the continued support through my graduate program. I thank Pennebaker and Gosling lab members—especially Lindsay Graham, Ryan Boyd, Cindy Chung, Gabriella Harari, Rebecca North, Carson Sandy, Yitai Seih, and Kate Blackburn—for several years of educational bliss.

I thank my family for supporting me during my adventure in education (a special thanks to Garrett, Crina, Dad, my late Mom, and Amber). I would not have made it without Amber Ferrell-Reeves staying by my side. Crina Silasi-Mansat, thank you for sharing your life with me, and for being EMsF. I thank local Austin musicians for

providing soul medicine in the form of life-changing music, especially Phil Hurley, Josh Zee, Teal Collins Zee, Barbara Nesbitt, Lonnie Trevino Jr., Phil Bass, and Chris Beall. What an amazing ride on the waves of the Longhorn seas.

The Effectiveness of Synchronous Massive Online Courses at The University of Texas at Austin

Jason Don Ferrell, Ph.D.

The University of Texas at Austin, 2017

Supervisor: James W. Pennebaker

Is online education an effective and viable alternative to face-to-face education? The purpose of this dissertation was to evaluate the effectiveness of online education at The University of Texas at Austin (UT-Austin). The dissertation focused on Synchronous Massive Online Courses (SMOCs) at The University of Texas at Austin since 2012. This dissertation analyzed the extent to which course effectiveness varies as a function of lecture environment, comparing SMOCs to similar face-to-face (FTF) courses.

In total, 25,726 students across 53 courses at UT-Austin were included in analyses. Researchers compiled all relevant student and course data archived in university databases and merged that with course data compiled from archived course syllabi. Then, Hierarchical Linear Modeling was used to test how (a) final course grades vary as a function of lecture environment (SMOC or FTF), controlling for socioeconomic status, scholastic aptitude, and course exam frequency, (b) subsequent semester grades vary as a function of lecture environment (SMOC or FTF), controlling for socioeconomic status, scholastic aptitude, and course exam frequency, and (c) course completion rates vary as a function of lecture environment (SMOC or FTF), controlling for socioeconomic status, scholastic aptitude, and course exam frequency.

The primary goal of this project was to examine the effectiveness of SMOCs in comparison to FTFs. Course effectiveness was operationally defined with three objective outcomes: final course grades, subsequent semester GPAs, and course completions. Findings show that there were no significant differences between SMOCs and FTFs on any of these objective measures. That is, SMOCs neither outperform nor underperform FTFs in final grades, subsequent semester GPAs, or course completions.

Because previous studies propose that increasing exam frequency may reduce SES-based achievement gaps (e.g., Pennebaker, Gosling, & Ferrell, 2013), and there are some mixed results in the literature about the effectiveness of frequent testing (e.g., Bell, Simone, & Whitfield, 2015), a secondary goal of this dissertation focused on the interaction of SES and exam frequency in the context of course effectiveness outcomes. Exam frequency interacted with lecture environment; such that for FTFs, there was no substantial difference in final course grades by exam frequency; however, for SMOCs, students with more exams had higher final course grades than students with fewer exams. The highest final grades were earned by students in SMOCs that provided the highest exam frequencies (while accounting for control variables). Exam frequency also interacted with socioeconomic status (SES); such that for lower SES students, when exam frequencies are lower the probabilities of course completion are lower than when exam frequencies are higher; and when exam frequencies are higher, the probabilities of course completion are higher than when exam frequencies are lower. For higher SES students, the probabilities of course completion did not vary by exam frequency. Given these findings, increasing exam frequencies in course structures is recommended.

Looking across a wide range of course topics and courses, and large number of students, this dissertation provides evidence that SMOCs are as effective as FTFs on objective course outcomes, both short- and long-term. This includes final course grades, subsequent semester GPAs, and course completion rates as course effectiveness measures. Economically, SMOCs are able to reach thousands of students by relying on fewer faculty without the need for large classrooms. At the same time, it frees faculty to teach more and smaller upper division courses. Although the results of the SMOC and FTF courses were generally similar, the additional payoffs of the SMOCs make them a promising tool for the future of undergraduate education. If the high standard of educational course effectiveness is based in the traditional FTF course, then a comparable SMOC course meets that high standard.

Table of Contents

List of Tables

List of Figures

List of Illustrations

The Effectiveness of Synchronous Massive Online Courses

at The University of Texas at Austin

Online education touches millions of undergraduate students each year. Across the United States in 2014, out of approximately 20 million total undergraduate students, around 32% took at least one online course, 8% took exclusively online courses, and 7% were enrolled in completely online degree programs (National Center for Education Statistics, 2014). Locally, in the Fall 2017 semester, approximately 16% of students at UT-Austin will take online courses.

Given the sheer numbers of students and instructors working through the education process in an online environment, any effectiveness—or ineffectiveness—of the process and resulting implications are important to understand. For example, online environments may affect students' learning in ways not observed in FTF classrooms. Additionally, online environments may constrain instructors' teaching methods and materials in ways that affect the learning process. Online education environments may change student-to-student and student-to-instructor interactions. Scientists are only beginning to focus on how learning, teaching, and social interaction processes in online environments are different from those in traditional FTF environments.

Also, many universities and other organizations are opening up various online courses to anyone in the world with internet access. These open-education systems afford an education opportunity that, previously, many people could not access (e.g., low socio-economic-status people; working adults; long distance learners). Because online courses are reaching and affecting undergraduate and graduate students, instructors, administrators, institutions, and society as whole, a better understanding of online versus FTF education effectiveness is warranted.

History of Online Education

The first effective online education system was developed specifically for internet use in 1994, the Computer Assisted Learning Campus (CALC; Morabito, 2015). CALC created relatively small online courses because the internet was an emerging technology that afforded a space for education outside of a centralized classroom. Currently, CALC is a private business that continues to offer high-school and college online courses.

After CALC, other important developments followed. For example, Jones International University received national accreditation with online courses in 1999, which was an important step in legitimizing online education as an alternative to FTF education (Hickey, 2014). Massachusetts Institute of Technology (MIT) began offering most of their online courses for free to anyone with internet access in 2002 (Dumbald, 2014). Later, Khan Academy, which was created in 2006, produced high-quality video mini-lectures (i.e., the majority were less than 10 minutes), mostly on mathematics topics, which were freely available on YouTube (Noer, 2012). Khan Academy currently publishes videos, courses, and assessment materials for about 31 million students in a wide variety of subject materials (Kashyap, 2016), and creates online education material at primary, secondary, and higher education levels. In less than a decade, online education evolved from a non-accredited system available only to people privileged with reliable internet and enough wealth to afford it from a private institution to a fully accredited and free system available to anyone in the world with internet access.

The first Massive Open Online Course (MOOC) was created in 2007 by ALISON (Dumbald, 2014). ALISON, which is based in Ireland, is currently home to about 7 million students (ALISON, 2016), and is the largest MOOC provider outside of the United States with many students in Africa, India, and Indonesia (Mesquita & Peres, 2015). Currently, some of the largest MOOC providers are Coursera, Udacity, and edX. Coursera was started at Stanford in 2011 and gathered much attention when over 160,000 students enrolled in their *Introduction to Artificial Intelligence* MOOC (Waldrop, 2013). Coursera is now the most popular MOOC provider in the world, offers over 840 courses from over 140 universities, and has over 10 million students enrolled. Udacity also evolved from Stanford's *Introduction to Artificial Intelligence* MOOC, but went in another direction by focusing on technology-based courses. In 2013, Udacity (in collaboration with San Jose State University) offered MOOCs for college credit, and then a completely MOOC-based master's degree in collaboration with AT&T and Georgia Institute of Technology (Onink, 2013).

MOOCs quickly became a popular method of education delivery, with The New York Times claiming 2012 to be "The year of the MOOC" (Pappano, 2012). In the same year, MIT and Harvard collaborated in the creation of edX (MIT Office of Digital Learning, 2016). One goal in the creation of edX was to provide an *open-access* application that other universities could use to host their own online courses. Pioneers in online education quickly learned that it is technologically and economically challenging to develop a reliable online education application for large courses. In this direction, edX helped many universities in their pursuits to create small and large online courses because the edX application was freely open to use. edX is the only large-scale MOOC application/provider that is both open-source and nonprofit. Currently over 300 courses are available from over 90 universities using edX, including UT-Austin (edX, 2016).

Development of the Synchronous Massive Online Course

The University of Texas at Austin (UT-Austin) has also been a pioneer in online education. The first Synchronous Massive Online Course (SMOC) was an *Introduction to Psychology* course streamed from UT-Austin to more than 800 students in 2012 (Straumsheim, 2013). This SMOC was co-taught and co-developed by James W. Pennebaker and Samuel D. Gosling, and the SMOC system was co-developed and managed by Jason D. Ferrell. The *Introduction to Psychology* SMOC used a proprietary application developed specifically to (a) stream high-quality, real-time lectures to large numbers of students, (b) facilitate high levels of interaction between students and online activities, fellow classmates, undergraduate student mentors, graduate student teaching assistants, and instructors and professors, and (c) integrate a unique and complex daily exam system designed to maximize learning. Data from these *Introduction to Psychology* SMOCs reveal that the system not only reduces the traditional achievement gap between lower and higher socio-economic status students, but also—in comparison to students that did not experience the SMOC—improves student performance in other courses during the current and subsequent semesters (Pennebaker, Gosling, & Ferrell, 2013).

The UT-Austin SMOC application has now been transferred from a standalone proprietary application to the Canvas learning management system (see https://www.canvaslms.com/). UT-Austin developed individual proprietary applications (e.g., streaming and interactive video, personalized activities, communication systems, unique and complex daily exam system, etcetera) that operate seamlessly within the Canvas applications system to fulfill customizable needs for various courses. UT-Austin currently offers SMOCs in Psychology, Anthropology, Chemistry, Government, Economics, English, History, Classical Civilization, and Art History—each using various pedagogical methods and applications. Additionally, UT-Austin currently offers asynchronous, on-demand, and self-paced online courses; with some using Canvas applications and others using edX applications.

Effectiveness of Online Education

Online Versus FTF Course Effectiveness

Before online courses were created, researchers compared the effectiveness of various methods of distance education (DE; i.e., interacting from a remote location) to traditional classroom education. A meta-analysis of studies published from 1940-1990 showed no difference in effectiveness between DE and traditional classroom methods, as measured by midterm and final grades (Machtmes & Asher, 2000). Additionally, a meta-analysis of over 200 studies published from 1992-2001 showed that web-based-learning courses (not necessarily online) entail not only more favorable student attitudes towards them, but also result in similar, and sometimes better, knowledge gains than traditional courses do (Chumley-Jones, Dobbie, & Alford, 2002).

Similarly, Bernard et al. (2014) examined the results of over 200 studies comparing the effectiveness of DE and FTF courses. DE course outcomes were similar to FTF course outcomes when looking at course-level and standardized testing, final grades, course evaluations, and course retention. However, Bernard and colleagues found that communication methods and prevalence are more important than the learning media in predicting course effectiveness. Overall, there is consensus in the literature that most methods of DE are as effective as classroom education (Bernard et al., 2014; Chumley-Jones, Dobbie, & Alford, 2002; Machtmes & Asher, 2000).

With the growing popularity of online courses, a need for research focusing on their effectiveness, especially in comparison with the FTF courses, has emerged. Similar to DE, there is consensus in the literature that most online courses are as or more effective than FTF courses. For example, participants in a fully interactive e-learning group performed significantly better when tested on the course material than participants in an FTF group (Zhang, 2005). Additionally, physicians receiving online training regarding chronic pain management showed a pre- to post-test increase in effectiveness whereas physicians receiving FTF training did not (Harris, Elliot, Davis, Chabal, Fulginiti, & Fine, 2007).

Furthermore, MIT compared both versions of their *Physics Mechanics* course (MOOC and FTF) finding that MOOC students performed slightly better on exams than the FTF students (Rayyan et al., 2016). Hugenholtz, de Croon, Smits, van Dijk, and Nieuwenhuijsen (2008) showed that online courses did not differ from FTF courses in pre- to post-testing about mental health issues. Online training for the classification of ulcers led to better test performance for nursing students when compared to FTF training; however, the effect was not found for certified nurses (Beekman, Schoonhoven, Boucque, Van Maele, & Defloor, 2007). Lastly, a meta-analysis of 51 studies comparing online to FTF courses showed that online courses are as or more effective as FTF courses when comparing learning outcomes (Means, Toyama, Murphy, Bakia, & Jones, 2009).

Overall, results showing that there are not significant differences in distance and online courses from FTF courses, regarding effectiveness, are plentiful (Bernard et al., 2004; Chumley-Jones et al., 2002; Fortune, Shifflett, & Sibley, 2006; Herman & Banister, 2007; Koory, 2003; Machtmes & Asher, 2000; Means et al., 2009; Tallent-Runnels et al., 2006; Warren & Holloman, 2005; Weber & Lennon, 2007). However, surprisingly, opinions are still mixed on the issue, with almost 25% of academic leaders arguing that the online courses are inferior to FTF courses in terms of learning outcomes (Allen & Seaman, 2013).

Predictors of Online Course Effectiveness

What factors play a role in the effectiveness of online education? Various factors have been proposed. For example, data from 44,000 students in 120 universities showed that institution ranking and student engagement influence learning in online courses (Hu & Kuh, 2003). Students from higher ranked universities or with higher levels of engagement are more successful in online courses than students from lower ranked universities or with lower levels of engagement. Similarly, top contributing students to five social tools in an online course (i.e., questioning and answering, forums, Facebook, Twitter, and MentorMob) had higher final grades than low contributing students (Alario-Hoyos, Munoz-Merino, Perez-Sanagustin, Delgado Kloos, & Parada, 2016). A large-scale study of SMOCs found that randomly assigning students to small groups (i.e., 2-6 students) for short concept-relevant discussions is an effective social engagement activity; however, it seems that engagement—of any kind—is the mechanism for improved learned, and not necessarily the small group activities (Boyd, Pennebaker, Ferrell, & Georgiev, 2015).

Engagement, along with intelligence and conscientiousness, stand as “pillars” of academic success in meta-analyses of various individual differences that predict academic success (von Stumm, Hell, & Chamorro-Premuzic, 2011). In general, the more students engage with a course, the better they perform in the course. Extending these findings to online SMOC courses, engagement predicts final grades even after controlling for intelligence, conscientiousness, negative affect, socio-economic status, sociability, and thinking style (Ferrell, Tucker-Drob, Yarkoni, Gosling, & Pennebaker, in preparation).

Participation and engagement seem to have a reliable influence on the learning outcomes of online courses. De Barba, Kennedy, and Ainley (2016) also included the concepts of motivation (internally generated interest) and situational interest (externally generated interest), along with participation, as possible factors influencing learning outcomes in online courses. Results showed not only that motivation and participation predicted learning, but also that situational interest acts as a mediator for both motivation predicting learning and for participation predicting learning.

Completion rate (i.e., percent of students completing a course) is a commonly used course effectiveness measure. Completion rate is a relatively challenging metric in the context of MOOCs because there are virtually no consequences for dropping the course at any time. The average completion rate of MOOCs is 10-20% (North, Richardson, & North, 2014). For example, the completion rate of 12 Coursera MOOCs offered between 2013 and 2014 was approximately 11% (de la Garza, Sancho-Vinuesa, & Zermeno, 2015). While substantially accentuated in online courses, completion rate is a challenge faced by FTF courses as well (Adamopoulos, 2013).

Frequent Testing

While frequent testing is common advice for improving learning and performance, research concerning the benefits of frequent testing shows mixed results. Some research suggests a clear benefit of frequent testing on learning outcomes. For example, frequent testing was associated with greater delayed recall of presented information (Roediger & Karpicke, 2006), better retrieval of information (Carpenter & DeLosh, 2006), and higher exam scores not only for college students (Cone, 1990; Kling, McCorkle, Miller, & Reardon, 2005; Powell, 1977), but also for medical students (Larsen, Butler, & Roediger, 2013) and middle-schoolers (McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011). These learning outcomes have been observed regardless of the course topic (McDaniell et al., 2011) or type of exam questions (e.g., multiple choice, free response, or hybrid format; Smith & Karpicke, 2014).

Other research, however, shows a circumstantial benefit of frequent testing. For example, frequent testing was beneficial only in addition to a thorough review of the tested material (Brothen & Wambach, 2004). In summary, because there are mixed results in the literature about the effectiveness of frequent testing (Bell, Simone, & Whitfield, 2015), while previous studies propose that increasing exam frequency may have such positive effects such as reducing SES-based achievement gaps (e.g., Pennebaker, Gosling, & Ferrell, 2013), frequent testing and SES will be important components of the current dissertation.

Overreliance on Course Instructor Surveys

The end of semester Course Instructor Survey (CIS) is the most common method to evaluate course effectiveness (Clayson, 2009; Davis, 2009; Stark and Freishtat, 2014). However, CISs are problematic from several perspectives. There is converging evidence that CISs do not validly measure course effectiveness (Galbraith, Merrill, & Kline, 2012; Kornell & Hausman, 2016; Marsh, 2007; Stark & Freishtat, 2014; Stroebe, 2016; Uttl, White, & Gonzalez, 2017).

In a large-scale analysis of 116 courses, CIS ratings were negatively related to learning outcomes (Galbraith, Merrill, & Kline, 2012). A recent meta-analysis of previous meta-analyses on the relationship between CISs and learning reveals that contrary to more than 75 years of widespread belief that students learn more from instructors that receive higher CISs, in fact they do not (Uttl, White, & Gonzalez, 2017). Furthermore, there is a negative relationship between effectiveness and CIS ratings when learning from one course is measured in a follow-up course (Carrell & West, 2010; Kornell & Hausman, 2016; Stroebe, 2016).

Students' grade expectations influence CIS ratings, such that the lower the actual grade is from the expected grade, the lower the CIS ratings (Worthington, 2002). Additionally, CISs are influenced by the instructor's age, attractiveness, ethnicity, and sex (Kornell & Hausman, 2016; Marsh, 2007; Stark & Freishtat, 2014). Given the lack of evidence that CISs reflect course effectiveness regarding learning, and their subjective and biased nature, the current dissertation will not analyze CISs but instead use more objective measures of course effectiveness.

Dissertation Project

Is online education an effective and viable alternative to face-to-face education? The purpose of the current dissertation was to evaluate the effectiveness of online education at The University of Texas at Austin (UT-Austin). The dissertation focused on Synchronous Massive Online Courses (SMOCs) at UT-Austin from since 2012. The primary goal of this dissertation focuses on how course effectiveness varies as a function of lecture environment, comparing SMOCs to similar face-to-face courses (FTFs), while accounting for socioeconomic status (SES), scholastic ability, and exam frequency.

Scholastic ability will be used solely as a control variable. Because previous studies propose that increasing exam frequency may reduce SES-based achievement gaps (e.g., Pennebaker, Gosling, & Ferrell, 2013), and there are some mixed results in the literature about the effectiveness of frequent testing (e.g., Bell, Simone, & Whitfield, 2015), a secondary goal of this dissertation will focus on the interaction of SES and exam frequency in the context of course effectiveness outcomes.

The current dissertation was part of and funded by Project 2021 at UT-Austin. Project 2021 is a university-wide project that examines how we think about and deliver undergraduate education. Project 2021 researchers test innovative, research-based ideas about how to best bring about meaningful changes in education at a large scale, and relies heavily on technological advancements and the use of reliable and valid assessments of education. The implications for the current dissertation in the context of Project 2021 are that if SMOCs are as or more effective than traditional courses, then we will have evidence that we can teach larger numbers of students with fewer professors and teaching assistants. This frees up resources and time that could be devoted to more experiential learning opportunities and research productivity, all while maintaining the high standards of education. Additionally, this would be a great marketing tool regarding SMOC effectiveness in the university's pursuits to reach a broader population of students from outside the university.

Method

Participants

The dataset contained 25,726 students from 53 courses at UT-Austin. See Appendix 1 for a list of included courses and their descriptive statistics. Accounting for students that completed more than one course, there were 21,206 unique students. Of these, 54.37% were female and 45.63% were male. The mean age was 19.67 years (*SD* = 2.21 years), and is accurate to within +/- one year because of the method age was stored in university databases. Student classification was 36.28% Freshman, 31.69% Sophomore, 17.02% Junior, 14.97% Senior, and 0.04% Masters/Doctoral/Professional.

Variables of Interest

See Table 1 for descriptive statistics on all variables.

Final grade. Final course grade at the student level was indicated as 0 = F, 1 = D, 2 = C, 3 = B, 4 = A in the test course. Plus and minus grade distinctions were not considered. CR (i.e., credit/no credit) and X (i.e., incomplete) grades accounted for 1.34% of all final grades, and were excluded from Final Grade analyses.

Subsequent semester GPA. Subsequent semester grade point average (GPA) was indicated at the student level by the mean final letter grade earned in all courses the semester subsequent to the test course.

Course completion. Course completion was indicated at the student level in the test course as 0 = Dropped, 1 = Completed. Dropped courses are recorded after the 12^{th} day of class. Course drops before the 12^{th} day of class are not archived by the university and therefore were not included in this dataset.

Exam frequency. Exam frequency for every test course was calculated by weighting the raw exam frequencies as such:

exams(% of final grade) + # quizzes(% of final grade) + # final exams(% of final grade).

These exam data were coded from course syllabi.

Socioeconomic status. Socioeconomic status (SES) was indicated at the student level with a mean composite score taken from the education level of the student's mother and father, both on 7-point scales (1 = some high school, to 7 = Graduate/Professional degree).

Scholastic aptitude. Scholastic aptitude was indicated at the student level with a sum composite score taken from verbal and quantitative scores on the Scholastic Aptitude Test (SAT; Educational Testing Service, 2014).

Lecture environment. Lecture environment was indicated at the course level in the test course as 0 = FTF, 1 = SMOC.

Table 1: *Variable Descriptive Statistics*

Variable	n	Mean	SD	Median	Min	Max	Binary n %
Final Grade	24,141	2.87	1.82	3.00	0.00	4.00	
Subsequent Semester GPA	21,047	3.12	0.75	3.27	0.00	4.00	
Weighted Exam Frequency	25,726	7.65	8.37	3.30	0.00	22.88	
Socioeconomic Status	23,920	5.26	1.55	6.00	1.00	7.00	
Scholastic Aptitude	19,967	1217	157	1220	400	1600	
Lecture Environment	25,726						45.78% in FTF 54.22% in SMOC
Course Completion	25,726						95.17% Completed 4.83% Dropped

Note: N for complete dataset is 25,726.

Procedure

This project was IRB approved, and is recognized as having exempt status (under 45 CFR 46.101(b)(4)) by The University of Texas at Austin Institutional Review Board (reference number IRB 2012-07-0064). That is, this research is considered exempt from review and the need for written informed consent because it is educational research using archival data.

Researchers identified and included all SMOC courses at UT-Austin with a minimum of 200 students. Then, the last four same-topic FTF courses taught by the same professor/s were identified and included. If the professor did not teach the same-topic FTF course as the SMOC course, the most similar topic was chosen. The two cases of this were David Buss's Evolutionary Psychology courses in place of Psychology of Sex courses because he did not teach FTF Psychology of Sex courses, and Patrick

McDonald's International Relations courses in place of US Foreign Policy courses because he did not teach FTF US Foreign Policy courses. See Appendix 1 for descriptive statistics on the included 53 courses across eight course topics.

Next, researchers compiled all relevant student and course data archived in university databases and merged that with course data compiled from archived course syllabi. Researchers then cleaned and prepared the data for statistical analyses. All analyses were completed in the R statistical environment (R Project, 2016).

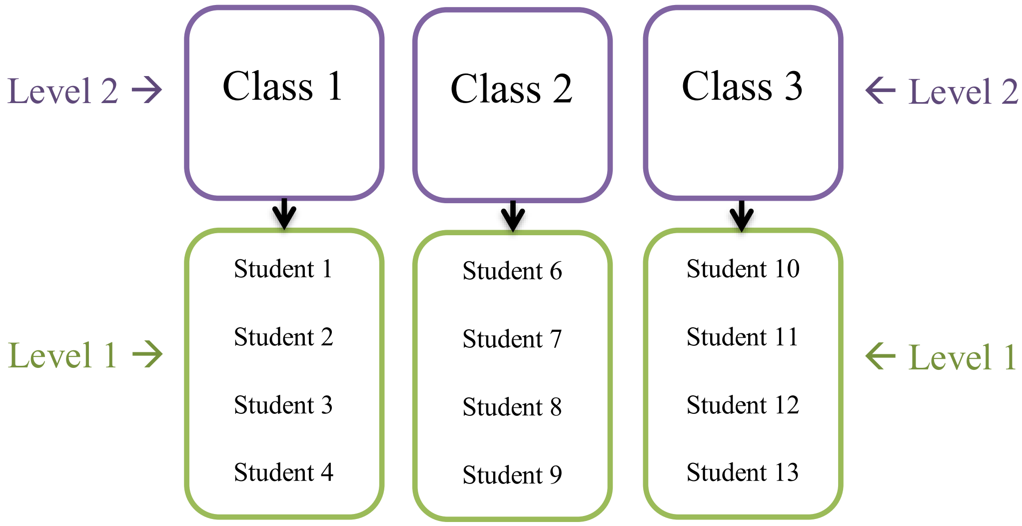
Results

Analytical Strategy

The following analyses were planned before data collection to partially fulfill requirements of the current dissertation. The analytical plan was pre-registered with and reviewed by the dissertation committee before data collection. The statistical reasoning and information was compiled from multiple sources (e.g., Field, Miles, & Field, 2012; Goldstein, 2011; Kline, 2011; Raudenbush, 1993; Raudenbush & Bryk, 2002; Tabachnick & Fidell, 2007; Woltman, Feldstain, MacKay, & Rocchi, 2012).

Justification for Hierarchical Linear Modeling. Hierarchical Linear Modeling (HLM) is a family of analyses also referred to as Multilevel Modeling, Mixed Effects Modeling, Random Effects Modeling, Random Coefficient Modeling, and Covariance Components Modeling. HLM is an extension of standard regression and is designed to properly model hierarchical (i.e., nested) data. For example, when students are nested within classes (see Illustration 1), it is important to partition the variability accounted for by students (i.e., Level 1) and classes (i.e., Level 2). That is, mean grades (i.e., *intercepts* in regression terms) and the relationships between a predictor and mean grades (i.e., *slopes* in regression terms) may vary between classes (i.e., a contextual variable), and ANOVA and standard regression analyses cannot properly account for this variability at multiple levels.

Illustration 1. A hierarchical data structure with two levels. For example, Class 1 could be Introduction to Psychology, and Class 2 could be American Government.



In the context of the current dissertation, if we wanted to test if course final grades vary as a function of lecture environment (SMOC or FTF) we could analyze the data with a between-subjects ANOVA using lecture environment (IV) and grades (DV). We could also run the same analysis in a standard regression framework by dummy coding lecture environment (predictor) to test if it predicts grades (criterion). The processing and results of these two models would be identical (i.e., they are doing the same thing).

However, this approach analyzes the variability at Level 1 only, which is problematic for several reasons. Hierarchical contextual variables (i.e., in this case, different classes) create dependencies in the data that violate the assumption of independence between cases. For example, data between Student 1 and Student 2 correlates at higher levels than data between Student 1 and Student 6 does (from Illustration 1) because Student 1 and 2 are in the same course and Student 6 is in a different course (e.g., instructors influence students in one course in similar ways;

students from one class influence each other; students with similar interests enroll in the same course; etcetera). HLM both (a) tests if the criterion varies at Level 2, and (b) if it does, properly models that variation. That is, if grades vary as a function of class (e.g., Introduction to Psychology compared to American Government and the remaining six classes), results from the ANOVA/standard-regression will be untrustworthy (e.g., Type I and II errors are more likely).

HLM tests our original question: do grades vary as a function of lecture environment (SMOC or FTF) by properly partitioning variability between student and class levels? HLM allows intercepts and slopes to vary across hierarchical levels (e.g., Level 2 in our example, the eight different classes), does not assume independence between cases (because it models the dependence), does not assume homogeneity of regression slopes for covariates (because it models variation in slopes between different groups), and handles unbalanced designs (i.e., unequal Ns) and missing data with ease (because of the way it estimates parameters). HLM also allows predictors to be entered at any level, and can test for within and between level interactions. For example, with HLM we can properly test if SES or intelligence (both Level 1), or exam frequency (Level 2), moderates the relationship between lecture environment (SMOC or FTF) and grades.

HLM Results

The structure of the current dissertation's data is such that it requires HLM modeling at three levels: 25,726 students at Level 1 which are nested within 16 courses at Level 2 (SMOC and FTF sections of each course-topic) which are nested within 8 course-topics at Level 3 (Introductory Psychology, Government, Economics, etcetera). For example, all FTF Introductory Psychology sections are one ‘course’, and all SMOC Introductory Psychology sections are one course. HLM properly controls for the correlations in data caused by students having a similar professors and experiences for FTF (hence, they are all one ‘course’), and students having a particular professors and experiences for SMOC. That is, students’ data in one ‘course’ are not independent from one another, so we should not use analyses that rely on them being independent (e.g., ANOVA) and will instead use HLM. This means variability between courses, and between course-topics, will be properly accounted for in all analyses.

Final grades. The first question we attempted to answer is: how do final course grades (criterion) vary as a function of lecture environment (predictor; SMOC or FTF)? Mean final course grades are displayed in Figure 1.

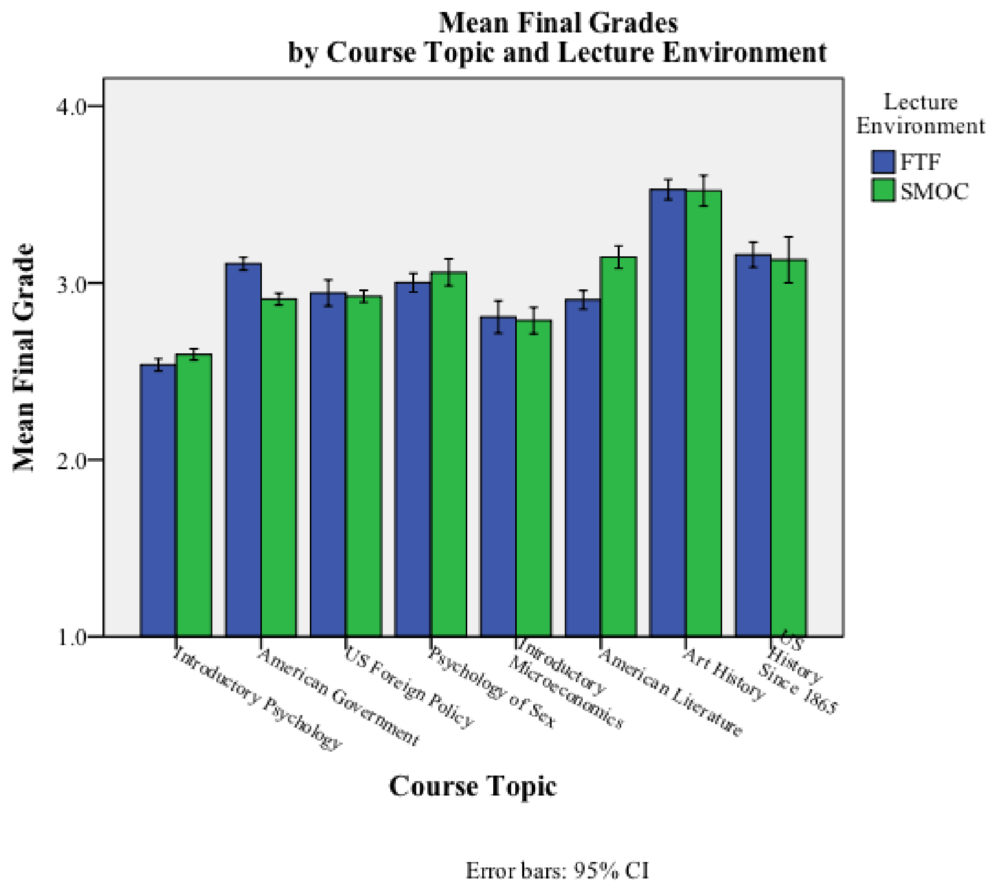


Figure 1. Mean final course grades by course topic and lecture environment, where error bars represent 95% confidence interval of the mean.

The first step compared a baseline model where only the fixed intercept was entered, with a random intercept model where the intercept was allowed to vary at Levels 2 and 3. A chi-square test revealed a significant difference between models ($p < .001$). In addition to the nested data structure suggesting HLM is needed, the random intercept model better fit the data, which means the contextual variables' (courses at Level 2 and course-topics at Level 3) variance require modeling in HLM and were modeled as such (i.e., random intercepts) moving forward.

The next step entered predictors into the model. The main predictor lecture environment (SMOC or FTF) was entered along with the covariates exam frequency, socioeconomic status (SES), and scholastic aptitude. Because there was particular interest in exam frequency as an explanatory variable, we also entered the exam frequency x SES, exam frequency x scholastic aptitude, and exam frequency x lecture environment interactions. That is, this model answers the question of how course grades vary as a function of lecture environment (SMOC or FTF), modeling and accounting for variance at Levels 2 and 3 (courses and course-topics), and also modeling and accounting for the relationships between exam frequency, SES, scholastic aptitude, exam frequency x SES, exam frequency x scholastic aptitude, and exam frequency x lecture environment.

Overall, 8.44% of the variance in the final course grades was accounted for by course topic (Level 3), 4.20% of the variance in the final course grades was accounted for by courses (Level 2), and 13.08% of the variance in the final course grades was accounted for by the predictors in the model. Therefore, 25.72% of the variance in the final course grades was accounted for by the model.

There was a significant Exam frequency x Lecture Environment interaction (b = 0.04, t = 4.98, p < .001); such that for FTFs, there was no substantial difference in final course grades by exam frequency; however, for SMOCs, students with more exams had higher final course grades than students with fewer exams. Importantly, the highest final grades were earned by students in SMOCs that provided the highest exam frequencies (while considering the control variables in the model). The Exam Frequency x Lecture Environment interaction is represented in Figure 2.

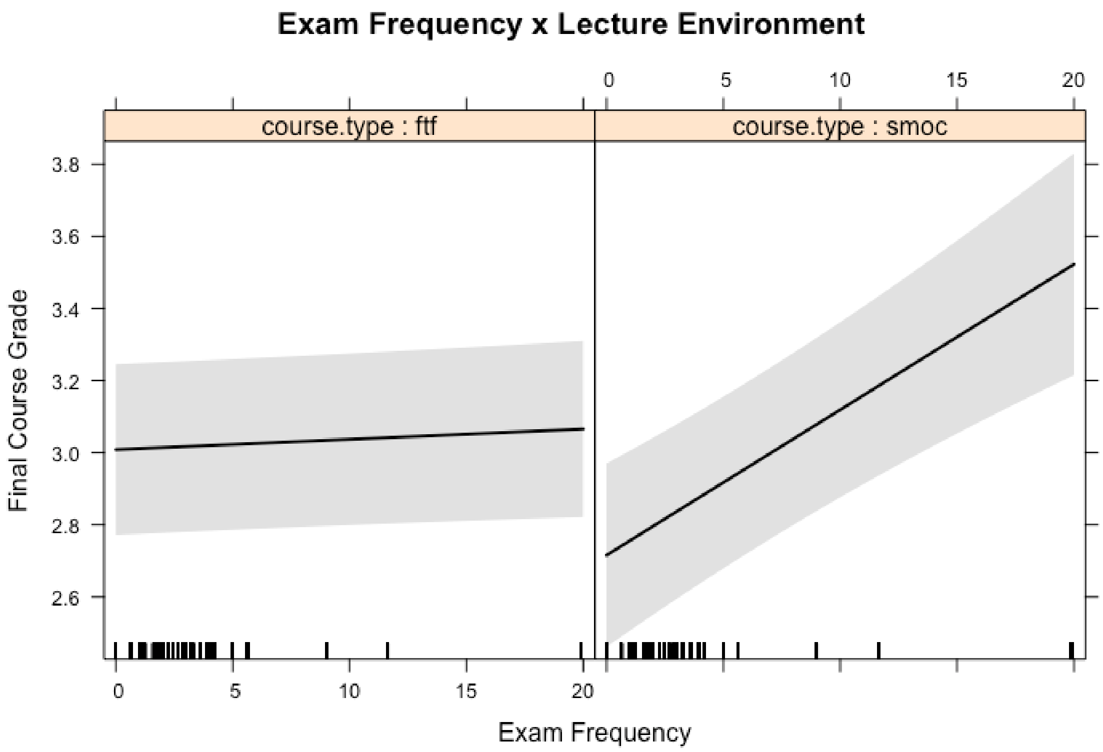


Figure 2. Predicting final course grades with the Exam Frequency x Lecture Environment interaction, where grey areas represent 95% confidence intervals. The highest level of exam frequency was observed only in some of the Introductory Psychology courses, N = 1485 in high-exam-frequency FTF sections, N = 4770 in high-exam-frequency SMOC sections.

There was also a significant Exam Frequency x Scholastic Aptitude interaction (b = 0.0004, t = 7.01, p < .001), but it is not central to the current dissertation's focus.

Additionally, SES positively predicted final course grades ($b = 0.05$, $t = 7.33$, $p < .001$), but did not interact with exam frequency ($p = .08$).

Subsequent semester grades. The next question we attempted to answer is: how do subsequent semester grades (criterion) vary as a function of lecture environment (predictor; SMOC or FTF)? Mean subsequent semester grades are displayed in Figure 3.

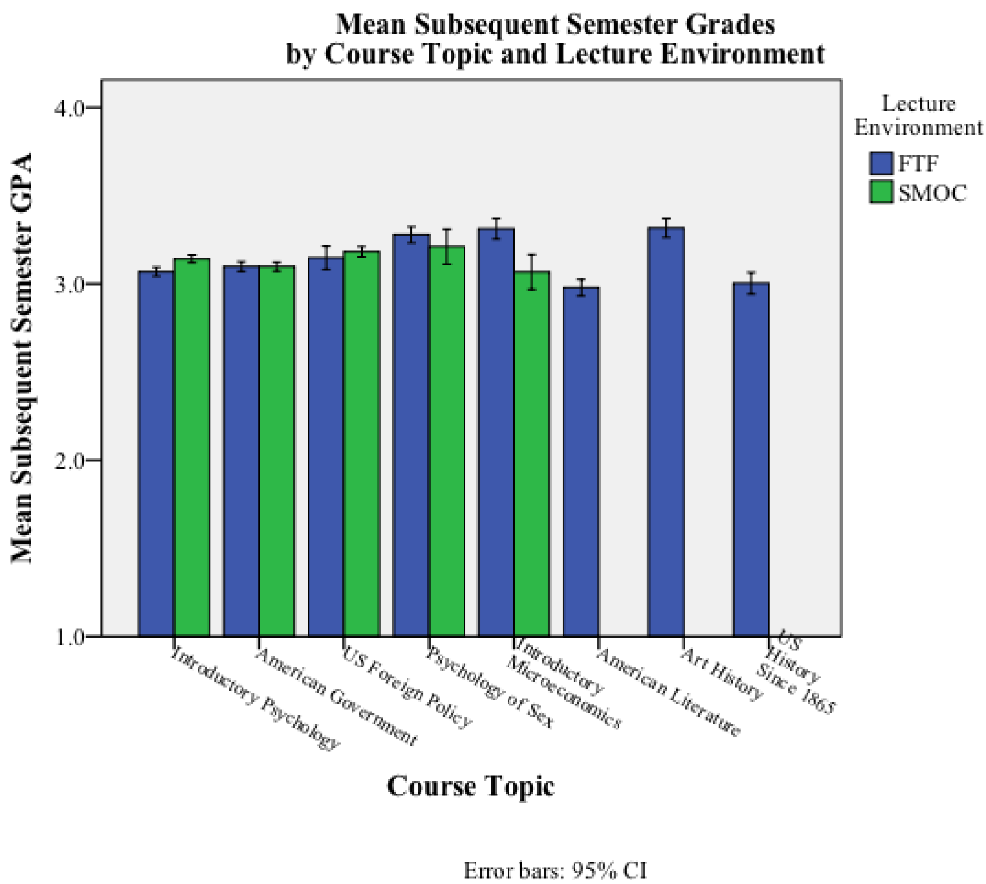


Figure 3. Mean subsequent semester grades by course topic and lecture environment, where error bars represent 95% confidence interval of the mean.

The same HLM process and model used to answer the first question was used to answer this one. The first step compared a baseline model where only the fixed intercept was entered, with a random intercept model where the intercept is allowed to vary at Levels 2 and 3. A chi-square test revealed a significant difference between models ($p < .001$). In addition to the nested data structure suggesting HLM is needed, the random intercept model better fit the data, which means the contextual variables' (courses at Level 2 and course-topics at Level 3) variance require modeling in HLM and were modeled as such (i.e., random intercepts) moving forward.

The next step entered predictors into the model. The main predictor lecture environment (SMOC or FTF) was entered along with the covariates exam frequency, socioeconomic status (SES), and scholastic aptitude. Because there was particular interest in exam frequency as an explanatory variable, we also entered the exam frequency x SES, exam frequency x scholastic aptitude, and exam frequency x lecture environment interactions. That is, this model answers the question of how subsequent semester GPA varies as a function of lecture environment (SMOC or FTF), modeling and accounting for variance at Levels 2 and 3 (courses and course-topics), and also modeling and accounting for the relationships between exam frequency, SES, scholastic aptitude, exam frequency x SES, exam frequency x scholastic aptitude, and exam frequency x lecture environment.

Overall, 0.78% of the variance in the subsequent semester grades was accounted for by course topic (Level 3), 0.72% of the variance in the subsequent semester GPA was accounted for by courses (Level 2), and 6.37% of the variance in the subsequent semester

GPA was accounted for by the predictors in the model. Therefore, 7.87% of the variance in the subsequent semester GPA was accounted for by the model.

There was no significant difference in subsequent semester GPA between SMOC and FTF courses (p = .51), and lecture environment did not interact with exam frequency (p = .65). There was a significant Exam frequency x Scholastic Aptitude interaction (b = 0.00001, t = 2.87, p = .004), but it is not central to the current dissertation's focus. Additionally, SES positively predicted subsequent semester GPA (b = 0.05, t = 8.90, p < .001), but did not interact with exam frequency (p = .96).

Course completion. The final question we attempted to answer is: how does course completion (binary criterion) vary as a function of lecture environment (predictor; SMOC or FTF)? Mean course completions are displayed in Figure 4.

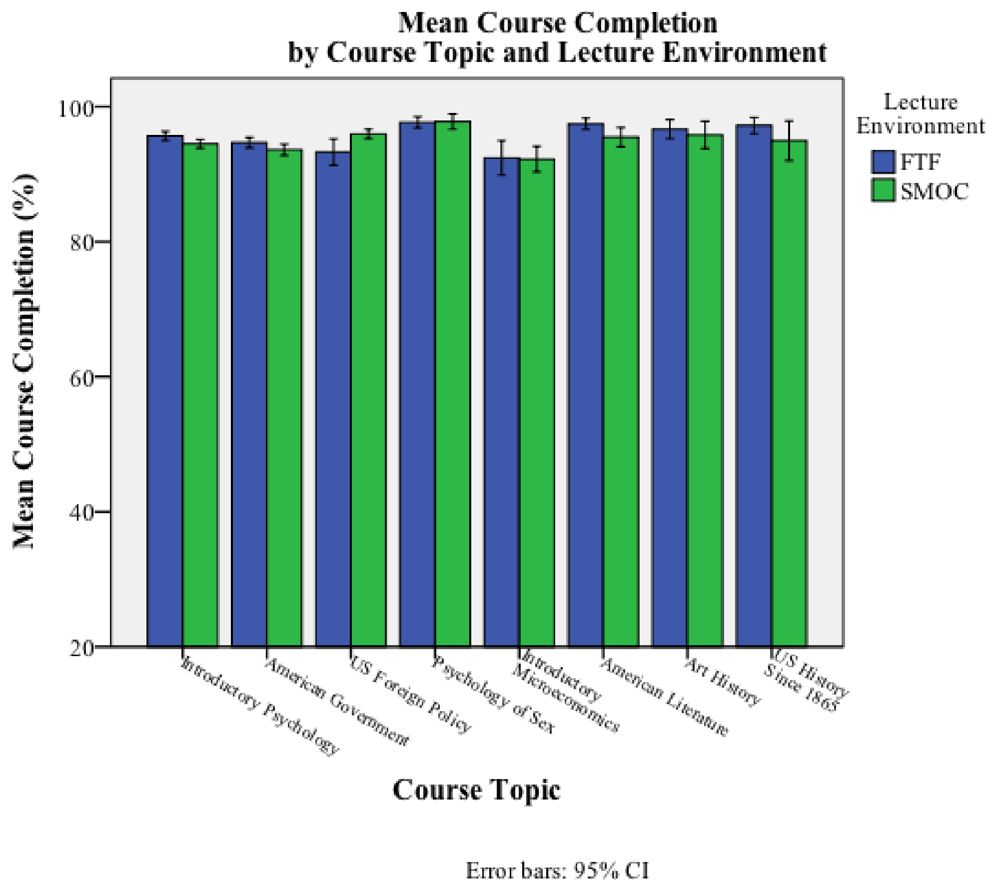


Figure 4. Mean course completions (%) by course topic and lecture environment, where error bars represent 95% confidence interval of the mean.

The same HLM process that was used to answer the first two questions was used to answer this one; however, because course completion is a binary variable, a logistic HLM was used. The first step compared a baseline model where only the fixed intercept was entered, with a random intercept model where the intercept is allowed to vary at Levels 2 and 3. A chi-square test revealed a significant difference between models ($p <$ .001). In addition to the nested data structure suggesting HLM is needed, the random intercept model better fit the data, which means the contextual variables' (courses at Level 2 and course-topics at Level 3) variance require modeling in HLM and were modeled as such (i.e., random intercepts) moving forward.

The next step entered predictors into the model. The main predictor lecture environment (SMOC or FTF) was entered along with the covariates exam frequency, socioeconomic status (SES), and scholastic aptitude. Because there was particular interest in exam frequency as an explanatory variable, we also entered the exam frequency x SES, exam frequency x scholastic aptitude, and exam frequency x lecture environment interactions. That is, this model answers the question of how the probability of course completion varies as a function of lecture environment (SMOC or FTF), modeling and accounting for variance at Levels 2 and 3 (courses and course-topics), and also modeling and accounting for the relationships between exam frequency, SES, scholastic aptitude, exam frequency x SES, exam frequency x scholastic aptitude, and exam frequency x lecture environment.

There was no significant difference in the probability of course completion between SMOC and FTF courses ($p = .40$), and lecture environment did not interact with exam frequency ($p = .49$). There was a significant Exam Frequency x SES interaction ($b = -0.006$, $z = -2.24$, $p = .025$); such that in courses with higher exam frequencies, lower SES students had higher probabilities of course completion, whereas higher SES students had no substantial difference in probability of course completion. This is important because it signals that a simple intervention—increasing exam frequency—is related to *increasing* course completions specifically for low-SES students. The SES x Exam

Frequency interaction is represented in Figure 5. There was also a significant Exam Frequency x Scholastic Aptitude interaction (b = 0.0001, z = 6.42, p < .001), but it is not central to the current dissertation's focus.

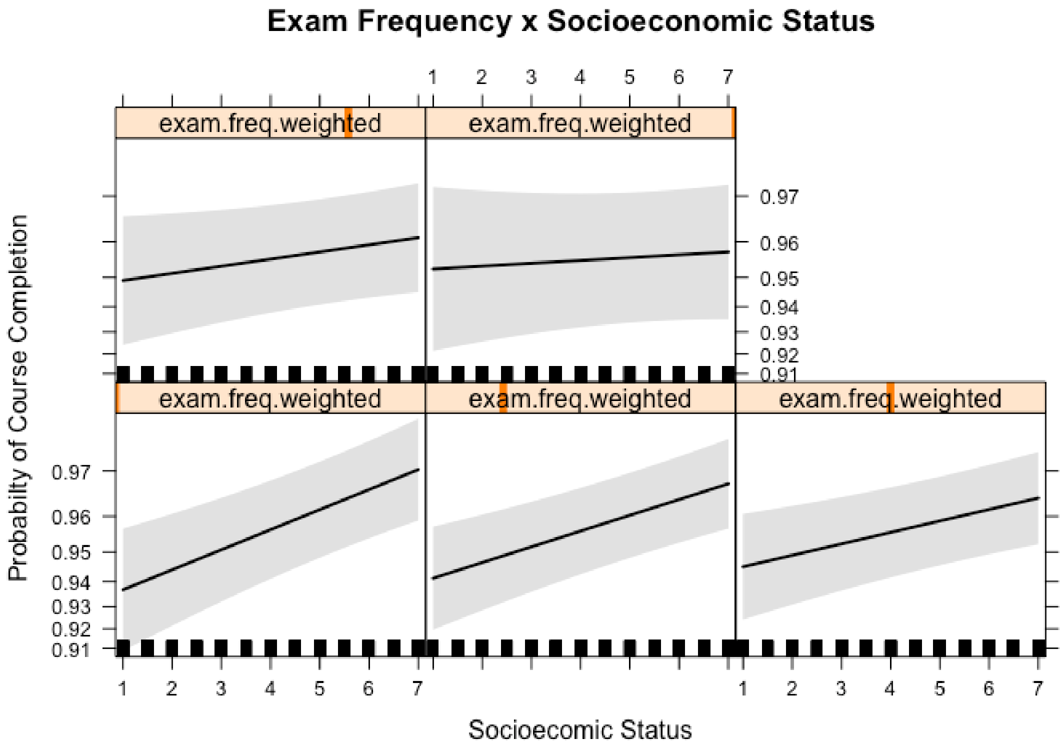


Figure 5. Predicting course completion with the Exam Frequency x SES interaction, where grey areas represent 95% confidence intervals, and where exam frequency is split into 5 quintiles. The plot in lower-left corner is the first quintile (designated by the vertical red line in the header of that box, all the way to the left), the lower-center plot is the second quintile, the lower-right plot is the third quintile (note the vertical red line is now centered in the header of that box), the upper-left plot is the fourth quintile, and the upper-center plot is the fifth quintile (note the vertical red line in the header of that box, all the way to the right). The highest level of exam frequency was observed only in some

of the Introductory Psychology courses, N = 1485 in high-exam-frequency FTF sections, N = 4770 in high-exam-frequency SMOC sections.

Discussion

The main goal of this project was to examine the effectiveness of SMOCs in comparison to FTFs. Course effectiveness was operationally defined with three objective outcomes: final course grades, subsequent semester GPAs, and course completions. Looking at main effects, there were no significant differences between SMOCs and FTFs on any of these objective measures. That is, these analyses of 25,726 students, aggregating across 53 courses, reveal that SMOCs neither outperform nor underperform FTFs in final grades, subsequent semester GPAs, or course completions.

Final course grades differed by lecture environment (SMOC versus FTF) only when factoring in the interaction with exam frequency. That is, final course grades in FTFs did not vary as a function of exam frequency. However, final course grades in SMOCs did vary as a function of exam frequency, such that SMOCs with more exams resulted in higher final course grades than SMOCs with fewer exams. Importantly, with variance appropriately partitioned at the hierarchical level, and with control variables in place, SMOCs with the most exams resulted in the highest final course grades of any courses, including FTFs; which is a valuable finding because we cannot see this by looking at mean final grades alone (e.g., looking at Figure 1). Given this finding, I recommend that courses structure themselves around relatively high exam frequencies.

Subsequent semester GPAs did not differ between SMOCs and FTFs, after accounting for variation by course topics, exam frequency, SES, and scholastic aptitude. This finding provides some evidence to contradict popular perceptions that online courses necessarily have negative long-term effects (e.g., inhibiting a student's ability to learn skills necessary for success in future courses). That is, students that experience SMOCs are just as successful in their next semester grade performances as students that experience FTF courses. Additionally, this finding holds even after controlling for the number of hours taken the next semester. Importantly, this finding can help counter arguments that first-semester or first-year students' educational outcomes might be harmed by attending SMOCs.

Course completions did not differ between SMOC and FTF courses, after accounting for variation by course topics, exam frequency, SES, and scholastic aptitude. Given some perceptions that online courses generally have higher dropout rates than FTF courses do, this finding provides some evidence that UT-Austin SMOCs do not. Indeed, SMOCs at UT-Austin do not exhibit the relatively high dropout rates sometimes observed in MOOCs at other institutions. Furthermore, just as many students in SMOCs complete courses as students in similar FTF courses do.

The creation of the SMOC at UT-Austin—an Introduction to Psychology course, co-taught and co-developed by James W. Pennebaker and Samuel D. Gosling, and co-developed and managed by Jason D. Ferrell—included a daily-exam intervention. Pennebaker, Gosling, and Ferrell (2013) hypothesized that frequent exams may explain positive outcomes in their large FTF courses, but a large-scale analysis of exam frequency has not been completed on SMOC courses. Also, there are some mixed results in the literature about the effectiveness of frequent testing (e.g., Bell, Simone, & Whitfield, 2015). If frequent exams are helping students in some way, in what ways and for whom are they helping? To examine these questions, the current dissertation used course effectiveness models (i.e., the same models that tested for differences by lecture environment) that included interactions between exam frequency and two individual difference variables that are related to the course effectiveness measures—scholastic aptitude and socioeconomic status. Because Pennebaker, Gosling, & Ferrell (2013) proposed that frequent testing may reduce SES-based achievement gaps, this dissertation was particularly focused on the relationships between exam frequency and SES.

Findings from the current dissertation show that exam frequency interacts with SES to predict course completion. Focusing on lower SES students, when exam frequencies are lower the probabilities of course completion are lower than when exam frequencies are higher; and when exam frequencies are higher, the probabilities of course completion are higher than when exam frequencies are lower. For higher SES students, the probabilities of course completion did not vary by exam frequency (i.e., course completion was relatively high for any exam frequency). Because almost 95% of students completed the courses in this dataset (i.e., the base-rate of dropouts is relatively low), future studies should replicate these analyses in college courses with much higher dropout rates.

This finding has far-reaching implications in the context of educational interventions that substantially help low SES students because not dropping out of a course helps facilitate not dropping out of college and thereby possibly increasing lifetime income and life satisfaction. Because course completion has important short- and long-term consequences for both students and administrators, the findings regarding exam frequency are promising. The findings show that increasing exam frequency is related to higher course completions for low-SES students (note: controlling for scholastic-ability). I recommend increasing exam frequency because these data provide some evidence that it will help keep disadvantaged students in class, and possibly, in college.

One major strength of the current dissertation is the large scale of the study, including a large sample of students, a large number of courses, a large variety of course topics, and a large variety of professors and teaching pedagogies. This is important because the power in the statistical analyses is relatively high and the external validity in terms of generalizability is relatively high. However, these findings from SMOCs may not generalize to all forms of online education (e.g., asynchronous, on-demand, self-paced, etcetera). Another strength of the current study is the analytical approach that included relatively complex multi-level modeling, which increases statistical validity and trustworthiness of the results. The study also benefitted from the relatively large number of control variables in the models, accounting for person variables, environment variables, and person x environment interactions, which increases the models' representations of these large interacting systems of variables as they operate in reality. Lastly, the current study is high in ecological validity because all data are taken from real-world courses.

Despite the strengths of the study, there are several limitations. Clearly, the study is correlational (i.e., students are not randomly assigned to conditions) and so we cannot make causal inferences from the results and cannot be certain that unmeasured variables are not causing the observed relationships. As difficult as it may be, future studies should randomly assign students to conditions wherever possible. Another weakness is the lack of additional information about exams. For example, it is likely that concept-mastery and application-based exams promote learning better than memorization and definition-based exams do, but we do not have that data to factor into these analyses. Given the apparent importance of testing frequency, future studies should code and record exam variables that could differentially affect course effectiveness. Furthermore, professors could create multiple versions of exams (application- and definition-based) and randomly assign students to conditions to test for short- and long-term differences in course effectiveness.

A third limitation of the current study is its relatively low number of person variables. For example, the personality trait *conscientiousness* is discussed in the literature as a large contributor to student success in college, and therefore it would be quite informative to include that data about students in these models. In the future, universities should collect additional education-relevant data about students, such as conscientiousness, because it would help inform our understanding of educational outcomes.

One of the study's strengths is also its weakness. With such a large sample size, we have a great deal of power to detect subtle effect sizes. Consequently, we were able to identify some significant main effects and interactions that predicted grades and course completion. It is important to emphasize that the effect sizes of some of these effects were quite small, sometimes accounting for less than 1% of the variance. In these cases, we should typically interpret effect sizes in the context of practical significance instead of p-values in the context of statistical significance.

Despite the complexity of the study and its shortcomings, the project is important in pointing to a new course delivery model. The SMOC is able to reach far more students in a given semester than traditional FTF courses. The course evaluations are generally equivalent. Economically, SMOCs are able to reach thousands of students by relying on fewer faculty without the need for large classrooms. At the same time, it frees faculty to teach more and smaller upper division courses and participate more in scholarship. Additionally, SMOCs open up the possibility of broadcasting high-standard education to people outside of the university. Although the results of the SMOC and FTF courses were generally similar, the additional payoffs of the SMOCs make them a promising tool for the future of undergraduate education.

Conclusion

Because SMOCs are reaching and affecting undergraduate and graduate students, instructors, administrators, the university, and society as whole, the results from this dissertation can be used to inform decisions about the future of undergraduate education. Looking across a wide range of course topics and courses, and large number of students, this dissertation provides evidence that SMOCs are as effective as FTFs on objective course outcomes, both short- and long-term. This includes final course grades, subsequent semester GPAs, and course completion rates as course effectiveness measures. This dissertation encompasses the first large-scale study of UT-Austin SMOCs showing that we can teach larger numbers of students with fewer professors and teaching assistants—which frees up resources and time that can be devoted to more experiential learning opportunities and research productivity—all while maintaining high standards of education. If the high standard of educational course effectiveness is based in the traditional FTF course, then a comparable SMOC course meets that high standard.

Appendix 1

Course Topic	Course Name	Semester	Lecture Environment	N	Exam Freq.	Weighted Exam Freq.	Professor/s
Art History							
	ARH 301	2013 Spring	FTF	75	13	3.90	Johns
	ARH 301	2013 Fall	FTF	174	3	1.65	Johns
	ARH 301	2014 Spring	FTF	208	8	1.90	Johns
	ARH 301	2014 Fall	FTF	175	8	1.90	Johns
	ARH 301	2016 Spring	SMOC	384	28	11.65	Johns
Masterworks of American Literature							
	E 316K	2012 Spring	FTF	240	7	1.15	Carton
	E 316K	2012 Summer	FTF	68	3	1.20	Carton
	E 316K	2012 Fall	FTF	285	9	2.65	Carton
	E 316K	2013 Fall	FTF	242	11	1.60	Carton
	E 316K	2014 Spring	FTF	473	0	0.00	Carton, Hutchinson
	E 316K	2014 Summer	FTF	69	2	0.60	Hutchinson
	E 316K	2016 Spring	SMOC	820	2	1.00	Carton, Hutchinson
Introductory Microeconomics							
	ECO 304K	2014 Fall	FTF	429	12	2.45	Houghton
	ECO 304K	2015 Spring	SMOC	265	15	2.90	Houghton
	ECO 304K	2016 Spring	SMOC	508	15	2.85	Houghton
American Government							
	GOV 310L	2011 Fall	FTF	388	3	2.25	McDaniel
	GOV 310L	2012 Spring	FTF	951	2	2.00	Shaw, McDaniel
	GOV 310L	2012 Fall	FTF	791	3	1.80	Theriault, McDaniel
	GOV 310L	2013 Spring	FTF	1080	15	5.63	Shaw, McDaniel
	GOV 310L	2013 Fall	SMOC	562	10	3.30	Shaw, McDaniel
	GOV 310L	2014 Spring	SMOC	981	10	3.00	Shaw, McDaniel
	GOV 310L	2014 Fall	SMOC	891	12	4.20	McDaniel, Albertson
	GOV 310L	2015 Spring	SMOC	959	10	4.00	Shaw, McDaniel
US Foreign Policy							
	GOV 312L	2013 Spring	FTF	384	3	1.80	Moser
	GOV 312L	2014 Fall	SMOC	838	14	3.00	Moser, McDonald

	GOV 312L	2015 Spring	SMOC	1073	15	3.60	Moser, McDonald
	GOV 312L	2015 Fall	SMOC	1003	3	1.80	Moser
	GOV 360N	2011 Fall	FTF	168	4	0.63	McDonald
	GOV 360N	2012 Fall	FTF	60	3	1.25	McDonald
	GOV 360N	2013 Spring	FTF	19	0	0.00	McDonald
	GOV 360N	2014 Spring	FTF	22	0	0.00	McDonald
US History Since 1865							
	HIS 315L	2012 Fall	FTF	215	0	0.00	Suri
	HIS 315L	2014 Spring	FTF	215	0	0.00	Suri
	HIS 315L	2015 Spring	FTF	283	0	0.00	Suri
	HIS 315L	2016 Spring	SMOC	218	0	0.00	Suri
Introductory Psychology							
	PSY 301	2008 Fall	FTF	981	4	3.00	Pennebaker, Gosling
	PSY 301	2010 Fall	FTF	1014	4	3.00	Pennebaker, Gosling
	PSY 301	2011 Fall	FTF	974	26	22.31	Pennebaker, Gosling
	PSY 301	2012 Fall	FTF	511	26	22.88	Pennebaker, Gosling
	PSY 301	2012 Fall	SMOC	856	26	22.88	Pennebaker, Gosling
	PSY 301	2013 Fall	SMOC	742	26	22.88	Pennebaker, Gosling
	PSY 301	2014 Fall	SMOC	1603	26	22.88	Pennebaker, Gosling
	PSY 301	2015 Fall	SMOC	1569	26	19.89	Pennebaker, Gosling
Psychology of Sex							
	PSY 334E	2012 Spring	FTF	86	4	4.00	Buss
	PSY 334E	2012 Fall	FTF	95	4	4.00	Buss
	PSY 334E	2103 Fall	FTF	94	4	4.00	Buss
	PSY 334E	2014 Spring	FTF	89	4	4.00	Buss
	PSY 341K	2011 Spring	FTF	449	5	5.00	Meston
	PSY 341K	2012 Spring	FTF	189	4	4.00	Meston
	PSY 341K	2013 Spring	FTF	188	4	3.20	Meston
	PSY 341K	2014 Spring	FTF	93	4	3.20	Meston
	PSY 346K	2015 Spring	SMOC	299	4	4.00	Meston, Buss
	PSY 306	2016 Spring	SMOC	378	30	9.00	Meston, Buss

References

Adamopoulos, P. (2013, December). *What makes a great MOOC? An interdisciplinary analysis of student retention in online courses.* Paper presented in the 34th International Conference on Information Systems (ICIS), Milan, Italy.

Alario-Hoyos, C., Munoz-Merino, P. J., Perez-Sanagustin, M., Delgado Kloos, C., & Parada, H. A. (2016). Who are the top contributors in a MOOC? Relating participants' performance and contributions. *Journal of Computer Assisted Learning, 32,* 232-243.

ALISON. (2016). ALISON home page. Retrieved from https://alison.com/.

Allen, I. E., & Seaman, J. (2013). *Changing course: Ten years of tracking online education in the United States*. Newburyport, MA: Pearson, and Sloan Consortium.

American College Testing. (2015, August). Online materials for the American College Test. Retrieved from http://www.actstudent.org.

Bell, M. C., Simone, P. M., & Whitfield, L. C. (2015). Failure of online quizzing to improve performance in introductory psychology courses. *Scholarship of Teaching and Learning in Psychology, 1,* 163–171.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123-144). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W.

Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56-64). New York, NY: Worth Publishers.

Boyd, R. L., Pennebaker, J. D., Ferrell, J. D., & Georgiev, N. (2015, July). Group discussions and academic performance: Exploring the classroom factors that contribute to student success. In *Proceedings of the Tenth Annual Meeting of the Interdisciplinary Network for Group Research (INGRoup)*, 1-8.

Beekman, D., Schoonhoven, L., Boucque, H., Van Maele, G., & Defloor, T. (2007). Pressure ulcers: E-learning to improve classification by nurses and nursing students. *Journal of Clinical Nursing, 17,* 1697-1707.

Bernard, R. M., Abrami, P. C., Lou, Y., Borokhovski, E., Wade, A., Wozney, L., ...Huang, B. (2004). How does distance education compare with classroom instruction? A meta-analysis of the empirical literature. *Review of Educational Research, 74,* 379-439.

Brothen, T., & Wambach, C. (2004). The value of time limits on Internet quizzes. *Teaching of Psychology, 31*, 62– 64.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative processing explanation of the testing effect. *Memory & Cognition, 34*, 268 –276.

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy, 118,* 409-432.

Cone, A. L. (1990). Frequency of testing in criterion-based learning. *Psychological Reports, 67,* 396-399.

Cronbach, L. J., & Snow, R. E. (1969). *Individual differences in learning ability as a function of instructional variables.* Stanford, CA: Stanford University Press.

Chumley-Jones H. S., Dobbie, A., & Alford, C. L. (2002). Web-based learning: Sound educational method or hype? A review of the evaluation literature. *Academic Medicine, 77,* 86-93.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education, 31*, 16-30.

Davis, B. G. (2009). *Tools for teaching* (2nd ed.). San Francisco, CA: John Wiley & Sons.

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

De la Garza, L. Y. A., Sancho-Vinuesa, T., & Zermeno, M. G. G. (2015). Atypical: Analysis of a Massive Open Online Course (MOOC) with a relatively high rate of program completers. *Global Education Review, 2,* 68-81.

Domjan, M. (2015). *The principles of learning and behavior* (7th ed.). Stamford, CT: Cengage Learning.

Dumbald, B. (2014). A brief history of online learning. Retrieved from http://www.straighterline.com/blog/brief-history-online-learning-infographic/.

Educational Testing Service. (2014, May). Online materials for the Scholastic Aptitude Test. Retrieved from https://www.ets.org/highered.

edX. (2016). edX: About us. Retrieved from https://www.edx.org/about-us.

Ferrell, J. D., Tucker-Drob, E. M., Yarkoni, T., Gosling, S. D., & Pennebaker, J. W. (in preparation). A multivariate model of academic success: A large-scale analysis of individual differences that promote and prevent success.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R.* Los Angeles, CA: Sage.

Fortune, M. F., Shifflett, B., & Sibley, R. E. (2006). A comparison of online (high tech) and traditional (high touch) learning in business communication courses in Silicon Valley. *Journal of Education for Business, 81*, 210-214.

Galbraith, C. S., Merrill, G. B., & Kline, D. M. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and Bayesian analyses. *Research in Higher Education*, *53*, 353-374.

Goldstein, H. (2011). *Multilevel statistical models (*4th ed.). New York, NY: John Wiley & Sons.

Harris, J. M., Elliot, T. E., Davis, B. E., Chabal, C., Fulginiti, J. V., & Fine, P. G. (2007). Educating generalist physicians about chronic pain: Live experts and online education can provide durable benefits. *American Academy of Pain Medicine, 9,* 555-563.

Herman, T., & Banister, S. (2007). Face-to face versus online coursework: A comparison of costs and learning outcomes. *Contemporary Issues in Technology Education, 7*, 318-326.

Hickey, R. (2014). The history of online education. Retrieved from https://www.petersons.com/college-search/online-education-history.aspx.

Hu, S., & Kuh, G. D. (2003). Maximizing what students get out of college: Testing learning productivity model. *Journal of College Student Development, 44,* 185-203.

Hugenholtz, N. I. R., de Croon, E. M., Smits, P. B., van Dijk, F. J. H., & Nieuwenhuijsen, K. (2008). Effectiveness of e-learning in continuing medical education for occupational physicians. *Occupational Medicine*, *58,* 370-372.

Kashyap, S. (2016). How Khan Academy plans to crack the Indian education market. Retrieved from http://yourstory.com/2016/05/khan-academy/.

Kline, R. B. (2011). *Principles and practice of Structural Equation Modeling* (3rd ed.). New York, NY: The Guilford Press.

Kling, N., McCorkle, D., Miller, C., & Reardon, J. (2005). The impact of testing frequency on student performance in a marketing course. *Journal of Education for Business, 81*, 67-72.

Kolb, D. A. (1984). *Experiential learning.* Englewood Cliffs, NJ: Prentice Hall.

Kornell, N., & Hausman, H. (2016). Do the best teachers get the best ratings? *Frontiers in Psychology*, *7,* 1-8. doi:10.3389/fpsyg.2016.00570

Koory, M. A. (2003). Differences in learning outcomes for the online and F2F versions of "An Introduction to Shakespeare". *Journal for Asynchronous Learning Networks, 7*, 18-39.

Larsen, D. P., Butler, A. C., & Roediger, H. L. III. (2013). Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Medical education, 47,* 674-682.

Machtmes, K., & Asher, W. (2000). A meta-analysis of the effectiveness of telecourses in distance education. *American Journal of Distance Education, 14,* 27-46.

Marsh, H. W. (2007). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J C. Smart (Ed.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp.319-384). New York, New York: Springer.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. III. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103,* 399-414.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2009). *Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies*. Report prepared for U.S. Department of Education Office of Planning, Evaluation, and Policy Development Policy and Program Studies Service Center for Technology in Learning.

Mesquita, A., & Peres, P. (2015). *Furthering higher education possibilities through Massive Open Online Courses.* Hershey, PA: Information Science Reference.

MIT Office of Digital Learning. (2016). MITx on edX (MOOCs). Massachussets Institute of Technology: Office of Digital Learning. Retrieved from http://odl.mit.edu/beyond-campus/mitx-edx-moocs.

Morabito, M. G. (2015). CALCampus origins. Retrieved from http://www.calcampus.com/calc.htm.

National Center for Education Statistics. (2014). Digest of education statistics. Retrieved from http://nces.ed.gov/programs/digest/d14/tables/dt14_311.22.asp?current=yes.

Noer, M. (2012). One man, one computer, 10 million students: How Khan Academy is reinventing education. Retreived from http://www.forbes.com/sites/michaelnoer/2012/11/02/one-man-one-computer-10-million-students-how-khan-academy-is-reinventing-education/#28fb6d703c05.

North, S. H., Richardson, R., & North, M. M. (2014). To adapt MOOCS, or not? That is no longer the question. *Universal Journal of Educational Research, 2,* 69-72.

Onink, T. (2013). Georgia Tech, Udacity shock higher ed with $7,000 degree. Retrieved from http://www.forbes.com/sites/troyonink/2013/05/15/georgia-tech-udacity-shock-higher-ed-with-7000-degree/#f12e44a7e895.

Pappano, L. (2012). The year of the MOOC. *The New York Times*, *2*(12). Retreived from http://www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html?pagewanted=all&_r=1.

Pennebaker, J. W., Gosling, S. D., & Ferrell, J. D. (2013). Daily online testing in large classes: Boosting college performance while reducing achievement gaps. *PLoS ONE, 8*: e79774. doi:10.1371/journal.pone.0079774

Powell, R. W. (1977). Grades, learning and student evaluation on instruction. *Research in Higher Education, 7,* 193-205.

R Project. (2016). The R Project for statistical computing. Retrieved from https://www.r-project.org/.

Raudenbush, S. W. (1993). Hierarchical linear models and experimental design. *Applied Analysis of Variance in Behavioral Science*, *137*, 459-496.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rayyan, S., Fredericks, C., Colvin, K. F., Liu, A., Teodorescu, R., Barrantes, A., ... & Pritchard, D. E. (2016). A MOOC based on blended pedagogy. *Journal of Computer Assisted Learning*, *32,* 190-201. doi:10.1111/jcal.12126

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning taking memory tests improves long-term retention. *Psychological Science*, *17*, 249-255.

Smith, M. A., Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory, 22*, 784–802.

Stark, P. B., & Freishtat, R. (2014). An evaluation of course evaluations. *ScienceOpen.* Center for Teaching and Learning, University of California, Berkley. Retrieved from https://www.scienceopen.com/document?id=6233d2b3-269f-455a-ba6b-dc3bccf4b0a8. doi:10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Straumsheim, C. (2013). Don’t call it a MOOC. *Inside Higher Ed*. Retrieved from https://www.insidehighered.com/news/2013/08/27/ut-austin-psychology-professors-prepare-worlds-first-synchronous-massive-online

Stroebe, W. (2016). Why good teaching evaluations may reward bad teaching: On grade inflation and other unintended consequences of student evaluations. *Perspectives on Psychologial Science, 11,* 800-816.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson.

Tallent-Runnels, M. K., Lan, W., Cooper, S., Ahern, T. C., Shaw, S. M. & Liu, X. (2006). Teaching courses online: A review of the research. *Review of Educational Research, 76,* 93-135.

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty’s teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54,* 22-42.

von Stumm, S., Hell, B., & Chamorro-Premuzic, T. (2011). The hungry mind: Intellectual curiosity is the third pillar of academic performance. *Perspectives on Psychological Science, 6,* 574–588.

Waldrop, M. M. (2013). Massive open online courses, aka MOOCs, transform higher education and science. Retrieved from http://www.scientificamerican.com/article/massive-open-online-courses-transform-higher-education-and-science/#.

Warren, L. L., & Holloman, H. L. (2005). On-line instruction: Are the outcomes the same? *Journal of Instructional Psychology, 32*, 148-150.

Weber, J. M., & Lennon, R. (2007). Multi-course comparison of traditional versus web-based course delivery systems. *The Journal of Educators Online, 4*, 1-19.

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to Hierarchical Linear Modeling. *Tutorials in Quantitative Methods for Psychology, 8*, 52-69.

Worthington, A. C. (2002). The impact of student perceptions and characteristics on teaching evaluations: A case study in finance education. *Assessment and Evaluation in Higher Education*, *27*. 49–64.

Zhang, D. (2005). Interactive multimedia-based e-learning: A study of effectiveness. *The American Journal of Distance Education, 2005,* 149-162.