Beyond Time-on-Task: The Relationship Between Spaced Study and Certification in MOOCs

Yohsuke R. Miyamoto

School of Engineering and Applied Sciences, Harvard University, USA

yrm@seas.harvard.edu

Cody A. Coleman

Massachusetts Institute of Technology, USA

Joseph Jay Williams, Jacob Whitehill, Sergiy Nesterko, Justin Reich

HarvardX at Harvard University, USA

ABSTRACT: A long history of laboratory and field experiments have demonstrated that dividing study time into many sessions is often superior to massing study time into few sessions, a phenomenon known as the "spacing effect." We use this well-established finding from the psychology literature as inspiration for investigating how students distribute their study sessions across an entire Massive Open Online Course (MOOC). Drawing on observational tracking log data from 20 HarvardX courses, we examine the relationship between students' allocation of their time in MOOCs and their performance. While controlling for the effect of total time, we show that the number of sessions students initiate is correlated with certification rate, across students in all courses. A one-unit change in session count is positively associated with an estimated 3.4% change in certification odds. When individual students spend similar amounts of time in multiple courses, they perform better in courses where that time is distributed among more sessions, suggesting that the benefit of spacing MOOC study sessions is independent of student characteristics. Our study demonstrates that well-established learning theories can be combined with massive new datasets and innovative approaches to learning analytics to advance our understanding of student practice and learning.

Keywords: MOOC, spacing effect, spaced practice, distributed practice, total time, sessions, online education

Editor's Note: As part of the Special Section on Learning Analytics & Learning Theory this article is followed by a short commentary on pp. 70-74 that discusses the challenges it faced and successes it achieved in drawing on and contributing to theory use in learning analytics.

1   INTRODUCTION

1.1   Early MOOC Research on Participation and Time-on-Task

Much of the early research in Massive Open Online Courses (MOOCs) has focused on measures of student participation, such as total time spent, number of click events produced, minutes of video

watched, or number of assignments completed. These studies have repeatedly observed two commonplace findings: 1) that the level of participation along one dimension of a course is a predictor of participation along other dimensions, and 2) that the level of participation is a predictor of better grades and course completion (Collins, 2013; DeBoer, Ho, Stump, & Breslow, 2014; Murphy, Gallagher, Krumm, Mislevy, & Hafter, 2014; Reich et al., 2014; Wilkowski, Deutsch, & Russell, 2014), outcomes often treated as proxies for learning. These results align with the total-time law in psychology: the amount of learning is a direct function of study time (Cooper & Pantle, 1967; Underwood, 1970).

Psychologists, however, have long known that not all uses of study time produce equal benefits. One striking exception to the total-time law is known as the *spacing effect:* for most learning outcomes, shorter, more spaced out study sessions are preferable to massed study sessions. We use this spacing effect research as inspiration for investigating the relationship between how students allocate their time in MOOCs and how they perform in their course. Specifically, while controlling for total time spent in MOOCs, we examine whether a higher number of study sessions is associated with better performance. The decades of established research in experimental psychology, education, and cognitive science should help guide new practical research (Benassi et al., 2014; Clark & Mayer, 2011; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Pashler et al., 2007; Williams, 2013). Here we demonstrate that well-established learning theories can produce important hypotheses for analysis leveraging massive new datasets and innovative new methods in learning analytics.

1.2   What is the Spacing Effect?

The spacing effect, initially documented by Herman Ebbinghaus in 1885, is the phenomenon where *distributed* presentations of material result in better long-term retention than that attained from *massed* presentations (all at once) of the same material, for a given amount of study time (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1989; Ebbinghaus, 1885; Greene, 1989; Melton, 1970). For example, the long-term retention of some element A is stronger following A1 and A2 (two presentations of A, each for 30 minutes) with a day in between (distributed practice) versus with no break (60 contiguous minutes of massed practice), even though the total time spent studying is the same in both cases. The dependability and robustness of this effect is demonstrated by its replication in a wide variety of experimental tasks over numerous studies (Melton, 1970; Moulton et al., 2006; Peterson, Wampler, Kirkpatrick, & Saltzman, 1963; Shea, Lai, Black, & Park, 2000; Underwood, 1970; Waugh, 1970; Young, 1967). Moreover, the quantitative advantage earned from a spaced schedule is remarkable, with reports that two spaced presentations are about twice as effective as two massed presentations (Hintzman, 1974; Melton, 1970), and the difference between them increasing further as a function of the number of presentations (Underwood, 1970).

Despite such potential for bolstering learning, the spacing effect has suffered a history of slow translation into standard educational practices (Dempster, 1988). In an effort to accelerate integration in practice, studies have appeared in recent years of real-world classroom demonstrations of the spacing

effect in many contexts, such as vocabulary learning (Bird, 2010; Carpenter, Pashler, & Cepeda, 2009; Gallo & Odu, 2009; Kang, Lindsey, Mozer, & Pashler, 2014; Khajah, Lindsey, & Mozer, 2014; Lindsey, Shroyer, Pashler, & Mozer, 2014; Rohrer, 2009; Rohrer & Taylor, 2006; Sobel, Cepeda, & Kapler, 2011), fraction learning (Rau, Aleven, Rummel, & Pardos, 2014), and skill acquisition (Moulton et al., 2006; Stafford & Dewar, 2014). Technology has enabled what is perhaps the most visible application of the spacing effect today: spaced repetition software (SPS). These are systems of digital flashcards that help sequence the presentation of study material in a spaced manner (Godwin-Jones, 2010).

1.3   The Current Study in Relation to Previous Studies of Spacing

A central insight emerging from the spaced practice literature is that while the total time invested in learning matters, there can be better and worse ways to allocate student time. Our study of student time in MOOCs builds upon this insight, with two important conceptual and methodological differences from previous research.

First, the outcomes that we examine in this study are quite different from those found in the spaced practice literature. Most research on spaced practice has been focused on narrowly defined, discrete outcomes, such as retention of vocabulary facts. In this study, our dependent variable is certification in 20 MOOCs, each of which varies considerably in the types of assessment, level of rigour, and grading schemes. The mechanisms by which spacing study sessions might affect student performance in a MOOC may be quite different from the mechanisms by which distributed practice might affect student vocabulary retention. The advantage of investigating a more general outcome measure is that any findings might be generalized broadly to a variety of common online learning situations. The disadvantage of our approach is that the activities and outcomes we investigate are in more of a "black box" relative to many study designs in the spaced practice literature, and our research represents a conceptual leap from the tradition of spaced practice research. We highlight this distinction by referring to the phenomenon of MOOC students allocating their time across multiple sessions as "spaced study sessions" rather than "spaced practice."

Second, much of the research on spaced practice takes advantage of experimental designs that can demonstrate causal relationships and internal validity in specific circumstances. Complementary to this work, our data allows us to observe large numbers of students who study diverse course content and spend their time across multiple courses in diverse ways. While we cannot draw causal conclusions from our data, we can investigate the effects of spaced study sessions "in the wild," beyond settings where experimental psychologists control critical elements of course design or student time.

1.4   Research Questions

We begin our research with the hypothesis, inspired by spaced practice literature, that among students who spend similar amounts of time on a MOOC, those who distribute their time into more sessions will

perform better than those with fewer sessions. We test this hypothesis by addressing two research questions.

First, among all students, do students with more study sessions earn certificates at a higher rate, controlling for total time? We first address this question with a set of non-parametric binning and bootstrap analyses. Then, we conduct a logistic regression analysis that allows us to control for additional confounding factors such as "struggle," or the degree to which low-performing students may have lower session count due to spending more contiguous time on fewer sections of the course.

Second, with individual students who take two courses, do students earn certificates at higher rates in the course where they have more spaced study sessions? One concern with our analysis to the first research question is that we cannot account for unobserved differences between students. To address this concern, we take advantage of the enormous population of MOOC students and conduct a second investigation to examine within-student variation exclusively. We examine only the subset of students who take two courses, spend similar amounts of standardized (z-scored) time in both courses, and space their time over different numbers of sessions.

We find evidence, both between students and within students, that spaced study sessions are associated with better performance in MOOCs. While imperfect compared with experimental trials, our methods demonstrate how large datasets allow us to account for learner variation in observational studies. These findings suggest a simple way of improving student learning, agnostic to the nature of specific activities and course content: by providing interventions to encourage students to complete the course with more sessions for the amount of total study time they have available.

In the sections that follow we present our data, methods, and findings, and suggest a set of possible experimental interventions to encourage spaced study sessions in MOOCs.

2   RESEARCH DESIGN

2.1   Data Collection

To explore the relationship between student performance and spaced study in MOOCs, we examined the timing of click events recorded in the tracking logs of all 101,913 unique students (with a combined 127,868 course registrations) with non-zero grades in 20 HarvardX courses (Table 1).

HarvardX is an online learning initiative of Harvard University, and provides the 20 courses examined herein on the edX MOOC platform. The different types of online click events triggered by students range from page navigation, to lecture video plays, to problem submissions, as shown in Figure 1b, and the diverse set of topics of these courses range from biostatistics (Health in Numbers), to philosophy (Justice). Some courses were offered on multiple occasions, such as The Ancient Greek Hero, and

Justice, while many others were offered on a single occasion. ChinaX is a series of separate courses with an ongoing theme that continued for 10 modules (the first six of which finished in time for inclusion in this study), and thus tend to have overlapping, devoted students, resulting in high certification rates. Certification rates from its later modules also tend to be high because of attrition from earlier modules. Since we use certification as the outcome of interest in subsequent analysis, we restrict our investigation to students who answer at least one problem correctly, to exclude casual browsers and auditors (Ho et al., 2014).

Table 1. Course Information for Twenty 2012–2014 HarvardX Courses

Course title	Course code	Start date	End date	# students with non-zero grade	# certified students (% certified of those with non-zero grade)
The Ancient Greek Hero	CB22x	3/13/13	8/26/13	4663	1395 (30%)
The Ancient Greek Hero	CB22.1x	9/3/13	12/31/13	2625	727 (27%)
Justice	ER22x	3/2/13	7/26/13	11896	5265 (44%)
Justice	ER22.1x	4/8/14	7/17/14	6307	2483 (39%)
Unlocking the Immunity to Change	GSE1x	3/11/14	6/30/14	16641	1854 (11%)
Leaders of Learning	GSE2x	7/8/14	8/25/14	11191	3933 (35%)
Fundamentals of Clinical Trials	HSPH	10/14/13	2/14/14	6422	2406 (37%)
Health in Numbers: Quantitative Methods in Clinical & Public Health Research	PH207x	10/15/12	1/30/13	16541	4910 (30%)
United States Health Policy	PH210x	4/7/14	6/30/14	3530	759 (25%)
Human Health and Global Environmental Change	PH278x	5/15/13	7/25/13	6544	2711 (41%)
Data Analysis for Genomics	PH525x	4/7/14	6/30/14	4501	621 (14%)
Science and Cooking: From Haute Cuisine to Soft Matter Science	SPU27x	10/8/13	3/15/14	10274	1794 (17%)
The Political and Intellectual Foundations of China (ChinaX)	SW12x	10/31/13	12/23/13	8216	2016 (25%)
The Creation and End of a Centralized Empire (ChinaX)	SW12.2x	1/2/14	1/30/14	3624	1751 (48%)
Cosmopolitan Tang: Aristocratic Culture (ChinaX)	SW12.3x	2/13/14	3/6/14	2509	1565 (62%)
A New National Culture (ChinaX)	SW12.4x	3/20/14	4/10/14	2140	1230 (57%)
From Global Empire to Global Economy (ChinaX)	SW12.5x	4/24/14	5/8/14	1607	1117 (70%)
The Last Empire (ChinaX)	SW12.6x	5/22/14	6/19/14	1845	1108 (60%)
Global Health: Case Studies from a Biosocial Perspective	SW25x	2/25/14	5/31/14	3204	1266 (40%)
Tangible Things	USW30x	6/2/14	8/2/14	3588	1089 (30%)

2.2   Measures

2.2.1 Outcome: Certification

We use certification in a course as our outcome of interest. Certification was awarded to students who achieved a minimum level of performance on a combination of quiz scores, homework assignments, project results, and other assessments — usually 50 or 60 out of 100 points, a threshold that varies from course to course. We use certification as a proxy for learning, but since courses differ in their assessment structures and grading schemes, the specific meaning of certification varies. In some courses students earn points in a small number of higher-stakes assessments, such as the two midterms and one final exam in JusticeX. In other courses students earn points over many lower stakes assessments, such as the many short quizzes in HeroesX. Some courses attempt to replicate the rigours of their residential counterparts, and others are designed to be more accessible to a wider audience. It is important to note that most studies of spacing effects focus on measures of memory and retention, rather than this overall measure of course performance.

2.2.2 Time and Session

We are primarily interested in two predictors of certification that we extracted from the course tracking logs: 1) total time spent, and 2) the number of sessions among which their time was distributed. As shown in Figure 1a, a session was defined as a collection of click events separated by periods of inactivity that lasted more than 30 minutes, in accordance with the Google Analytics standard for defining sessions for website usage.[1] Total time spent was thus calculated by summing the lengths of all sessions by a student. A few example sessions shown in Figure 1b, taken from The Ancient Greek Hero and ChinaX, demonstrate the patterns of click events that might occur during a session.

Our metric of total time includes the time spent on all activities in the course. The patterns and frequencies of these event types can vary widely from course-to-course, so we took this simplified approach of treating all types of activity equally to keep our analyses agnostic to the wide course-to-course differences in content, structure, and certification requirements. This has the advantage that potential interventions that take advantage of our findings may be flexibly applied to a wide variety of courses.

---

[1] "How a session is defined in Analytics," https://support.google.com/analytics/answer/2731565?hl=en
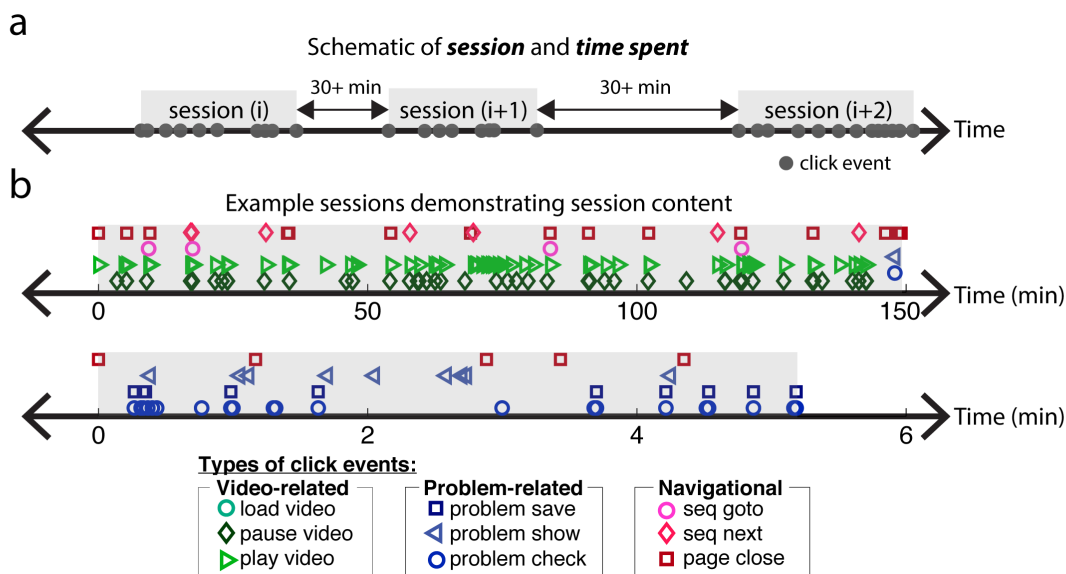
Figure 1. Illustrations of session and total time. (a) Schematic of the definition of session and total time. (b) Example sessions taken from the data, which demonstrate the variety and combinations of event types initiated by students. The top example demonstrates a session focused on video usage, interleaved with page navigation events. The bottom example demonstrates a session that focuses on quiz problems, with occasional page navigation. Note that the set of event types shown in these examples (9 event types) does not include all possible event types. (A comprehensive list and description of event types can be found at the edX research guide).[2]

2.2.2 Control Predictors for "Struggle"

One possible confound to our analysis is the degree to which students struggle with the course material. For instance, it may be that students with high total time and low session count spend more of their time struggling with the material — resulting in more massed study sessions. These same students might be expected to drop out more frequently and therefore earn certificates at lower rates. We attempt to account for this confound by including three additional control predictors. First, we include *median session date*, to examine where student work falls with the calendar progression of the course. To compute the median session date, we took the date of each session, quantified as the number of minutes from course start, and calculated the median of these values for each student. Second, we calculated the *fraction of course chapters accessed*, by simply dividing the number of course chapters accessed by a student by the total number of chapters available within the course. This provides a measure of a student's progression through the curriculum of a course. Third, we calculate *time spent per tenth of course,* which we derive by dividing the calendar time of each course into tenths, and then computing the total time spent in each tenth.

---

[2] "Alphabetical event list," http://edx.readthedocs.org/projects/devdata/en/latest/internal_data_formats/event_list.html#event-list

Consider two students who spend equal amounts of total time in a course. A "high-struggle" student would have an earlier median session date, lower fraction of course chapters accessed, and a higher proportion of their time spent per tenth in the earlier tenths of the course. A "low-struggle" student would have a later median session date, a higher fraction of course chapters accessed, and time more equitably distributed across all tenths of the course.

2.3 Data Analytic Plan

We address our first research question, concerning the across-student relationship between certification and session count, using two kinds of analysis. First, we use an exploratory, non-parametric approach that allows us to investigate the relationships among total time, session count, and certification without assumptions about functional form. We divide each student in each course into deciles of similar time on site, then plot the relationship between groups of students with high, medium, and low levels of session count. We confirm the insights from our non-parametric approach using a bootstrapped analysis, where we divide students into even finer bines (centiles), and then reassign certification labels (0 or 1) within each bin via bootstrap resampling with replacement. This maintains the relationship between total time and certification while removing the relationships of certification with all else. We then test whether the observed data deviates from the null distribution (see Figure 3 of Stafford & Dewar, 2014, for an analogous use of bootstrap).

These non-parametric approaches avoid assumptions about functional form, provide useful visual presentations of data, and allow for statistical testing of the relationship of session count to certification controlling for functional form. As disadvantages, however, they impose arbitrary bins on data and do not easily allow for the analysis of additional potential confounds. Therefore, we also address this first research question with a logistical regression analysis, where we predict the effect of session count on certification controlling for the effects of total time and our proxy measures for struggle.

To address our second research question about the within-student relationship between spaced study and certification, we examine a subset of students who enrolled in two courses, spent similar amounts of time in each course, and spaced their time differently between the two courses. We examine these relationships using an exploratory, non-parametric approach similar to that of our first research question.
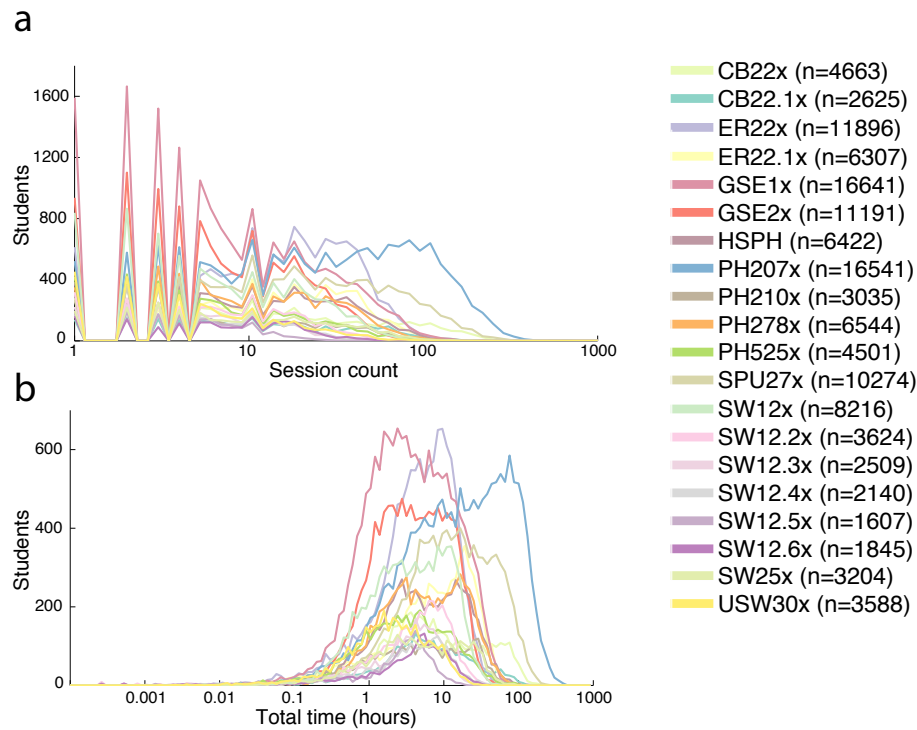
3 RESULTS

3.1 Exploratory Analysis of Total Time, Session Count, and Certification

Across all 20 courses, we found widely varying levels of certification rate, total time, and session count. The certification rates of students with non-zero grades in each course vary widely, from 11% to 70% (Table 1). In Figures 2a and 2b, we plot histograms of the number of students by log-scale session count and total course hours, respectively. These histograms reveal that total time and session count of

54

courses vary widely, with the median total time level of courses varying from as little as 2 hours (USW30x) to about 17 hours (PH207x), an over 8-fold difference, and median sessions from 5 sessions (SW12.5x) to 26 sessions (PH207x), an over 5-fold difference.
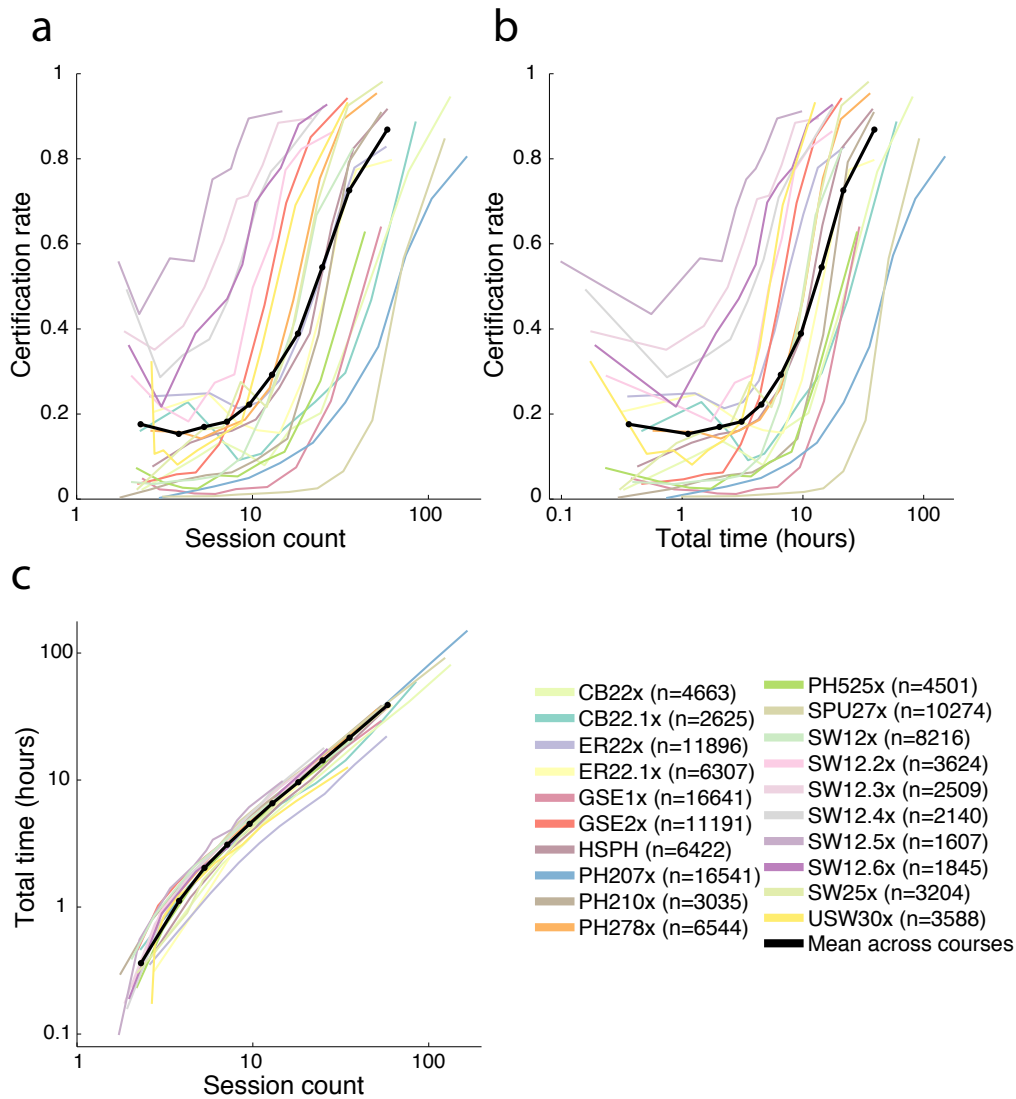

Figure 2. Exploratory analysis of total time, spacing, and certification rates in 20 MOOCs.(a) Histograms show students' session counts, with each colour line representing students from a different course, as detailed in the legend to the right. Note the log-scale x-axis, and also that the spikes at the low spacing levels are because the number of sessions can only take integer values. (b) Histograms show students' levels of total time for each of the 20 courses

A further look at the histograms of Figures 2a and 2b also reveals widely varying levels of total time and number of sessions across students within each course — for example, total time of individual students across all courses varied from as little as a few minutes to as much as over 100 hours, and session counts from 1 to over 100.

We next plotted total time and session count against certification rate. In Figure 3a, we divided students within each course into deciles based on their session counts from low to high, and computed each decile's certification rate to obtain each of the coloured lines in Figure 3a, with colours representing separate courses. The bold black line corresponds to the mean relationship between session count and certification rate across courses, which was obtained by averaging the session count and certification rate of each the deciles across the 20 courses. These relationships appear roughly sigmoidal, with small

effects on certification in small session counts of less than 10 sessions, followed by a sharp increase and later a gradual plateauing as the certification rate approaches its maximum of 1 (note the log-scale of the x-axis).

Figure 3. Exploratory analysis of total time, session count, and certification rate in MOOCs. (a) Certification rates are plotted as a function of session count. Students from each course were divided equally into deciles of session counts, and the average certification rate for each decile is plotted as a function of session count for that decile, to form the coloured lines, one colour for each course. The mean certification rates and session counts were then averaged across all 20 courses for each decile to obtain the bold black line, representing the mean relationship between certification rate, and session count across courses. (b) Certification rates are plotted as a function of total time, after dividing students into deciles of total time in an analogous fashion to session count in (a). (c) Total time is plotted as a function of session count after dividing students into deciles of session counts, as in (a).

When we analyzed students' total time versus certification in the same manner as with session count (Figure 3b), here too we found a strong positive correlation with certification rate, consistent with expectations from the MOOC literature (Collins, 2013; DeBoer et al., 2014; Murphy et al., 2014; Reich et al., 2014; Wilkowski et al., 2014). This relationship appeared to be very similar to that of session count from Figure 3a, where certification rates increased in a sigmoidal fashion also with respect to total time. Notice the dip that occurs in certification rate in some courses from the lowest decile to the next, for both session count and total time; this may reflect the behaviour of some students using a secondary account to earn a certificate in a single session, after initial exposure to test questions on a primary account (a behaviour that in certain contexts might be described as cheating).

Our analyses thus far of Figure 3a and 3b suggest a positive relationship of certification, our proxy for learning, with session count and with total time. However, Figure 3c, which plots total time versus session count, depicts one challenge of further interrogating these relationships. We found that total time and session count themselves were highly correlated with each other (Figure 3c), with a correlation coefficient of 0.87 ± 0.04 (mean $\pm$ s.d. across courses), indicating, as one might expect, that students who work for many sessions also tend to spend a lot of time doing so overall. This collinearity of our features makes it difficult to determine whether students' certification rates are driven by total time, by session count, or both.

3.2 Examining the Effect of Spaced Study via a Non-Parametric Analysis of Student-to-Student Comparisons

To explore the effect of session count on certification rate while controlling for the effect of total time, we divided students into deciles of total time and compared students with different session counts within each decile. Within each decile of total time, we formed terciles of low-, mid-, and high-spacing (session count) groups. Figure 4a, which shows the results of this analysis, provides compelling visual evidence for the benefit of the spaced study sessions in MOOCs, showing that the high-spacing subgroup (black) at each level of total time consistently exhibited higher certification rates than their corresponding low- and mid-spacing subgroup counterparts at each decile of total time (light gray and dark gray dots).
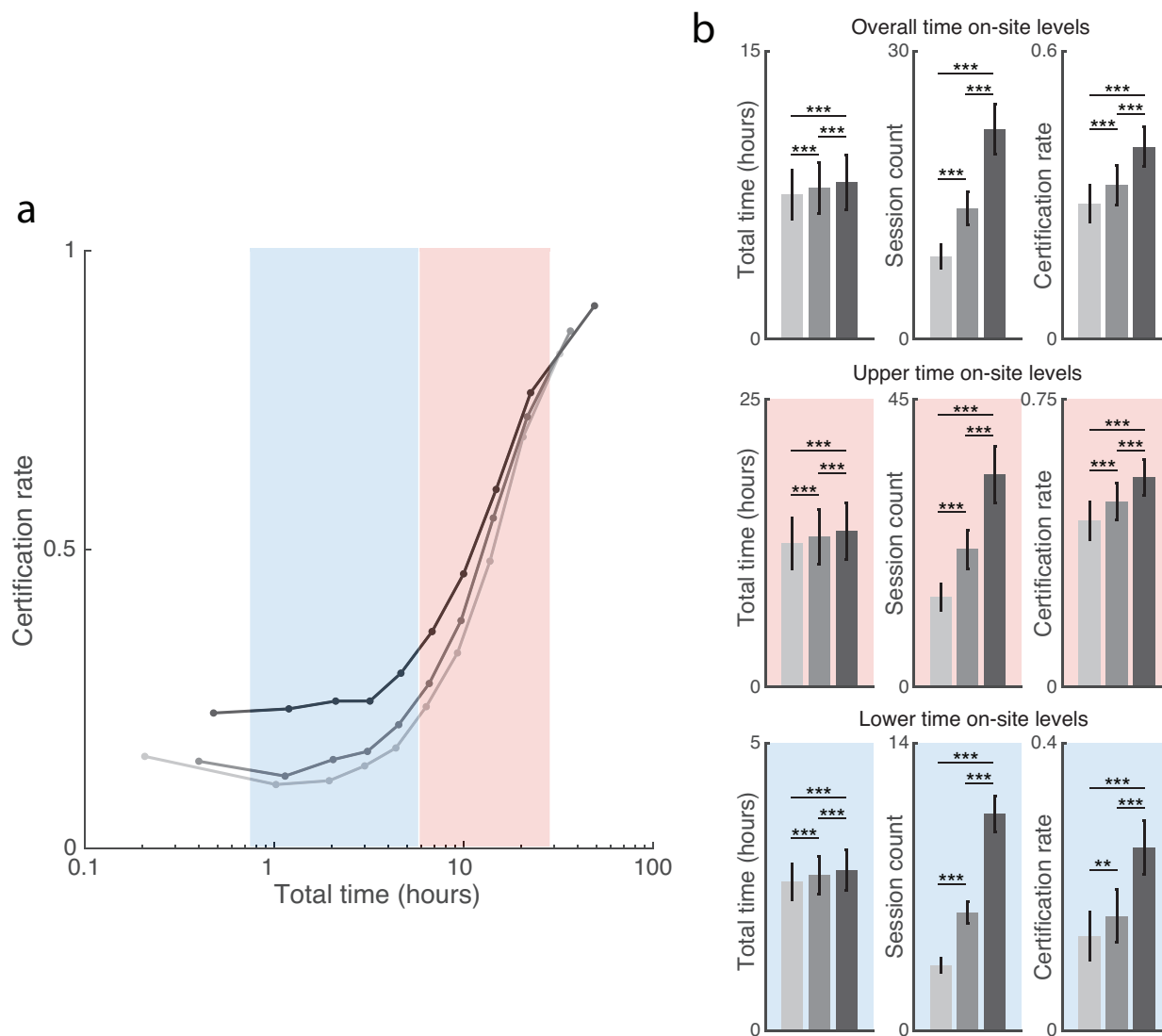
Figure 4. Across-student analysis of the effect of spaced study in MOOCs. (a) Certification rates are plotted as a function of total time (10 levels) and spacing (3 levels) to form 10 x 3 = 30 total groups of students taken from the 20 HarvardX courses in the analysis. First, students within each course were divided equally into deciles of total time, and then students of each decile were divided further into terciles of low-, mid-, and high-spacing groups (light, mid, and dark gray dots). The certification rates and total time levels of the 30 groups were then averaged across the 20 courses to form what is shown here. (b) The top bar graphs show the average total time, session count, and certification rate aggregated across the overall deciles of total time for low- (light gray), mid- (mid gray) and high-spacing (dark gray) groups (excluding the lowest and highest deciles, whose spacing groups are not as well matched in their total time). The middle (shaded red) and bottom (shaded blue) bar graphs are analogous plots aggregating across only upper levels of total time (corresponding to the red region of (a)) and only lower levels of total time (corresponding to the blue region of (a)). Error bars indicate standard errors of the mean, across the 20 courses. The large error bars reflect the large variability across courses noted in Figure 2, but the statistical comparisons are highly significant because comparisons are evaluated in a paired fashion within-course. *: p < 0.05, **: p < 0.01, ***: p < 0.001

The bar graphs in Figure 4b summarize the relationships between spacing and certification at the various deciles of total time. The bar graphs of the top row summarize the total time, session count, and certification rate aggregated across both upper and lower deciles of total time. The light, medium and dark grey bars correspond to the average of the low-, mid-, and high-spacing groups. By design, these low-, mid-, and high-spacing groups have similar levels of total time (7.5 vs. 7.9 vs. 8.2 hours, Figure 4b, left panel of top row) and highly contrasting session counts (8.6 vs. 13.6 vs. 21.9 sessions, Figure 4b, middle panel of the top row). The top-rightmost bar graph shows that certification rates were significantly higher for the high-spacing (0.40) than for the mid-spacing group (0.32, Figure 4b, top right, paired t-test, $p < 0.001$), and both of these groups had rates significantly higher than that of the low spacing group (0.28, Figure 4b, top right, paired t-test, $p < 0.001$). Comparing the high- and low-spacing groups within these high levels of total time, we find that an increase of 13.3 sessions corresponds to about a 71% increase in certification odds (Low-spacing: $0.28 / (1 − 0.28) ≈ 0.39$ vs. High-spacing: $0.40 / (1 − 0.40) ≈ 0.67$, a 71% increase). This translates to an effect size whereby every additional session initiated corresponds to a roughly 4% multiplicative increase in certification odds ($\log (0.67 / 0.39) / 13.3 ≈ 0.04$).

The curves of Figure 4a also indicate that the strength of the effect of session count may be largest at low levels of total time (light blue region) rather than for high level of total time (light red region). The bar graphs of the middle and bottom rows of Figure 4b summarize the upper levels of total time (middle row of Figure 4b, shaded red) separately from the lower levels of total time (bottom row of Figure 4b, shaded blue). The difference in certification rate between spacing groups is largest at the lower deciles of total time, suggesting that the potential benefit of spaced study may be most impactful at these lower levels.

3.3 Non-Parametric Bootstrap Analysis of Student-to-Student Comparisons

To test our non-parametric analysis, we compared the observed data to a bootstrapped null distribution that we created by dividing up the data into fine bins (centiles) based on total time, and then reassigning the certification labels within each bin via bootstrap resampling with replacement. This process preserves the relationship between total time and certification, but eliminates the relationship between session count and certification beyond what can be predicted through total time. This is because the bootstrapping procedure preserves the conditional dependence of certification on the bin of total time and also preserves the relationship between session count and total time, but removes the relationships of certification with all else. Therefore, observing data that deviates from this null distribution would indicate that session count predicts certification at a level beyond what would be expected through its relationship with total time (see Figure 3 of Stafford & Dewar, 2014, for an analogous use of bootstrap).

Figure 5 shows the results of applying this bootstrap analysis. The gray shaded regions show the 99% confidence envelopes of the bootstrapped null relationship between certification and total time (Figure 5a) and certification and session count (Figure 5b), while the blue trace shows the observed relationship

in the data. These envelopes were calculated by simulating 9999 null datasets, and then drawing the confidence bounds at each point along the trace large enough that only 1% of the datasets (from 10,000 datasets, including the observed dataset) exceeded the bounds at any point along the overall trace (thus adjusting for multiple comparisons). To provide a closer and more sensitive look at how the observed data deviates from the null, we de-trended the observed data curves and null confidence envelopes by subtracting off the mean of the null relationship (Figure 5c, d). Figure 5c shows that the observed relationship between certification and total time is indistinguishable from the null distribution, as intended by design. In sharp contrast, the relationship between certification and session count is well beyond that which would be expected from the null distribution, in favour of an alternative hypothesis in which session count has a positive effect on certification.
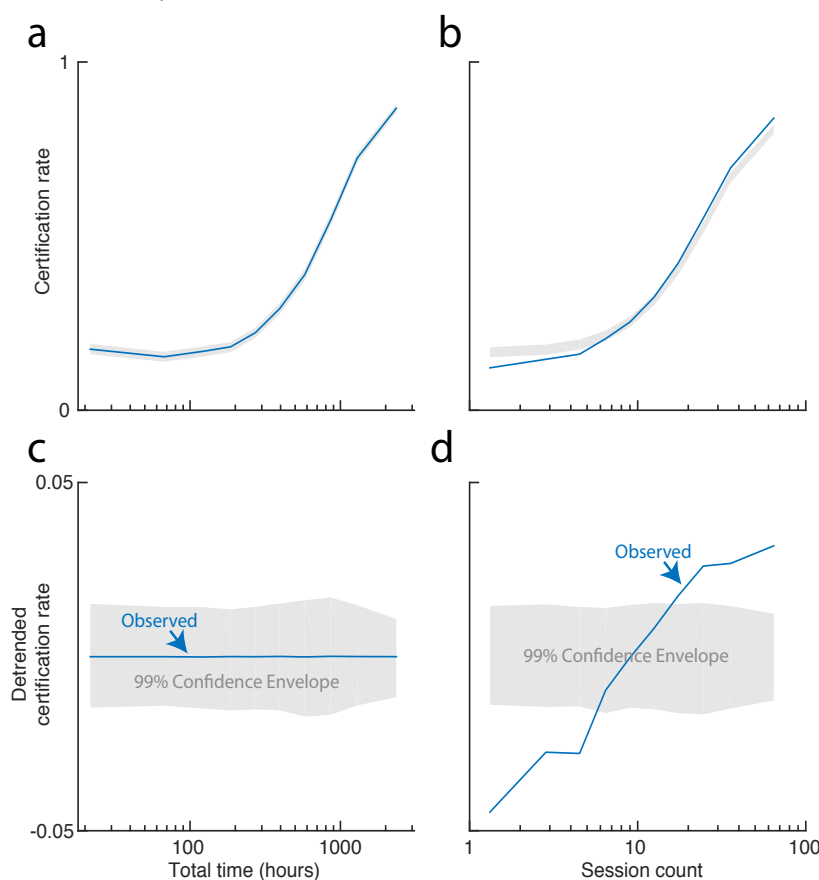
Figure 5. The bootstrap hypothesis test validates the effect of session count on certification while accounting for total time. (a) Relationship between certification rate and total time for the observed data (blue) and for the bootstrapped null distribution (gray: 99% confidence envelopes); (b) Relationship between certification rate and session count for the observed data and for the bootstrapped null distribution, analogous to (a). (c) and (d) convey the same information as (a) and (b) except that the observed data curves and null confidence envelopes are detrended by subtracting away the mean null prediction to more sensitively verify the following: the observed relationship between certification and session count deviates considerably from the bootstrapped null distribution, which preserves the relationship between certification and total time.

3.4   Logistic Regression Analysis of Student-to-Student Comparisons

To avoid arbitrary binning of data and to allow for the analysis of additional control predictors, we complemented our previous analyses with a logistic regression analysis. Along with session count and total time, we explored the main effects of three proxies for struggle: median session date, fraction of chapters accessed, and time spent per tenth of course.

Table 2 displays a taxonomy of fitted logistic regression models predicting certification. This regression analysis used the z-scores of the features computed within each course in order to achieve comparability between different courses, which could have different distributions of session count or total time. Models A–D examine the effects of total time and session count on certification controlling for course effects, and then Models E–H include our proxies for struggle. In line with the non-parametric and bootstrapped analyses, we found that the effect of session count has a statistically significant positive association with certification rate while accounting for the effect of total time (Table 2, model D). Given that one unit of the session count z-score corresponds roughly to 20 sessions (the average standard deviation across courses is 20 sessions), the 0.67 logit coefficient of model D corresponds to a 3.4% increase in certification odds for every additional session initiated (exp(0.67 * 1/20) ≈ 1.034). That this effect size closely matches that of the non-parametric analysis (4.0% increase in certification odds per unit session increase) suggests internal consistency of these analyses.

Furthermore, we consistently see a significantly positive effect for session count even when accounting for the possible confounds of median session date, fraction of chapters accessed, and time spent per tenth of course (model E through H). The effect of session count on certification remains strong, indicating that its effect is not driven by the potential confounds that we explored here. Other confounds may exist, given the observational nature of this study.

Table 2. Taxonomy of logistic regression models for student-to-student analysis. Values represent the logit coefficients of the regression models. Note that an intercept, not shown, was included in all models. *: p < 0.05, **: p < 0.01, ***: p < 0.001

	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
Total time	1.49***		0.96***	1.10***	0.91***	1.31***	0.77***	0.93***
Session count		1.45***	0.61***	0.67***	0.73***	0.72***	0.79***	0.78***
Random course effects	No	No	No	Yes	Yes	Yes	Yes	Yes
Centroid of session dates					0.69***			0.56***
Fraction of chapters accessed						-0.35***		-0.13***
Time spent during 1st tenth							-8.4e-4***	-3.3e-4***
Time spent during 2nd tenth							-6.9e-4***	-5.1e-4***

Time spent during 3rd tenth	-4.7e-4***	9.5e-5
Time spent during 4th tenth	-1.8e-5	-2.1e-4**
Time spent during 5th tenth	2.1e-4**	-7.3e-5
Time spent during 6th tenth	3.4e-4***	-1.4e-5
Time spent during 7th tenth	9.1e-4***	1.9e-4*
Time spent during 8th tenth	1.3e-3***	6.3e-4***
Time spent during 9th tenth	1.9e-3***	7.3e-4***
Time spent during final tenth	3.3e-3***	1.1e-3***

3.5 Identifying a Benefit of Spaced Study via Within-Student Comparisons

We found that students who distribute their time into a larger number of sessions had higher levels of certification than students who spent similar total time divided into fewer sessions, and that this effect remained whilst accounting for other features of student behaviour such as median session date, fraction of chapters accessed, and time spent per tenth of course. However, these correlational observations could be confounded by other unobserved variables: it could simply be that high-performing students also tend to have large numbers of sessions, but the *act* of distributing time into a larger number of sessions may not be a causal factor in better performance. To investigate this question, we analyzed our data more finely to examine whether the effect of sessions can be observed within individuals.

We examined whether individual students with different levels of session counts across different courses show higher certification rates in courses where they space out their time more. Here we take advantage of our data set of over 100,000 total students to examine the subset of students enrolled in multiple courses. Looking within the overlapping students of a particular pair of courses, some may have distributed their time into more sessions in one than the other, while maintaining their level of total time. Although this may happen rarely, the massive scale of the data set gives us the ability to observe a sizeable number of such events. Thus we can examine if a student's change in session count is associated with a corresponding change in performance.

In this analysis, we focused on students who maintained similar levels of total time in relation to their peers within each pair of courses, so that we could look at the effect of spacing while controlling for levels of total time. We only included students who remained within the same decile of total time in both courses to minimize changes in levels of total time across pairs of courses. Because this method substantially reduces the number of students we can include in our analysis, we only considered pairs of courses with at least 250 overlapping students before filtering students based on their levels of total time to maintain some level of reliability for examining each course-pair. This threshold allowed us to

consider 45 out of the 190 possible pair-wise comparisons that could be made across courses. We further eliminated redundant comparisons that might occur when students were enrolled in more than two courses to avoid double-counting student comparisons; such comparisons were eliminated in reverse chronological order such that the comparisons from the earlier courses were retained. As a result, 2554 out of the over 100,000 students were included in this within-student analysis.
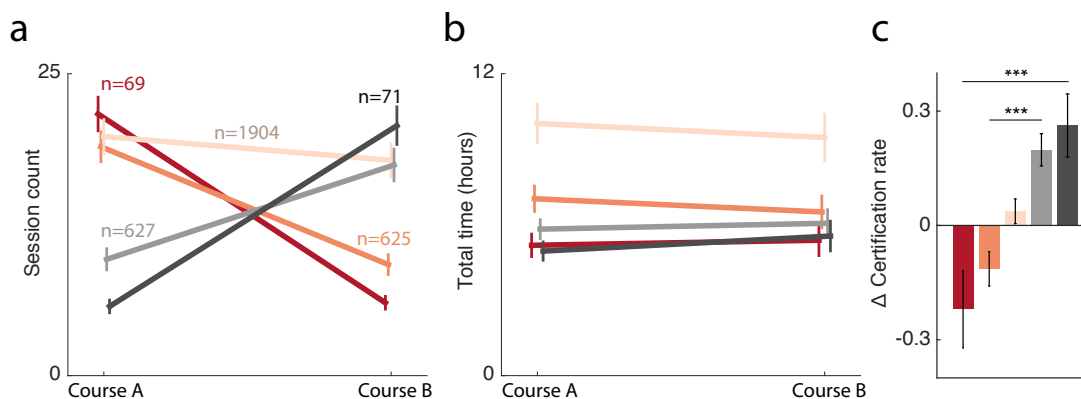
Figure 6. Within-student analysis of the effect of spaced study in MOOCs. (a) Session counts are shown for 5 different variations of changes from one course to the other within a course-pair. Dark gray represents students who greatly increased their session count from low to high terciles between courses. Terciles were computed within each course, to account for course-to-course differences in session counts. Light gray represents those who mildly increased their session count from low to mid, or mid to high terciles between the pair of courses. Red and dark orange represent students who greatly and mildly decreased their session count, respectively, in an analogous manner. Light orange represents those who remained within the same tercile for both courses. (b) The mean total time levels are shown corresponding to the 5 different variations of session count changes in (a). By construction, these levels of total time were largely maintained between the courses within each course-pair, in contrast to the session counts of (a). (c) Certification rates are shown for the 5 different variations of session count changes corresponding to the colours of (a) and (b). Error bars indicate the standard error of the mean computed across all course-pairs. ***: p < 0.001

As in the analysis of Figure 4, we divided students of each course into 3 groups — low-, mid-, and high-session groups, corresponding to the low, mid, and high terciles of session counts relative to their peers within each course. Terciles were computed within each course separately, to account for course-to-course differences in session counts. We then compared the changes in certification rate between courses for different paths traversed among these terciles, resulting in 5 different paths all shown in Figure 6a: 1) those who greatly decreased from high session counts to low, shown as red, 2), those who mildly decreased from high to mid, or mid to low session counts, shown as dark orange, 3) those who largely maintained their session counts within the same tercile, shown as pale orange, 4) those who mildly increased their session counts from low to mid, or mid to high, shown as light gray, and finally 5)

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

63

those who greatly increased their session counts from low to high, shown as dark gray. All groups were computed separately for each course-pair, and then averaged together across course-pairs to obtain the total time levels, session counts, and delta certification rates in Figure 6. Although these students on average could have big changes in session count (Figure 6a), they largely maintained their level of total time between the pair of courses (Figure 6b), as we intended through the construction of these groups. Figure 6c shows the changes in certification rate for these five groups, providing further evidence for the benefits for spaced study sessions in MOOCs. The progression of these 5 paths from greatly decreased (red) to greatly increased (dark grey) session counts corresponds to a monotonically increasing progression in certification rate, with a strong correlation between them of 0.99 ($p < 0.001$) for these 5 points, in contrast to the substantially weaker correlation ($r = 0.48$, $p = 0.41$) between total time and session count (as intended by the design of this analysis, which attempts to isolate the effect of session count). Perhaps most striking is the difference in change of certification rate between the greatly decreasing (red) and greatly increasing (dark gray) session count paths (Figure 6c, $p < 0.001$), with a difference between the changes of almost 0.5. A similar comparison can be made for the mildly decreasing and mildly increasing paths in light orange and light grey ($p < 0.001$), with a difference in certification rate of over 0.2. Taken together, these results suggest a benefit of the act of spacing study time in MOOCs, measured within the behavioural changes of individual students.

4   DISCUSSION

Motivated by the psychology literature on the spacing effect, we examined student behaviour in HarvardX with the hypothesis that for a given level of study time, a higher number of study sessions would be associated with higher certification rates. In agreement with this hypothesis, our analysis comparing different students within courses revealed that those with higher levels of spacing had a higher chance of achieving certification than those who did not, while controlling for their total time spent. Interestingly, we found that this benefit may be greater for students at low total time levels. Furthermore, by examining students enrolled in multiple courses, we discovered that students who increased their spacing levels from one course to another while maintaining their levels of total time demonstrated a corresponding benefit in their tendency to achieve certification. These findings arose from analyses of a variety of courses, and our metrics — total time and session count — were agnostic to the specific course activities, content, and structure. Taken together, these results strongly suggest benefits to distributing study time into a larger number of sessions, which may have flexible applications for improving student performance in MOOCs. We also hope that this study encourages MOOC researchers to go beyond simple correlations of activity metrics to outcomes (since virtually any action is positively correlated with other actions and performance), and instead examine more closely better and worse ways for students to invest their scarce time.

4.1 Differences Between the Current Results and Spacing Effects Studied in Psychology

Our findings cohere with previous research on spaced practice, but there are important differences in our study and previous literature. While our observational study allows us to examine for evidence of spaced study sessions "in the wild" in diverse settings without parameters controlled or affected by experimentalists, our research designs cannot prove that the effect of spacing of the current findings is directly due to the conventional spacing effect that is recognized by the psychology literature. The spacing effect, as typically recognized, is the benefit in long-term retention following the spaced presentations of particular items of knowledge. However, it is unclear to what extent certification in a course reflects long-term retention as opposed to short-term retention, for example, since student activity is largely self-scheduled; in fact, massed presentation has sometimes been shown to be more beneficial for short-term retention (Peterson, Hillner, & Saltzman, 1962; Peterson, Saltzman, Hillner, & Land, 1962). Our outcome measure, certification, is less precisely defined than long-term retention, but may represent more holistic dimensions of achievement and learning.

The effect of spacing identified in the current study, therefore, likely reflects a separate mechanism for success in a learning environment. For instance, the spacing effect identified here might be more related to motivation, rather than retrieval of memories per se. It may be that students can stay more excited about learning when their interactions with MOOCs are spaced out. Regardless of the underlying nature of this spacing effect, however, the findings from this study cohere with previous findings on spaced practice, indicate practical advantages to spacing out total time in MOOCs, and motivate practical steps toward leveraging these benefits.

4.2 Interventions for Applying Spaced Practice and Spaced Study Sessions to MOOCs

Recent years have seen a growing demand for ways to apply the benefits of spacing into practice, in light of criticism of the disconnect between psychological research and educational practice, perhaps best described by the title of an article by Frank Dempster in the *American Psychologist* (1988), "The spacing effect: A case study in the failure to apply the results of psychological research." Approaches that have recently gained in popularity for applying the spacing effect, however — such as through flashcard systems of spaced repetition software — may not be an ideal one to apply to the wide variety of course structures and course topics present in MOOCs. Topics can range from computer science to philosophy, to guitar learning and entrepreneurship, whose content might be neither easily nor appropriately converted to a sequence of flashcards by students. Another recent approach, by a start-up company named SpacedEd (Lambert, 2009), has been to call upon instructors to design courses specifically in a spacing-friendly format of a list of questions and answers, to offer online education that is designed around the spacing effect. These approaches both require a costly overhead in which either the students or the instructors must figure out a way to mold the curriculum into a spacing-friendly

format. Directly applying this spacing effect to MOOCs could require redesigning of course content and structure.

The present results suggest the potential benefit of designing and applying relatively simple and inexpensive interventions to students taking MOOCs that encourage them to distribute their time into a larger number of shorter sessions, rather than a small number of long sessions. Possible effective interventions might include, for example, incentives for more frequent sessions, such as a daily login reward, or perhaps the sending of reminders for logging in, as well as reminders to take breaks in the middle of a session. One useful idea for motivating students to increase the number of their sessions might be the division of assignments into smaller modules that appear more frequently in time. For instance, many courses release their content weekly or bi-weekly, but courses might consider releasing certain elements mid-week, or even daily. Course developers might divide papers and projects into smaller sub-assignments, release course materials in smaller increments multiple times a week rather than a large release once a week, or assign smaller and more frequent homework and quizzes rather than lengthy midterm and final exams.

Such experimental interventions could validate the causal relationship between spaced study sessions in MOOCs and certification. The most useful experimental designs will test competing theories of mechanisms for the effects of spaced study in MOOCs, and these experiments will help clarify whether the value of spacing study sessions in MOOCs has any conceptual link to the value of spacing practice for memorizing vocabulary, beyond the general finding that some uses of time are better than others. No doubt we will see new advances in open online learning and perhaps entirely new generations of learning technologies. The findings here serve as an important reminder that well-established learning theory can be put to the service of new technologies. We have over a century of research on learning theories related to spaced practice. Course and platform developers should attend to findings over the last century of educational research. The field of learning analytics, too, can be greatly enriched when we turn to learning theory to identify what is worth tracking, investigating, and analyzing.

REFERENCES

Benassi, V., Overson C., & Hakala, C. (Eds.). (2014). *Applying science of learning in education: Infusing psychological science into the curriculum*. Washington, DC: American Psychological Association. Retrieved from http://teachpsych.org/ebooks/asle2014/index.php

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics*, *31*(4), 635–650. http://dx.doi.org/10.1017/S0142716410000172

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of US history facts. *Applied Cognitive Psychology, 23*(6), 760–771. http://dx.doi.org/10.1002/acp.1507

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulleti*, *132*(3), 354–380. Retrieved from http://escholarship.org/uc/item/3rr6q10c

Clark, R. C., & Mayer, R. E. (2011). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. Hoboken, NJ: John Wiley & Sons.

Collins, E. D. (2013). *SJSU plus augmented online learning environment* (Pilot project report). Retrieved from http://www.sjsu.edu/chemistry/People/Faculty/Collins_Research_Page/AOLE Report Final Version_Jan 1_2014.pdf

Cooper, E. H., & Pantle, A. J. (1967). The total-time hypothesis in verbal learning. *Psychological Bulletin*, *68*(4), 221–234. http://dx.doi.org/10.1037/h0025052

DeBoer, J., Ho, A. D., Stump, G. S., & Breslow, L. (2014). Changing "course": Reconceptualizing educational variables for massive open online courses. *Educational Researcher, 43*(2), 74–84. http://dx.doi.org/10.3102/0013189X14523038

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*(8), 627–634. doi:10.1037/0003-066X.43.8.627

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review*, *1*(4), 309–330. http://dx.doi.org/10.1007/BF01320097

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. http://dx.doi.org/10.1177/1529100612453266

Ebbinghaus, H. (1885). *Memory* (H. A. Ruger & C. E. Bussenius, Trans.). New York: Teachers College Press.

Gallo, M. A., & Odu, M. (2009). Examining the relationship between class scheduling and student achievement in college algebra. *Community College Review*, *36*(4), 299–325. http://dx.doi.org/10.1177/0091552108330902

Godwin-Jones, R. (2010). Emerging technologies from memory palaces to spacing algorithms: Approaches to second language vocabulary learning. *Language, Learning & Technology*, *14*(2). Retrieved from http://lsa-cmsf5test.lsa.umich.edu/german/hmr/531/llt/emerging.pdf

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(3), 371–377. http://dx.doi.org/10.1037/0278-7393.15.3.371

Hintzman, D. L. & Solso, R. L. (1974). Theoretical implications of the spacing effect. *Theories in Cognitive Psychology: The Loyola Symposium*, 77–99

Ho, A., Reich, J., Nesterko, S., Seaton, D., Mullaney, T., Waldo, J., & Chuang, I. (2014). *HarvardX and MITx: The first year of open online courses* (HarvardX and MITx Working Paper No. 1). Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2381263

Kang, S. H., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: Should spacing be expanding or equal-interval? *Psychonomic Bulletin & Review*, *21*(6), 1544–1550. http://dx.doi.org/10.3758/s13423-014-0636-z

Khajah, M. M., Lindsey, R. V., & Mozer, M. C. (2014). Maximizing students' retention via spaced review: Practical guidance from computational models of memory. *Topics in Cognitive Science*, *6*(1), 157–169. http://dx.doi.org/10.1111/tops.12077

Lambert, C. (2009, November). Learning by degrees. *Harvard Magazine*, 11. Retrieved from http://harvardmagazine.com/2009/11/spaced-education-boosts-learning

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological Science*, *25*(3), 639–647. http://dx.doi.org/10.1177/0956797613504302

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 596–606. doi:10.1016/S0022-5371(70)80107-4

Moulton, C.-A. E., Dubrowski, A., MacRae, H., Graham, B., Grober, E., & Reznick, R. (2006). Teaching surgical skills: What kind of practice makes perfect?: A randomized, controlled trial. *Annals of Surgery*, *244*(3), 400–409. http://dx.doi.org/10.1097/01.sla.0000234808.85789.6a

Murphy, R., Gallagher, L., Krumm, A., Mislevy, J., & Hafter, A. (2014). *Research on the use of Khan Academy in schools*: *Research Brief.* Menlo Park, CA: SRI Education. Retrieved June 20, 2014 from http://www.sri.com/sites/default/files/publications/2014-03-07_implementation_briefing.pdf

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning* (IES practice guide*). Retrieved form the website of the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education at http://ies.ed.gov/ncee/wwc/pdf/practice_guides/20072004.pdf

Peterson, L. R., Hillner, K., & Saltzman, D. (1962). Supplementary report: Time between pairings and short-term retention. *Journal of Experimental Psychology*, *64*(5), 550–551. http://dx.doi.org/10.1037/h0046079

Peterson, L. R., Saltzman, D., Hillner, K., & Land, V. (1962). Recency and frequency in paired-associate learning. *Journal of Experimental Psychology*, *63*(4), 396–403. http://dx.doi.org/10.1037/h0043571

Peterson, L. R., Wampler, R., Kirkpatrick, M., & Saltzman, D. (1963). Effect of spacing presentations on retention of a paired associate over short intervals. *Journal of Experimental Psychology*, *66*(2), 206–209. http://dx.doi.org/10.1037/h0046694

Rau, M. A., Aleven, V., Rummel, N., & Pardos, Z. (2014). How should intelligent tutoring systems sequence multiple graphical representations of fractions? A multi-methods study. *International Journal of Artificial Intelligence in Education*, *24*(2), 125–161. http://dx.doi.org/10.1007/s40593-013-0011-7

Reich, B. J. F., Emanuel, J. P., Nesterko, S. O., Seaton, D. T., Mullaney, T. P., Waldo, J. H., … Ho, A. D. (2014). HeroesX: The ancient Greek hero (*Spring 2013 Course Report*). Retrieved from http://dash.harvard.edu/handle/1/11988100

Rohrer, D. (2009). The effects of spacing and mixing practice problems. *Journal for Research in Mathematics Education, 40*, 4–17.

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practise on the retention of mathematics knowledge. *Applied Cognitive Psychology*, *20*(9), 1209–1224. http://dx.doi.org/10.1002/acp.1266

Shea, C. H., Lai, Q., Black, C., & Park, J.-H. (2000). Spacing practice sessions across days benefits the learning of motor skills. *Human Movement Science*, *19*(5), 737–760. http://dx.doi.org/10.1016/S0167-9457(00)00021-X

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*(5), 763–767. http://dx.doi.org/10.1002/acp.1747

Stafford, T., & Dewar, M. (2014). Tracing the trajectory of skill learning with a very large sample of online game players. *Psychological Science*, *25*(2), 511–518.

Underwood, B. J. (1970). A breakdown of the total-time law in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 573–580. http://dx.doi.org/10.1016/S0022-5371(70)80104-9

Waugh, N. C. (1970). On the effective duration of a repeated word. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 587–595. http://dx.doi.org/10.1016/S0022-5371(70)80106-2

Wilkowski, J., Deutsch, A., & Russell, D. M. (2014). Student skill and goal achievement in the mapping with Google MOOC. *Proceedings of the First ACM Conference on Learning @ Scale Conference,* 3–10. http://dx.doi.org/10.1145/2556325.2566240

Williams, J. J. (2013). Improving learning in MOOCs with cognitive science. In Z. A. Pardos, E. Schneider (Eds.) *AIED 2013 Workshops Proceedings of the Workshop on Massive Open Online Courses (Moocshop)* (Vol.1, pp. 49–54). Retrieved from http://ceur-ws.org/Vol-1009/aied2013ws_volume1.pdf

Young, J. L. (1966). *Effects of intervals between reinforcements and test trials in paired-associate learning (*Technical Report No. 101, Psychology Series). Stanford, CA: Institute of Mathematical Studies in the Social Sciences.