Evaluating Retrieval Practice in a MOOC:

How writing and reading summaries of videos affects student learning

Tim van der Zee
Graduate School of Teaching (ICLON), Leiden University
Leiden, Netherlands
t.van.der.zee@iclon.leidenuniv.nl

Dan Davis
Web Information Systems, Delft University of Technology
Delft, Netherlands

Nadira Saab
Graduate School of Teaching (ICLON), Leiden University
Leiden, Netherlands

Bas Giesbers
Rotterdam School of Management, Erasmus University Rotterdam
Rotterdam, Netherlands

Jasper Ginn
Online Learning Lab, Leiden University
Leiden, Netherlands

Frans van der Sluis
Online Learning Lab, Leiden University
Leiden, Netherlands

Fred Paas
Early Start Research Institute, University of Wollongong;
Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam
Rotterdam, Netherlands

Wilfried Admiraal
Graduate School of Teaching (ICLON), Leiden University
Leiden, Netherlands

ABSTRACT

Videos are often the core content in open online education, such as in Massive Open Online Courses (MOOCs). Students spend most of their time in a MOOC on watching educational videos. However, merely watching a video is a relatively passive learning activity. To increase the educational benefits of online videos, students could benefit from more actively interacting with the to-be-learned material. In this paper two studies (n = 13k) are presented which examined the educational benefits of two more active learning strategies: 1) Retrieval Practice tasks which asked students to shortly summarize the content of videos, and 2) Given Summary tasks in which the students were asked to read pre-written summaries of videos. Writing, as well as reading summaries of videos had a positive impact on quiz grades. Both interventions helped students to perform better, but there was no difference between the efficacy of these interventions. These studies show how the quality of online education can be improved by adapting course design to established approaches from the learning sciences.

CCS CONCEPTS

• **Applied computing** → **E-learning**; *Distance learning*; • **Human-centered computing** → Empirical studies in HCI;

LAK '18, March 7–9, 2018, Sydney, NSW, Australia

ACM ISBN 978-1-4503-6400-3/18/03…$15.00
https://doi.org/10.1145/3170358.3170382

ACM Reference format:
Tim van der Zee, Dan Davis, Nadira Saab, Bas Giesbers, Jasper Ginn, Frans van der Sluis, Fred Paas, and Wilfried Admiraal. 2018. Evaluating Retrieval Practice in a MOOC:. In *Proceedings of International Conference on Learning Analytics and Knowledge, Sydney, NSW, Australia, March 7–9, 2018 (LAK '18),* 10 pages.
https://doi.org/10.1145/3170358.3170382

1 INTRODUCTION

Online education has become strongly entrenched in the educational landscape, with more than one out of four students taking an online course [2]. Online courses which are freely accessible, such as Massive Open Online Courses (MOOCs), have opened up education. With an unprecedented scale, MOOCs are reaching millions of students around the world. In addition, open online education has also opened up a new frontier for research on learning, as we can track the behavior of countless students in an uninterrupted manner.

MOOC participants show a wide variety of engagement patterns. While a variety of studies have come up with different ways to cluster these behavior patterns [17, 22, 24, 25, 28, 33], most report three generally similar archetypes of student behavior. First, there is a minority of 'completers', students who watch all or most videos and complete most or all assignments. This group tends to be small, as only around 10% of the students who start a course complete it [27]. Secondly, there is much a larger group of 'auditors' or 'explorers'. These students watch some videos and do some of the readings but less frequently do assessments. While they do show varying levels of activity in a course, they tend not to complete the course nor obtain a certificate. A third group consists of 'disengagers' or 'dropouts'. They are active at the start of a course, but then show a marked decrease in activity and then dropout completely.

The minimal completion rates of MOOCs, as well as low learner satisfaction has been the focus of criticism and doubt about the value

of these courses [23, 30]. Furthermore, the instructional design quality of MOOCs has been criticized for being suboptimal [35], while this is an essential factor which contributes to learners' continued MOOC usage [51]. Similarly, using MOOCs as a platform to study learning processes does not come without its difficulties. The novel research opportunities come with a variety of challenges regarding the validity, generalizability, and evidential value of MOOC research [48]. In light of these challenges we present in this paper two experimental MOOC studies on improving student's learning gains through theoretically informed educational interventions.

1.1 The central role of videos in MOOCs

Educational videos have become increasingly popular in a variety of forms of education, but they are especially fundamental in MOOCs in which they are typically the very core content. For example, while all MOOCs make use of educational videos, only 82% have discussion fora and 69% have teaching guides and background readings [18]. There are more reasons to consider videos the core content of MOOCs. When asked about their intent to engage with videos and the associated assignments, a majority of MOOC students report that they plan to watch all the videos [10]. Other studies report that of all the educational resources available to them, students most often access videos, and spend most of their time (re)watching them [7, 21, 45]. While watching videos is the primary activity of most MOOC students, it is very common for students to not finish watching a video. In an analysis of over 800 videos, it was found that nearly half of the video viewing sessions are cut-off before the end of the video [31]. However, over 2/3 of these dropouts occurred at the very start of the video, suggesting that these students might not have intended to watch the video in the first place. When watching videos, MOOC participants predominantly stream the videos online instead of watching them offline [33]. There appear to be few differences in this video watching behavior between the minority of students who complete a course compared to the majority who only sample a part of a course.

As videos play such a fundamental role in MOOCs, it is essential to better understand which factors influence how students engage with, and learn from, videos. In addition, there is a need for evidence-based interventions which can increase student learning from videos.

1.2 Factors influencing students' video watching behavior

The instructional design of videos has been highlighted as an important quality criterion that drives a successful open online course [53]. A number of studies have investigated which factors correlate with how students engage with a video, such as their dwelling time: how long they spend watching a video. Students who are re-watching a video (as opposed to watching it for the first time) have been found to stop watching the video more often [31]. PowerPoint slide videos tend to have lower dwelling times than 'Khan academy' style videos (tablet drawing tutorials) [21]. Similarly, videos which show the instructor's face have higher dwelling times as opposed to videos that don't [21]. However, while students might engage differently with these different types of videos, this does not necessarily translate to differences in learning gains. For example, multiple studies have reported that the presence or absence of the teacher's face has no impact on how much students learn from a video [32, 50].

In terms of drop-outs, the length of a video seems to correlate with dropout rates, such that more students dropout in longer videos [21, 31]. While some have argued that this means that MOOC videos should be made shorter to reduce dropout rates [21], we are hesitant to make such a causal claim based on correlational data. That is, shorter and longer videos might differ on a wide variety of other instructional design characteristics which might actually be underlying this correlation. For example, in an analysis of over a hundred MOOC videos it was found that the complexity of the transcript explains almost a quarter of the variance in video dwelling time [47]. Importantly, there is a non-linear relationship, with both low-complex and high-complex videos showing an increased dwelling time. As such, in the absence of strong experimental evidence we are hesitant to directly interpret correlations between video characteristics and student behaviors as a straightforward causal relationship.

A different study looked not at complexity of the video transcript, but experimentally manipulated the visual design complexity of multiple MOOC videos [49]. Videos which use a more visually demanding design (e.g., presenting a lot of information simultaneously) reduced the students' ability to learn from the video, as they scored lower on subsequent quizzes. This study is based on a long tradition of multimedia research focused on which design features impact information processing, and as a consequence: how much students learn from a video [36]. This line of research, grounded in cognitive psychology, does not start with a consideration of MOOC characteristics but with those of the human cognitive architecture. When watching a video, students have to process a rich amount of information; working memory acts as a bottleneck for processing and integrating this information [3]. Videos that present too much irrelevant information, or present relevant information in a way that makes it hard to integrate, show substantially lower learning gains [37].

Up to now we have discussed mostly video design factors which affect how students engage with videos. In addition to these in-video factors, there are also extra-video factors which influence learning. Videos are never presented in isolation but are part of a larger curriculum. To further understand how we can improve learning from videos it is important to appreciate *how* videos are being embedded in a course. That is, given that a MOOC learner watches videos (as most do [7, 45]), what kind of learning activity should follow the video to increase how much knowledge students will retain from it? In the following section we will focus specifically on activities and interventions which have been shown to increase how much students learn from educational materials. Specifically we will focus on retrieval practice (learning by remembering information) and writing summaries.

1.3 Retrieval practice

Many online courses make use of in-video questions, weekly quizzes, graded assignments and other types of tests. Often, such tests are used primarily to assess what a learner has learned and/or to assign a grade. However, answering quiz questions is not only useful to get insight into how much a student has learned, it also positively affects learning. This so-called testing effect or retrieval practice has accumulated a large body of evidence for over a century [1, 44].

Ever since, the testing effect has been repeatedly studied and further validated in a variety of settings and with a variety of materials [29, 41, 43]. Retrieval practice does not depend on an aspect of feedback that students might get from answering quiz question, as it is beneficial for learning even when no kind of feedback is given [5, 11]. While the precise mechanism underlying retrieval practice remains unclear, the act of retrieving memory from information seems to strengthen this memory and slow down forgetting. As such, benefits of retrieval practice over other study strategies are typically not visible when knowledge is tested immediately after learning, but it does lead to enhanced long-term retention, such as when students are tested a week after the learning phase [42]. This aligns well with MOOCs, given that these feature weeks of content and are often positioned as being useful to lifelong learning and career development. When compared to other popular study strategies, such as note-taking and rereading texts, retrieval practice typically comes out as the strongest learning strategy [29, 39]. Importantly, the benefits of retrieval practice are not limited to text-based materials, but also extend to video materials [26].

The majority of the literature on retrieval practice are set in the context of laboratory or classroom settings, while only a few focus on MOOC materials or are set in a MOOC context. A previous study which attempted to apply retrieval practice to MOOCs failed to find beneficial effects [16]. In this study, retrieval cues were added after every final video of each course week, just before the weekly quiz. The fact that students who engaged with these cues did not show any benefits is surprising, but there are several plausible explanations. First, it might be the case that retrieval practice simply does not work very well in the context of a MOOC. However, there is evidence that MOOC videos are suitable for retrieval practice, as a recent lab-based study with MOOC videos showed that watching the video once followed by a retrieval test led to better performance than simply rewatching the video [52]. Secondly, the retrieval cues were positioned right before the weekly quizzes, but benefits of retrieval practice are commonly much more pronounced when there is a substantial delay between the retrieval cue and the test [42]. The learn-test delay might have been minimal for many students, which will obfuscate an effect. Thirdly, the retrieval cues were only positioned after a single video, while the weekly quizzes assessed information from all the videos of the week. As such, a potential positive effect of retrieval practice of the single video might not have been strong enough to be visible in the quiz results.

1.4 Writing summaries

One way of having students use retrieval practice is by having them write what they have learned from memory, such as writing a summary of a video they have previously seen. Writing has been previously highlighted as a method of learning [4], especially when using a computer [20]. More specifically, writing an abstract or summary of what has been learned is an effectively strategy for learning with benefits for comprehension, retention, and reading and writing abilities [19]. Similar to retrieval practice, summarizing is an effective learning strategy partly because it requires reconstruction of knowledge [46]. Students who summarize outperform students who use underlining instead - even when the they summarize and underline the exact same information [46]. Furthermore, students who under perform at tests due to test-anxiety do not suffer from impaired performance when they are asked to write a summary instead [40].

The *Interactive, Constructive, Active, and Passive framework* (ICAP) predicts that as students become more engaged with the learning materials, from passive to active to constructive to interactive, their learning will increase [13]. In this framework, watching a video or reading a text are defined as *passive* modes of engagement based only on receiving. In contrast, answering comprehension questions or summarizing concepts are *constructive* modes of learning based on (knowledge) generation. Writing a summary goes beyond mere *manipulation* of information when students are required to go beyond what is given and generate inferences that go beyond what is presented in the video.

Combined, the research on retrieval practice and summaries provide us with promising educational interventions which can be applied in the context of MOOCs.

1.5 Applying retrieval practice to MOOCs

Although tests are relatively common in online courses it is debatable whether we can assume that those tests necessarily promote retrieval practice and, by extension, meaningful long-term learning. To be able to handle the large amount of learners, many MOOCs employ multiple-choice tests for automatic grading. However, multiple-choice tests only require the student to *passively recognize* the correct answer, while production tests such as short-answer questions require *active reconstruction* of knowledge from memory. This is why production tests generally outperform recognition tests in terms of learning gains [9, 38]. Combined with the fact that watching a video is an inherently passive learning activity, active reconstruction of knowledge can be expected to increase learning gains over passive recognition of information in a multiple-choice quiz.

In online courses with thousands of learners, such as in most MOOCs, it is often not doable to give individual feedback. Given that the benefits of retrieval practice are not dependent on the presence of feedback [5, 8], this opens the possibility to use scalable, content-independent cues to trigger retrieval practice. One option is to ask learners to generate their own questions about the content, which triggers retrieval practice, and improves learning [14, 15]. Here we will focus on instructing students to summarize the most important content of videos as a way to trigger retrieval practice and promote learning.

1.5.1 Current Study. For operationalizing retrieval practice in a MOOC context, two design goals were kept in mind: 1) scalability and 2) effectiveness. For scalability, it is vital that the retrieval cue can be used within any MOOC, is independent of the specific content, and can be employed without requiring much oversight of the course instructors. In order to maximize the effectiveness, the literature on retrieval practice was consulted to come up with a design which can expected to be effective. As a result, the current study has operationalized retrieval practice as follows: after every video the learners are asked to shortly summarize the content of the video. This cue is independent of the specific content of the video, functions without feedback, and is as such easily scalable and employable in any online course. Furthermore, based on the literature discussed earlier this cue can be expected to be effective. In short: it is expected to trigger learners to retrieve the most important

information from memory, requires the production of knowledge, and the retrieval cue is repeated for each new learning object.
While we could have studied the effects of having students write summaries compared to students who do not write summaries, this is not necessarily the most informative comparison. That is, such a design boils down to comparing students who do more with students who do less, which is by itself not very informative. As such, we included a third condition by presenting a portion of the students with pre-written summaries of the videos. Previous research has shown that this is also a valid study strategy which increased learning [34]. As such, given students a pre-written summary can act as a meaningful comparison group to evaluate the effectiveness of instructing students to write a summary.

1.5.2 Hypotheses. Based on the described literature we propose the following hypotheses regarding the effects of writing and reading summaries on quiz grades:

(1) Reading summaries has a positive impact on quiz grades.
(2) Writing summaries has a positive impact on quiz grades.
(3) Writing summaries has a larger impact on quiz grades than reading summaries.

While we do not have any firm hypotheses on the effects of these interventions on dropout rates and students' overall engagement with the course, this is something we will consider to evaluate the practical benefits. In the two experiments described below we tested the three hypotheses. The first experiment acted as a pilot to test the technical feasibility of the intervention, while the much larger second experiment acted as a more powerful test of the hypotheses. We will first present the method and results of each experiment individually, and then discuss the overall results.

2 STUDY 1

2.1 Study 1 Method

2.1.1 Course selection. We selected the MOOC "Terrorism and Counterterrorism: Comparing Theory and Practice" of Leiden University to conduct our study. We chose this course because it extensively uses educational videos and lacks tasks and tests in-between videos (assessments do appear at the end of each course week). The study took place in the two course weeks following the initial introductory section. These two weeks featured a total of 11 videos, each approximately 5 to 15 minutes long.

2.1.2 Design. This study used a sequential cohort design with two iterations, each of which lasted for 6 weeks. In the first iteration we asked students after every video to write a brief summary of the video they had just seen ('Retrieval Practice Version'). To this end, we presented the learner with the question "Please shortly summarize the video for yourself in about three sentences". This prompt was presented on a new page after each video. A text box was presented underneath the question in which they could write the summary.
During the second iteration the students instead received a written summary of each video ('Given Summary Version'). Each of these three-sentence long summaries was presented after its respective video. These summaries were written by the main author in collaboration with the teaching staff of the course. We instructed the learners to "Please read the summary of the video carefully.". A check box was presented under the summary so the learners could indicate whether they had read the summary.

2.1.3 Participants. Participation in a MOOC is completely self-directed. The MOOC learners were free to skip the questions regarding writing or reading a summary. For the purpose of this study, all the learners who wrote or read at least one summary were included in the analyses.

2.2 Study 1 Results and Discussion

All the relevant data for this study is openly available at https://osf.io/qz5m6/.

2.2.1 Descriptive statistics. A total of 823 learners wrote or read at least one summary. Figure 1 shows engagement with both interventions for every individual video in the first two weeks of the course. This information is also presented in Table 1. As is typical for average engagement levels in MOOCs, it decreases substantially over time. There appears to be no difference in the amount of learners who write or read summaries, despite the fact that writing a summary demands a higher investment by the student. Due to a technical issue we did not obtain any data for the first video in week 1, nor the final quiz grade (but these are included in the follow-up study).

The means and standard deviations of the length (in characters) of the written summaries can be found in Table 2. Note that the instruction for the students was to "Please shortly summarize the video for yourself in about three sentences". As such, the average length of about 230 characters per summary is appropriate, and reflect that the students followed the instruction. While the total number of written summaries decreases over time, the average length of the summaries does not. In other words, while fewer students were writing summaries, the length of the summaries remained consistent. Exploratory analyses using the length of the summaries as a predictor or as a mediator for learning or engagement did not result in anything noteworthy, and will not be reported.

2.2.2 Quiz scores. We were interested in the quiz grades of the respective week as a result of either writing or reading video summaries. For that purpose we computed Pearson correlations between the amount of summaries that the learners read or wrote in a week, and their score on the relevant quiz. These correlations are shown in Table 3. Writing summaries was positively correlated with the respective weekly quiz, $\rho = 0.087$ $p = 0.018$ in week 1, and $\rho = 0.137$, $p = 0.004$ in week 2. Similar results were found for reading summaries: $\rho = 0.138$, $p < 0.001$ in week 1, and $\rho = 0.138$, $p = 0.003$ in week 2.

These results show that the amount of read or written summaries correlate with the quiz grades. To test whether these correlations are significantly different based on the intervention type, we performed Fisher r-to-z transformation. However, there was no significant difference between the correlation pairs in week 1 ($z = 1.00$, $p = 0.317$) nor in week 2 ($z = 0.02$, $p = 0.984$). So while the amount of engagement with the interventions (whether active retrieval or passive rereading) positively correlated with the students' quiz grades, we

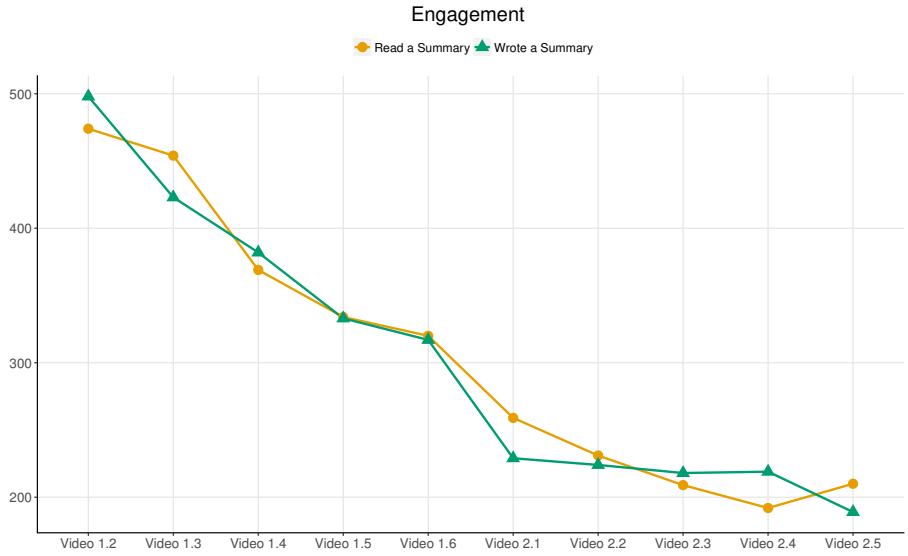

Figure 1: Engagement with the interventions in Study 1

Table 1: Number of students who read or wrote summaries in Study 1

	Video 1.2	Video 1.3	Video 1.4	Video 1.5	Video 1.6	Video 2.1	Video 2.2	Video 2.3	Video 2.4	Video 2.5
Read Summary	474	454	369	334	320	259	231	209	192	210
Wrote Summary	498	423	382	333	317	229	224	218	219	189

Table 2: Length of the written summaries in characters in Study 1.

	Mean	Standard Deviation
Video 1.2	235.00	165.30
Video 1.3	180.37	113.93
Video 1.4	252.52	170.24
Video 1.5	237.67	148.81
Video 1.6	217.02	135.54
Video 2.1	255.05	189.02
Video 2.2	246.66	157.50
Video 2.3	240.52	150.58
Video 2.4	238.68	150.72
Video 2.5	224.65	158.86

Note. Lengths are given in characters.

observe no significant difference between these two interventions' effect on quiz scores.

3 STUDY 2

3.1 Study 2 Method

The first experiment followed a cohort study design, which comes with various limitations. After the study was completed, Coursera offered the option to randomly allocate learners to different versions of a course running in parallel. This opportunity was used to replicate the first study with a more rigorous methodology and more participants.

Study 2 took place in the same MOOC as study 1 and also focused on the first two weeks of the course. The main difference between the two studies was that in study 2 the learners were randomly allocated to one of three conditions: 1) Given Summary, 2) Retrieval Practice, and 3) a Control condition, which received neither. Other than the randomized allocation, the conditions were identical to those in the first study.

3.2 Study 2 Results and Discussion

All the relevant data for this study is openly available at https://osf.io/qz5m6/.

3.2.1 Learner participation. A total of 12.444 learners participated in the study. After random allocation to the three conditions, there were 4.169 participants in the control condition, 4.128 in the given summaries condition, and 4.146 in the retrieval practice condition. Figure 2 summarizes the learner progress in terms of how far the learners progressed in terms of taking the quiz of week 1, week 2, and the final test.

These data were analyzed with a Repeated Measures ANOVA, to see whether the student drop-outs can be predicted by condition and/or time (start, week 1, week 2, final quiz). Mauchly's Test of Sphericity indicated that the assumption of sphericity had been violated, $\chi^2(5) = 7245.09$, $p < .001$, therefore degrees the freedom were corrected using Greenhouse-Geisser estimates of spherecity,

Table 3: Pearson correlation matrix for writing/reading summaries and quiz scores in Study 1.

	Week 1 Grade	Week 2 Grade
Number of read summaries (week 1)	0.138 (n = 769, p < 0.001)	
Number of read summaries (week 2)		0.138 (n = 471, p = 0.003)
Number of written summaries (week 1)	0.087 (n = 743, p = 0.018)	
Number of written summaries (week 2)		0.137 (n = 431, p = 0.004)

Note. These analyses only include students who have written/read summaries *and* completed a weekly/-final quiz. As such, the sample sizes for these analyses are lower than those reported in Figure 1 and Table 1 as these also include students which did not do any of the quizzes.

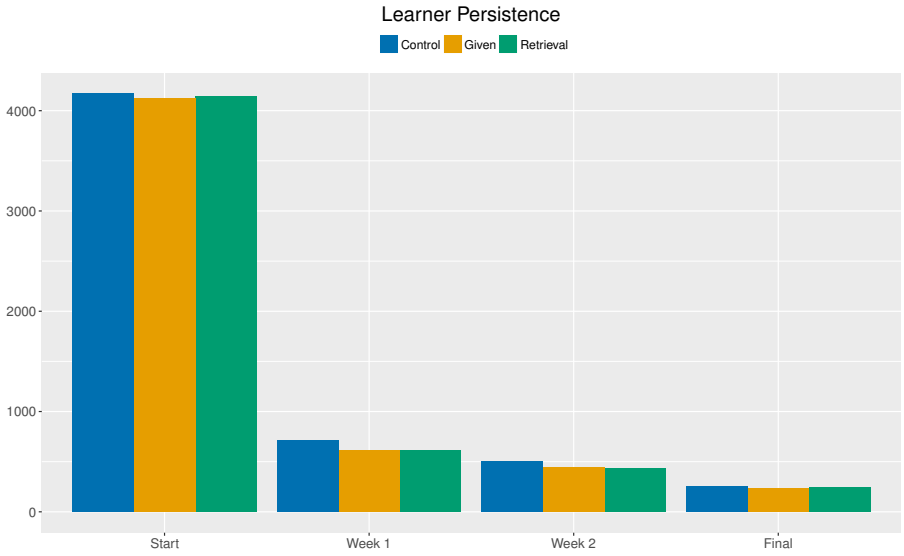

Figure 2: Amount of students who were active at the start of the course, the week 1 quiz, week 2 quiz, and at the final quiz in Study 2.

Table 4: Number of students who were active at the start of the course, week 1 quiz, week 2 quiz, and at the final quiz in Study 2.

Condition	Start	Week 1	Week 2	Final
Control	4169	709	501	259
Given	4125	614	439	230
Retrieval	4146	616	436	245

ε = .714. There was a very strong effect of time on learner dropouts, F(2.142, 26644.88) = 61664.47, p < .001, partial η^2 = .832. In other words, student attrition increased substantially as the course progressed each week.

There was also a significant albeit very weak between-subjects effect of condition, F(2, 12440) = 3,478, p = 0.031, partial η^2 = .001. Similarly, there was a significant but very weak condition by time interaction, F(4.284, 26644.88) = 3.273, p = .009, partial η^2 < .001. This effect of the condition and its interaction with time is driven by the higher number of learners in the control condition in the first two weeks. At the quiz of week 1, there are significantly more learners in the control condition than the given summary condition, 709 vs 614, χ^2 = 7.037, p = 0.008, as well as in the second week, 501 vs 439, χ^2 = 3.947, p = 0.047, but not at the final quiz, 259 vs 230, χ^2 = 1.536, p = 0.215. In other words, there are more early dropouts in the given summary condition than in the control condition, but this difference appears to diminish with time. The same pattern is found when comparing the control and retrieval conditions at week 1, 709 vs 616, χ^2 = 7.165, p = 0.007, at the quiz of week 2, 501 vs 436, χ^2 = 4.685, p = 0.032, but there is no difference at the final quiz, 259 vs 245, χ^2 = 0.336, p = 0.562. The intervention can thus be considered a filter—able to identify learners most likely to persist deeper in the course due to their willingness to engage with more course materials early on.

Note that the above analyses include *all* learners, not just those that engaged with the interventions. This was done to test the overall impact of *implementing* these interventions on cohorts of students. In the next section we zoom in specifically on those students who interacted with either intervention.

3.2.2 Engagement with intervention. Not all learners engaged with the interventions in the two experimental conditions. This is typical for MOOCs, where the majority of learners show very low

levels of activity and there are high drop-out rates. In Figure 3 we observe how many hundreds of learners engage with either intervention and to what extent this engagement declines over time. There is a noticeable drop in engagement at the transition from week 1 to 2. Overall, substantially more learners in the 'given summary' condition engage with the intervention than the 'write a summary' condition, most likely because the latter requires more effort from the learners.

Table 6 shows the means and standard deviations for the length (in characters) of the written summaries. Interestingly, while there are progressively fewer learners who actually write a summary, the average length of the summaries remains stable over time. This is consistent with the result of Study 1. The average length of almost 300 characters per summary is appropriate given that the students were instructed to summarize the videos in "about three sentences". Exploratory analyses using the length of the summaries as a predictor or as a mediator for learning or engagement did not result in anything noteworthy, and will not be reported.

3.2.3 Quiz scores. When we look at all learners in the three conditions, we detect no differences in quiz scores. That is, there is no difference at the week 1 quiz, $F(2, 1938) = 1.457$, $p = 0.233$, nor at the week 2 quiz, $F(2, 1375) = 2.905$, $p = 0.055$, nor at the final quiz, $F(2, 733) = 0.289$, $p = 0.749$.

However, as noted above, not all learners engaged with the intervention. As such, we computed Pearson correlations between the amount of summaries that the learners read or wrote in a week, and their score on the relevant quiz. These correlations are shown in Table 7.

Learners who wrote or read more summaries of the videos also scored higher on the weekly quizzes. With correlations between 0.121 and 0.323 these are noticeable correlations. However, the amount of written or read summaries correlate less strongly with the final quiz grade; $\rho = 0.115$, $p = 0.016$ for the total amount of read summaries, and $\rho = 0.087$, $p = 0.218$ for the total amount of written summaries. In other words, reading more summaries in the first two weeks of the course significantly predicts a higher grade on the final quiz, while this is not true for writing more summaries.

To test whether these correlations are significantly different based on the intervention type, we performed Fisher r-to-z transformations. However, there was no significant difference between the correlation pairs in week 1 ($z = 0.29$, $p = 0.386$) nor in week 2 ($z = -1.59$, $p = 0.056$), nor for the final quiz ($z = 0.33$, $p = 0.370$).

Contrary to our hypothesis, writing summaries of videos does not appear to lead to better quiz scores than passively reading provided summaries. This is a surprising result, given the extensive literature on how active memory retrieval outperforms passive reading as a learning strategy. However, while the two interventions do not different from each other, they do appear to both have a positive impact on quiz grades.

4 GENERAL DISCUSSION

In this paper we described two studies investigating the effects of prompting learners to either read or write a summary after every instructional video. As merely watching a video is a passive learning activity, these more active post-video assignments were expected to increase knowledge retention. Based on the retrieval practice literature we furthermore expected that writing a summary would be more beneficial for learning than reading a given summary. Both studies show the more summaries students read or wrote, the higher their quiz scores. This is a promising result, as the interventions were specifically designed to be easy to implement in a wide variety of MOOCs (i.e., they are domain-independent) and require little to no oversight from the course instructor.

Interestingly, there was no apparent difference in the efficacy of writing versus reading summaries - there was no distinction in how strongly they correlated with quiz grades. However, there are two relevant practical differences between the two interventions. First, providing students with a summary ensures that all students have similar access to a high quality summary of the video. This also promotes the inclusiveness and accessibility of a course. Secondly, substantially more students *read* a summary as opposed to *writing* one. In other words, while both interventions appear to be equally effective, providing pre-written summaries increases *how many* students benefit from the intervention.

4.1 Does giving learners summaries of the videos help them learn?

Our first hypothesis was that reading summaries has a positive impact on quiz grades. These two experiments both provide evidence in favor of affirming this hypothesis, but this comes with a number of caveats. Yes, there is a moderately strong, positive correlation between reading summaries and obtaining higher grades. That is, learners who read more summaries have higher grades both in the weekly quizzes as well as on the final course quiz. Notably, these findings were found in both studies despite the different methodology. This replication, and the large sample size speaks to the reliability of this finding.

An important caveat is that this is a correlational observation, and there are alternative interpretations. For example, it is possible that at least a portion of this correlation is explained by learner self-selection; high performers might have been more likely to engage with this intervention.

4.2 Does prompting learners to write summaries of videos help them learn?

Our second hypothesis was that writing summaries has a positive effect on quiz grades. Similar to the previous hypothesis, these experiments provide evidence to suggest that this is indeed the case. In both studies there were consistent correlations between the amount of written summaries and the quiz grades. This finding comes with the same strengths (large sample size and successful replication) and caveats (it being a correlational finding) as mentioned above. An important difference with reading summaries, is that the amount of written summaries did not significantly correlate with the final course grade.

In both studies, the students received no guidance in how they should summarize the videos. As not all students might have been able to produce summaries of sufficient quality, future research could investigate the effects of more direct instruction, which has been shown to increase the quality of written summaries [6, 12].

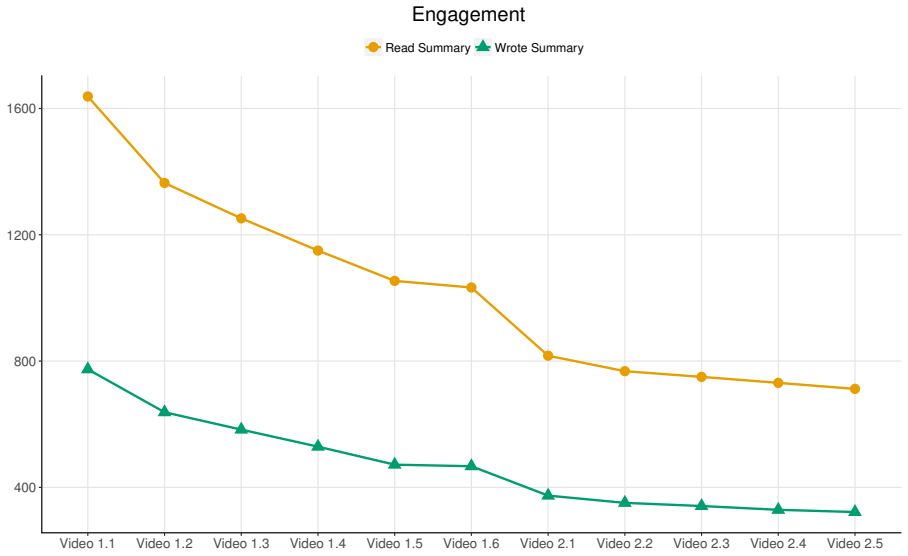

Figure 3: Engagement with the interventions in Study 2

Table 5: Number of students who read or wrote summaries in Study 2.

	Video 1.1	**Video 1.2**	**Video 1.3**	**Video 1.4**	**Video 1.5**	**Video 1.6**	**Video 2.1**	**Video 2.2**	**Video 2.3**	**Video 2.4**	**Video 2.5**
Read Summary	1638	1364	1252	1150	1054	1033	817	768	750	731	712
Wrote Summary	774	638	583	529	472	467	374	351	341	329	322

Table 6: Length of the written summaries in characters in Study 2.

	Mean	Standard Deviation
Video 1.1	288.95	160.996
Video 1.2	328.47	246.835
Video 1.3	234.05	133.545
Video 1.4	315.21	186.945
Video 1.5	319.85	287.396
Video 1.6	275.16	155.905
Video 2.1	303.90	182.279
Video 2.2	291.89	215.311
Video 2.3	285.75	194.491
Video 2.4	304.00	208.482
Video 2.5	305.47	224.081

Note. Lengths are given in characters.

4.3 Do learners benefit more from writing or reading summaries?

Our third and final hypothesis was that writing summaries has a larger impact on quiz grades than reading summaries. Both studies failed to provide evidence for this hypothesis. In each study, the correlations of both interventions were similar and never significantly different from each other. That is, while both interventions are positively correlated with weekly quiz grades, these correlations were not statistically distinguishable from each other.

A relevant, but hard to interpret difference between the two interventions is that while the amount of read summaries was significantly correlated with the final quiz grade, the amount of written summaries was not. While these two correlations (between the amount of read/written summaries and final quiz grades) differ in strength, this difference was not significant. In short, we can conclude that 1) reading summaries and the final quiz grade are correlated, 2) there is no evidence that writing summaries and the final quiz grade are correlated, and 3) there is no evidence that reading summaries has a bigger impact on the final quiz grade than writing summaries.

Another difference between the two interventions, is that (at least in the much larger second study) much more learners will *read* a summary of a video as opposed to *writing* one themselves. This is something course instructors and designers should take in mind when designing or adapting a course.

This finding is surprising given the literature on retrieval practice; we expected a performance advantage based on writing summaries as opposed to merely reading given summaries. We propose several plausible interpretations for this discrepancy. First of all, it is typically found that the benefits of retrieval practice emerge when the retrieval event happens after some time has passed since the learning phase. However, due to the inherent user freedom of MOOCs it is hard to force learners to wait after a video before they write a

Table 7: Pearson correlation matrix for writing/reading summaries and quiz scores in Study 2.

	Week 1 Grade	Week 2 Grade	Final Grade
Number of read summaries (week 1)	0.136 (n = 1158, p < 0.001)		
Number of read summaries (week 2)		0.227 (n = 766, p < 0.001)	
Number of read summaries (total)			0.115 (n = 436, p = 0.016)
Number of written summaries (week 1)	0.121 (n = 368, p = 0.008)		
Number of written summaries (week 2)		0.323 (n = 339, p < 0.001)	
Number of written [summaries (total)			0.087 (n = 203, p = 0.218)

Note. These analyses only include students who have written/read summaries *and* completed a weekly/final quiz. As such, the sample sizes for these analyses are lower than those reported in Figure 3 and Table 5 as these also include students which did not do any of the quizzes.

summary about it. Secondly, it is possible that both interventions were sufficient helpful to ensure that learners would successfully complete the quizzes; such a ceiling effect could obscure any potential differences.

Ultimately, both present studies advance the literature by testing the efficacy of well-known instructional interventions in a large educational setting. While the nature of MOOC data should make us cautious about drawing causal inferences [48], it does provide us with valuable information about the potential effects of these educational interventions.

4.4 Unexpected findings

This study also comes with a cautionary tale. Both experimental interventions appear to have caused some more learners to drop-out in the first two weeks of the course. One possible interpretation is that this is due to the increased task demands and pressure to perform, which might cause more learners to decide that the course is not for them. Although we are not yet certain about the plausible mechanisms, it appears that even educational interventions which are relatively minor and completely optional can potentially affect dropouts in unexpected ways. However, it is important to note that the *overall* drop-out rate was equivalent in all conditions, leading us to assume that these interventions might merely cause drop-outs to happen *earlier*, and not necessarily to happen *more frequently*.

Ultimately, this is a question of perspective. While MOOCs have been frequently criticized for their high attrition rate, these numbers are strongly driven by their low barrier to entry. Consequently, many learners have been found to start a MOOC to simply 'check it out' but ultimately decide to leave. It is debatable whether this should be considered a loss for the learner or the course instructors. From this perspective, causing the drop-out peak to occur earlier in a MOOC could be considered a positive effect, potentially saving time and energy of the learners. However, both perspectives are partly based on speculative assumptions regarding learners' intentions and interests and would benefit from more research.

4.5 Future directions

MOOCs are a promising new environment to test educational interventions due to their scale and the non-intrusive data collection, which can give an unprecedented view of learner's behaviors. Nevertheless, there are various challenges with respect to the internal and external validity of MOOC research [48]. The open and free nature of MOOCs also act as a limiting factor to experimental research such as described in this paper. Given the open design of MOOCs and freedom for learners to self-direct their learning with full autonomy, it is difficult to distinguish artifacts from true (and potentially causal) relationships. Nevertheless, MOOC studies provide an unique opportunity to study and apply well known phenomena from the laboratory in large-scale, real-world educational settings. Relying on robust theories on learning from the educational and cognitive sciences, combined with the strengths of learning analytics, is a future direction which seems most promising.

REFERENCES

[1] Edwina E Abott. 1909. On the analysis of the factor of recall in the learning process. *The Psychological Review: Monograph Supplements* 11, 1 (1909), 159.

[2] IE Allen, J Seaman, R Poulin, and TT Straut. 2016. Online report card: Tracking online education in the United States. *Babson Park, MA: Babson Survey Research Group and Quahog Research Group, LLC* (2016).

[3] Alan Baddeley. 2003. Working memory: looking back and looking forward. *Nature Reviews. Neuroscience* 4, 10 (2003), 829.

[4] Robert L Bangert-Drowns, Marlene M Hurley, and Barbara Wilkinson. 2004. The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of educational research* 74, 1 (2004), 29–58.

[5] Robert L Bangert-Drowns, Chen-Lin C Kulik, James A Kulik, and MaryTeresa Morgan. 1991. The instructional effect of feedback in test-like events. *Review of educational research* 61, 2 (1991), 213–238.

[6] Thomas W Bean and Fern L Steenwyk. 1984. The effect of three forms of summarization instruction on sixth graders' summary writing and comprehension. *Journal of Reading Behavior* 16, 4 (1984), 297–306.

[7] Lori Breslow, David E Pritchard, Jennifer DeBoer, Glenda S Stump, Andrew D Ho, and Daniel T Seaton. 2013. Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment* 8 (2013).

[8] Andrew C Butler, Jeffrey D Karpicke, and Henry L Roediger III. 2007. The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied* 13, 4 (2007), 273.

[9] Andrew C Butler and Henry L Roediger III. 2007. Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology* 19, 4-5 (2007), 514–527.

[10] Jennifer Campbell, Alison L Gibbs, Hedieh Najafi, and Cody Severinski. 2014. A comparison of learner intent and behaviour in live and archived MOOCs. *The International Review of Research in Open and Distributed Learning* 15, 5 (2014), 235–262. http://www.irrodl.org/index.php/irrodl/article/view/1854

[11] Shana K Carpenter and EDWARD L DeLOSH. 2005. Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology* 19, 5 (2005), 619–636.

[12] Kuo-En Chang, Yao-Ting Sung, and Ine-Dai Chen. 2002. The effect of concept mapping to enhance text comprehension and summarization. *The Journal of Experimental Education* 71, 1 (2002), 5–23.

[13] Michelene TH Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist* 49, 4 (2014), 219–243.

[14] Beth Davey and Susan McBride. 1986. Effects of question-generation training on reading comprehension. *Journal of Educational Psychology* 78, 4 (1986), 256.

[15] Beth Davey and Susan McBride. 1986. Generating self-questions after reading: A comprehension assist for elementary students. *The Journal of Educational*

Research 80, 1 (1986), 43–46.

[16] Dan Davis, Guanliang Chen, Tim Van der Zee, Claudia Hauff, and Geert-Jan Houben. 2016. Retrieval practice and study planning in moocs: Exploring classroom-based self-regulated learning strategies at scale. In *European Conference on Technology Enhanced Learning*. Springer, 57–71.

[17] Rebecca Ferguson and Doug Clow. 2015. Examining engagement: analysing learner subpopulations in massive open online courses (MOOCs). In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. ACM, 51–58.

[18] Elia Fernández-Díaz, Carlos Rordíguez-Hoyos, and Adelina Calvo Salvador. 2017. The Pedagogic Architecture of MOOC: A Research Project on Educational Courses in Spanish. *The International Review of Research in Open and Distributed Learning* 18, 6 (2017).

[19] Yang Gao. 2013. The effect of summary writing on reading comprehension: the role of mediation in EFL classroom. *Reading Improvement* 50, 2 (2013), 43–47.

[20] Amie Goldberg, Michael Russell, and Abigail Cook. 2003. The effect of computers on student writing: A meta-analysis of studies from 1992 to 2002. *The Journal of Technology, Learning and Assessment* 2, 1 (2003).

[21] Philip J Guo, Juho Kim, and Rob Rubin. 2014. How video production affects student engagement: An empirical study of mooc videos. In *Proceedings of the first ACM conference on Learning Scale conference*. ACM, 41–50. https://doi.org/10.1145/2556325.2566239

[22] Sherif Halawa, Daniel Greene, and John Mitchell. 2014. Dropout prediction in MOOCs using learner activity features. *Experiences and best practices in and around MOOCs* 7 (2014), 3–12.

[23] Khe Foon Hew and Wing Sum Cheung. 2014. Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges. *Educational research review* 12 (2014), 45–58.

[24] Andrew Dean Ho, Isaac Chuang, Justin Reich, Cody Austun Coleman, Jacob Whitehill, Curtis G Northcutt, Joseph Jay Williams, John D Hansen, Glenn Lopez, and Rebecca Petersen. 2015. Harvardx and mitx: Two years of open online courses fall 2012-summer 2014. (2015).

[25] Andrew Dean Ho, Justin Reich, Sergiy O Nesterko, Daniel Thomas Seaton, Tommy Mullaney, Jim Waldo, and Isaac Chuang. 2014. HarvardX and MITx: The first year of open online courses, fall 2012-summer 2013. (2014).

[26] Cheryl I Johnson and Richard E Mayer. 2009. A testing effect with multimedia learning. *Journal of Educational Psychology* 101, 3 (2009), 621.

[27] Katy Jordan. 2015. Massive open online course completion rates revisited: Assessment, length and attrition. *The International Review of Research in Open and Distributed Learning* 16, 3 (2015).

[28] Tali Kahan, Tal Soffer, and Rafi Nachmias. 2017. Types of Participant Behavior in a Massive Open Online Course. *The International Review of Research in Open and Distributed Learning* 18, 6 (2017).

[29] Jeffrey D Karpicke and Henry L Roediger. 2008. The critical importance of retrieval for learning. *science* 319, 5865 (2008), 966–968.

[30] Hanan Khalil and Martin Ebner. 2013. "How satisfied are you with your MOOC?"-A Research Study on Interaction in Huge Online Courses. In *EdMedia: World Conference on Educational Media and Technology*. Association for the Advancement of Computing in Education (AACE), 830–839.

[31] Juho Kim, Philip J Guo, Daniel T Seaton, Piotr Mitros, Krzysztof Z Gajos, and Robert C Miller. 2014. Understanding in-video dropouts and interaction peaks inonline lecture videos. In *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 31–40.

[32] René F Kizilcec, Jeremy N Bailenson, and Charles J Gomez. 2015. The instructor's face in video instruction: Evidence from two large-scale field studies. *Journal of Educational Psychology* 107, 3 (2015), 724.

[33] René F Kizilcec, Chris Piech, and Emily Schneider. 2013. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge*. ACM, 170–179.

[34] Claudia Leopold, Elke Sumfleth, and Detlev Leutner. 2013. Learning with summaries: Effects of representation mode and type of learning activity on comprehension and transfer. *Learning and Instruction* 27 (2013), 40–49.

[35] Anoush Margaryan, Manuela Bianco, and Allison Littlejohn. 2015. Instructional quality of Massive Open Online Courses (MOOCs). *Computers & Education* 80 (jan 2015), 77–83. https://doi.org/10.1016/j.compedu.2014.08.005

[36] Richard E Mayer. 2002. Multimedia learning. *Psychology of learning and motivation* 41 (2002), 85–139.

[37] Richard E Mayer, Julie Heiser, and Steve Lonn. 2001. Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of educational psychology* 93, 1 (2001), 187.

[38] Mark A McDaniel, Janis L Anderson, Mary H Derbish, and Nova Morrisette. 2007. Testing the testing effect in the classroom. *European Journal of Cognitive Psychology* 19, 4-5 (2007), 494–513.

[39] Mark A McDaniel, Daniel C Howard, and Gilles O Einstein. 2009. The read-recite-review study strategy effective and portable. *Psychological Science* 20, 4 (2009), 516–522.

[40] Wilson Shun Yin Mok and Winnie Wai Lan Chan. 2016. How do tests and summary writing tasks enhance long-term retention of students with different levels of test anxiety? *Instructional Science* 44, 6 (2016), 567–581.

[41] Ludmila D Nunes and Yana Weinstein. 2012. Testing improves true recall and protects against the build-up of proactive interference without increasing false recall. *Memory* 20, 2 (2012), 138–154.

[42] Henry L Roediger and Andrew C Butler. 2011. The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences* 15, 1 (2011), 20–27.

[43] Henry L Roediger and Jeffrey D Karpicke. 2006. The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science* 1, 3 (2006), 181–210.

[44] Henry L Roediger and Jeffrey D Karpicke. 2006. Test-enhanced learning taking memory tests improves long-term retention. *Psychological science* 17, 3 (2006), 249–255.

[45] Daniel T Seaton, Yoav Bergner, Isaac Chuang, Piotr Mitros, and David E Pritchard. 2014. Who does what in a massive open online course? *Commun. ACM* 57, 4 (2014), 58–65.

[46] Arie Spirgel. 2011. *Assessing summary writing as a memory strategy*. The University of North Carolina at Greensboro.

[47] Frans Van der Sluis, Jasper Ginn, and Tim Van der Zee. 2016. Explaining Student Behavior at Scale: The Influence of Video Complexity on Student Dwelling Time. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM, 51–60.

[48] Frans van der Sluis, Tim Van der Zee, and Jasper Ginn. 2017. Learning about Learning at Scale: Methodological Challenges and Recommendations. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*. ACM, 131–140.

[49] Tim Van der Zee, Wilfried Admiraal, Fred Paas, Nadira Saab, and Bas Giesbers. 2017. Effects of subtitles, complexity, and language proficiency on learning from online education videos. *Journal of Media Psychology* (2017).

[50] Margot van Wermeskerken and Tamara van Gog. 2017. Seeing the instructor's face and gaze in demonstration video examples affects attention allocation but not learning. *Computers & Education* (2017).

[51] Ming Yang, Zhen Shao, Qian Liu, and Chuiyi Liu. 2017. Understanding the quality factors that influence the continuance intention of students toward participation in MOOCs. *Educational Technology Research and Development* 65, 5 (2017), 1195–1214.

[52] Paul Zhihao Yong and Stephen Wee Hun Lim. 2015. Observing the Testing Effect using Coursera Video-Recorded Lectures: A Preliminary Study. *Frontiers in psychology* 6 (2015).

[53] Ahmed Mohamed Fahmy Yousef, Mohamed Amine Chatti, Ulrik Schroeder, and Marold Wosnitza. 2014. What drives a successful MOOC? An empirical examination of criteria to assure design quality of MOOCs. In *Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference on*. IEEE, 44–48.