Effects of subtitles, complexity, and language proficiency on learning from online education videos

Zee, T. van der; Admiraal, W.F.; Paas, F.; Saab, N.; Giesbers, B.

Effects of Subtitles, Complexity, and Language proficiency on Learning from Online Education Videos

Tim van der Zee, Wilfried Admiraal
Leiden University

Fred Paas
Erasmus University Rotterdam, University of Wollongong

Nadira Saab
Leiden University

Bas Giesbers
Erasmus University Rotterdam

Abstract

Open online education has become increasingly popular. In Massive Open Online Courses (MOOCs) videos are generally the most used method of teaching. While most MOOCs are offered in English, the global availability of these courses has attracted many non-native English speakers. To ensure not only the availability, but also the accessibility of open online education, courses should be designed to minimize detrimental effects of a language barrier, for example by providing subtitles. However, with many conflicting research findings it is unclear whether subtitles are beneficial or detrimental for learning from a video, and whether this depends on characteristics of the learner and the video. We hypothesized that the effect of 2nd language subtitles on learning outcomes depends on the language proficiency of the student, as well as the visual-textual information complexity of the video. This three-way interaction was tested in an pre-registered experimental study. Using Bayesian analyses, no main effect of subtitles was found, nor any interaction. However, the student's language proficiency and the complexity of the video do have a substantial impact on learning outcomes.

Introduction

Open online education has rapidly become a highly popular method of education. The promise – global and free access to high quality education – has often been applauded. With a reliable Internet connection comes free access to a large variety of Massive Open Online Courses (MOOCs) found on platforms such as Coursera and edX. MOOC participants indeed come from all over the world, although participants from Western countries are still overrepresented (Nesterko et al., 2013). In all cases there are many non-native English speakers in English courses. This raises the question to what extent non-native English speakers can benefit from these courses, compared to native speakers. Open online education may be *available* to most, the content might not be as *accessible* for many due to language barriers. It is important to design online education in such a way that it minimizes detrimental effects of potential language barriers to increase its accessibility for a wider audience.

MOOCs typically feature a large number of videos which are central to the student learning experience (Guo, Kim, & Rubin, 2014; Liu et al., 2013). The central position of educational videos is reflected by students' behavior and their intentions: most students plan to watch all videos in a MOOC, and also spend the majority of their time watching these videos (Campbell, Gibbs, Najafi, & Severinski, 2014; Seaton, Bergner, Chuang, Mitros, & Pritchard, 2014). In this study, we investigate the impact of subtitles on learning from educational videos in a second language. Providing subtitles is a common approach to cater to diverse audiences and support non-native English speakers. The Web Content Accessibility Guidelines 2.0 (WCAG, 2008) prescribe subtitles for any audio media to ensure a high level of accessibility. Intuitively there seems nothing wrong with this advice and many studies have indeed found a positive effect of subtitles on learning (e.g., Markham, Peter, McCarthy, et al., 2001). However, a different set of studies provides evidence that subtitles can also hamper learning (e.g., Kalyuga, Chandler, & Sweller, 1999). In the current study the effects of subtitles will be further examined.

This paper is organized as follows: First, conflicting findings on the effects of subtitles on learning will be discussed. Secondly, a framework will be proposed which can explain these conflicting findings by considering the interaction between subtitles, language proficiency, and visual-textual information complexity (VTIC). In turn, an experimental study will be described which tests the main hypothesis of the framework.

Although this study is situated in an online educational setting, the results may also be of relevance for other media-orientated fields, such as film studies and video production.

Subtitles: Beneficial or Detrimental for Learning?

Research on the effects of subtitles typically differentiates between subtitles in someone's native language, called L1, versus subtitles in one's second language, or L2. A meta-analysis of 18 studies showed positives effects of L2 subtitles for language learning (Perez, Noortgate, & Desmet, 2013). Specifically, enabling subtitles for language learning videos substantially increases student performance on recognition tests, and to a lesser extent on production tests. Other studies have found similar positive effects of subtitles on learning from videos and there appears to be a consensus that subtitles are beneficial for learning a second language (e.g., Baltova, 1999; Chung, 1999; Markham, 1999; Winke, Gass, &

Sydorenko, 2013). However, these are all studies which focus on learning a language and not on learning about a non-linguistic topic in a second language. There are important differences between language learning, and what we will call content learning. When learning a language, practicing with reading and understanding L2 subtitles is directly relevant for this goal. In contrast, when learning about a specific topic, apprehending L2 subtitles is not a goal in itself but only serves the purpose of better understanding the actual content. As such, we would argue that findings from studies focusing on language learning are by themselves not convincing enough to be directly applied to content learning, as subtitles have a different relationship with the content and the learning goals.

In contrast to studies on language learning, there are only a few studies which investigated the effects of subtitles for content learning. These studies have shown positive effects for subtitles for content learning in a second language. For example, when watching a short Spanish educational clip, English-speaking students benefited substantially from Spanish subtitles, but even more so from English subtitles (Markham et al., 2001). Another study, focused on different combinations of languages, similarly show that students performed better at comprehension tests when watching a L2 video with subtitles enabled (Hayati & Mohmedi, 2011).

Although several studies did find positive effects of subtitles, a range of other studies yielded contradictory findings. For example, Kalyuga et al. (1999) found that narrated videos without subtitles are better for learning than videos with subtitles. In this study, subtitles were shown to lead to lower performance, an increased perceived cognitive load, and more reattempts during the learning phase (i.e., re-watching videos). This is in contrast with the earlier discussed studies, which showed positive effects of learning from videos with subtitles. In a different study on learning from narrated videos, two experiments showed that enabling subtitles led to lower knowledge retention and transfer (Mayer, Heiser, & Lonn, 2001). With Cohen's *d* effect sizes ranging from 0.36 to 1.20, the detrimental effects of subtitles in these studies were quite substantial. A range of other studies found similar evidence that for content- and language-learning alike, narrated explanations are typically better than showing only subtitles, or narration combined with subtitles. (Harskamp, Mayer, & Suhre, 2007; Mayer, Dow, & Mayer, 2003; Mayer & Moreno, 1998; Moreno & Mayer, 1999). Finally, some studies showed neither a positive nor a negative effect of subtitles on learning (e.g., Moreno & Mayer, 2002a).

Explaining Conflicting Findings on the Effects of Subtitles

The previously discussed literature provides a confusing paradox for instructional designers: are subtitles beneficial, detrimental, or irrelevant for learning? Here we will present an attempt to explain the conflicting findings using a framework built on theories of attention and information processing. In short, we propose that the conflicting findings can be integrated by considering the interaction between subtitles, language proficiency, and the level of visual-textual information complexity (VTIC) in the video.

Working memory limitations. An essential characteristic of the human cognitive architecture is that not every type of information is processed in an identical way. Working memory is characterized by having modality-specific channels, one for auditory and one for visual information (Baddeley, 2003). Both have a limited capacity for information, which can only hold information chunks for a few moments before they decay (Baddeley,

2003). During learning tasks, working memory acts as a bottleneck for processing novel information; as more cognitive load is imposed on the learner, less cognitive resources are available for the integration of information into long-term memory, effectively impairing learning (Ginns, 2006; Sweller, Van Merrienboer, & Paas, 1998). For novel information, the cognitive resources required for processing appears to be primarily dictated by measurable attributes of the information-in-the-world such as the amount of words and their interactivity (Sweller, 2010). As each channel has its own capacity it is generally more effective to distribute processing load between both channels, instead of relying only on one modality (Mayer, 2003). When two sources of information are presented in the same modality this can (more) easily overload our limited processing capacity (Kalyuga et al., 1999). This provides an explanation of why a range of studies found negative effects of subtitles when learning from videos, as both are sources of visual information.

Textual vs non-textual visual information. Up to now we did not distinguish between textual and non-textual visual information. As previously discussed, auditory and visual information are initially processed in separate channels. However, after this initial processing, any language presented either visually or verbally will be processed in the same working memory sub-component: the phonological loop (Baddeley, 2003). In contrast, non-textual visual information is processed in a different component, the visuospatial sketchpad. This notion can further clarify the earlier presented findings. Specifically, the presence of textual visual information, as compared to non-textual visual information becomes an important variable to account for. A video which contains three different sources of language - narration, subtitles, and in-video text - is likely to induce cognitive overload. If the visual information in the video does not have a language component we can expect a reduced, or no detrimental effect. More precisely, subtitles are expected to be detrimental to learning when a video already has a high level of VTIC. However, when a video has a relatively low level of VTIC, adding subtitles will not necessarily lead to cognitive overload. Should the addition of subtitles be desired, it then becomes necessary to ensure that the VTIC of a video is low enough to prevent detrimental effects due to cognitive overload. We propose two ways how the VTIC of an educational video can be manipulated whilst maintaining the educationally relevant content.

Amount of visual-textual information. The first, and most straightforward aspect of VTIC is the amount of visual-textual information shown in a video. That is, a video in which much more text is shown is arguably more complex to process than a video with much less text. However, while removing information which is vital to understand the topic of the video might reduce the complexity, it will also harm the educational value of the video. However, removing or adding visual-textual information which is not strictly relevant for the learning goals can be used to respectively decrease or increase the VTIC of a video. Given that such information does not benefit the student in mastering the learning goal, the validity of the video as an educational tool is fully maintained. For example, take a complex image such as a schematic representation of the human eye, with many labels referring to each individual part of the eye. Labels which are not relevant to the learning goals can be effectively removed, possibly greatly limiting the amount of visual-textual information presented to the student. Evidence for the beneficial effect of removing irrelevant information has been found by several studies and is typically referred to as the 'coherence effect' (Mayer et al., 2001; Moreno & Mayer, 2000; Butcher, 2006).

Presentation rate of visual-textual information. The second proposed component of VTIC is the presentation rate of the visual-textual information. As discussed earlier, working memory is limited in how much information it can hold and process at any given time. Therefore, introducing many concepts simultaneously risks overloading a student with more information than (s)he can effectively handle. This can be prevented by spreading the information over time, while maintaining the same of overall amount of information. For example, detrimental effects of subtitles disappear when verbal and written text explanation are presented *before* the visual information is shown (Moreno & Mayer, 2002b). With such a sequential presentation the student does not need to process the spoken word, written word, as well as the visual information simultaneously. Instead, first the narration and subtitles are processed, and only afterwards the visual information is shown. This effectively removes the role of split-attention effects as well as spreading out cognitive load over time, thus reducing the risk of cognitive overload. However, while this form of information segmentation makes videos easier to process and understand, it also increases the video duration which is often not a desired consequence. Visual-textual information can be segmented without affecting the video duration by only showing new information from the moment it is mentioned in the narration and becomes relevant. Using the previous example of a complex schematic image with many labels: at the start of a video segment the complex image can be shown without any labels, with labels becoming visible from the moment they are verbally discussed. In this format the total duration as well as the narration remain unchanged, while decreasing overall VTIC through a segmentation presentation style.

Split-attention effects. As discussed, subtitles add an additional source of information which needs to be processed, leaving less cognitive resources for learning processes. Additionally, subtitles also draw visual attention, such that less attention is spent on other - possibly important - aspects of the video. Like other cognitive resources, attention is limited. That subtitles can cause a so-called split-attention effect has been made clear by several eye-tracking studies. In general, viewers spend a substantial amount of time paying attention to subtitles (Schmidt-Weigand, Kohnert, & Glowalla, 2010). In a video with a lecturer and subtitles, non-native speakers spend 43% of the time looking at the subtitles (Kruger, Hefer, & Matthew, 2014). The finding that subtitles draw so much attention further signifies their importance. Even when in certain circumstances subtitles are beneficial for learning, it should be taken into account that students will have less attention for other visual information. In situations where subtitles do not significantly aid the learner, a substantial amount of attention will have been wasted. We propose two additional factors contributing to the VTIC of a video.

Attention Cuing. When presented with novel information, it can be difficult to immediately understand where to look. Profound differences in visual search and attention anticipation has been reported for expertise differences in many areas, such as in chess, driving, and clinical reasoning (Chapman & Underwood, 1998; Krupinski et al., 2006; Reingold, Charness, Pomplun, & Stampe, 2001). Given the already high attentional load present in visually complex videos, the presence of subtitles can be expected to have detrimental effects. However, to lower the attentional load, attention can be guided by using attentional cues such as arrows pointing to the most relevant area in a video, or by underling or highlighting these sections. Such attentional cues help novice learners to more effectively direct their attention when and where it is necessary (Boucheix & Lowe, 2010; Ozcelik, Arslan-Ari, &

Cagiltay, 2010), possibly lowering detrimental effects of subtitles.

Physical distances. The final proposed factor of VTIC relates to the physical organization of related information in a video. Specifically, the physical distance between a header (such as a label) and its referent. Non-trivial physical distances between headers and referents are detrimental for learning, as longer distances require more cognitive resources to hold and process information (Mayer, 2008). Additionally, longer distances can induce a split-attention effect, as the increased distances require more attention, which can thus not be spent on other, more relevant parts of the video (Mayer & Moreno, 1998). A split-attention effect can further explain the contradictory findings: subtitles will cause a split-attention in the presence of other visual information, such as graphics, texts, annotated pictures, or diagrams with textual explanations. Furthermore, physical distances can be manipulated to increase or decrease the VTIC without affecting the educational content itself. Using the earlier example of the complex image with labels: the physical distances between the labels and the respectively position in the image can be changed to manipulate the VTIC of a video.

The possible role of language proficiency. It is argued that subtitles (whether L1 or L2) are beneficial for the comprehension of L2 video content because they help students bridge the gap between their language proficiency and the target language (Chung, 1999; Vanderplank, 1988). More specifically, it is often easier to understand L2 written text over spoken word, as reading comprehension skills are typically more developed in students (Danan, 2004; Garza, 1991). Perez et al. (2013) report different learning gains based on L2 proficiency, although this study provides insufficient evidence to verify a moderating role of L2 proficiency. Furthermore, L2 subtitles typically draw more attention than subtitles in one's native language, presumably because L1 subtitles can be processed more automatically and require only peripheral vision (Kruger et al., 2014). A final reason to consider L2 proficiency as an influential factor is because information which is known by a person requires much less, or possible no cognitive resources to operate in working memory (Diana, Reder, Arndt, & Park, 2006; Sweller et al., 1998). As such, processing L2 subtitles can be expected to require less cognitive resources when a student has a higher L2 proficiency. At first sight, this appears to be in conflict with the argument that subtitles specifically help students with a *lower* L2 proficiency to bridge the language barrier. A possible integration of these findings would be that with a lower L2 proficiency subtitles do indeed require more effort to process, but can also aid learning only if no other visual information is present.

Putting the pieces together

Based on the discussed literature, we would argue that to better understand the effects of subtitles on learning it is essential to consider both language proficiency and the complexity of visual-textual information. For example, consider videos showing a teacher explaining a topic simultaneously with a written summary, annotated pictures, or diagrams with textual explanations. The inclusion of subtitles in such videos can be detrimental for learning, especially for students with a low English proficiency. The amount of different visual sources of information will put more strain on the limited capacity of the visual working memory channel. Not only do the subtitles potentially cause a cognitive overload, but also a split-attention effect.

Hypothesis

The main hypothesis of the proposed model is a three-way interaction effect, specifically:

- There is a three-way interaction effect between English proficiency, subtitles, and visual-textual information complexity (VTIC) on test performance. Additionally, we predict specific directions in this three-way interaction:
- For low-VTIC videos, lower English proficiency is related to a higher performance *gain* when subtitles are enabled, and
- For high-VTIC videos, lower English proficiency is related to a higher performance *loss* when subtitles are enabled.

Note that the hypotheses concern relative differences in performance *change*. No claim is made about absolute difference between students with different levels of English proficiency, or between videos with different levels of visual-textual information. The underlying reasoning is that the presented framework predicts different effects of subtitles depending on the amount of additional visual information and level of English proficiency, but it does not necessarily predict absolute differences.

Methods and Materials

Videos

A total of four types of videos have been used; videos with high/low Visual-Textual Information Complexity (VTIC), and with/without subtitles. To ensure ecological validity, actual videos from MOOCs from the Coursera platform have been used as base material; however, to make the videos usable for this experiment they have been extensively edited as further described below. A total of four videos have been used as raw material, which have been manipulated to create the four versions of each video, resulting in 16 videos. To manipulate the complexity of the videos, the four proposed VTIC components have been used as a guideline, as summarized in Table 1. All other video characteristics have been kept the same for each video. The duration of each video is approximately seven minutes, with no differences between the versions of each video. None of the videos in any version show the narrator or teacher. Each video was narrated by the same person to exclude a narrator-effect. In the versions with subtitles, the subtitles will are shown in the bottom part of the screen where it does not overlap with any other content. The narration and subtitles are verbatim identical. The video topics are: The Kidney, History of Genetics, The Visual System, and The Peripheral Nervous System.

English proficiency test

To test the hypothesis of the proposed model it was necessary to estimate the English proficiency of the participants. As the goal of this study is to generate results, which can be easily implemented in online education, a short and easy to implement test was preferred. With this in mind, the English proficiency placement test made by TrackTest has been used (TrackTest, 2016). TrackTest is a placement test which is used to estimate the user's English proficiency at the level of the widely used CEFR scales (Council of Europe, 2001). The CEFR identifies 6 levels, from A1 to C2, signifying beginner to advanced proficiency

Table 1

Overview of the manipulations to create more and less complex versions.

Complexity Factor	High complex version	Low complex version
Irrelevant information	Included	Removed
Segmentation	None	Segmented
Attentional cues	No cues	Cues
Physical distances	Increased	Minimal

levels. The TrackTest is adaptive, meaning that subsequent questions are based on the performance on earlier questions. In total, each participant was presented 15 multiple-choice questions, from a pool of 90. The test takes less than 10 minutes to complete. In a pilot test with 800 users who took the test twice, the test-retest reliability was satisfying with a Spearman's ρ of .736.

Procedure

Upon registration, each participant was randomly allocated to one of four counter-balance lists, which are presented in Table 2. The annotations C- and C+ refer to the video versions with decreased and increased levels visual-textual information complexity, respectively. Likewise, S+ and S- refer to videos with and without English subtitles, respectively. As shown, each participant views one video in each condition. Before the study, the participants were asked to rate their prior knowledge about each of the four topics. For example, regarding the video about the organization of the human eye, the participants were asked how much they know about the different parts and the organization of the human eye. For these questions 5-point Likert scales were used; participants who scored a 3 or higher (e.g. who self-report a moderate amount of prior knowledge) were excluded from the study, to exclude a confounding influence of expertise. The participants were allowed to watch a video only once. After every video, the participants were asked to rate how much mental effort they had to invest to understand the video, on a 9-point Likert scale. The difference in average mental effort ratings for the videos high and low in VTIC serve as a measure of manipulation success. Subsequently, they were presented with a knowledge tests of 10 multiple-choice questions, containing factual questions about the content of the video. After completing the questions participants continued with the next video, until all videos and tests were completed. Afterwards, the participants were asked to take the short English proficiency placement test. Finally, they completed some questions about possible technical issues while watching the videos; these questions were asked for quality assurance. No relevant technical issues were reported.

Participants

As this study focuses on online education, participants were recruited and tested online, using the Prolific platform (Prolific Academic, 2015). Participants were considered eligible when they are over 18 years of age and are non-native English speakers. Upon

Table 2
Counterbalance list.

List	Video 1	Video 2	Video 3	Video 4
1	C+ S+	C- S-	C+ S-	C- S+
2	C- S-	C+ S-	C- S+	C+ S+
3	C+ S-	C- S+	C+ S+	C- S-
4	C- S+	C+ S+	C- S-	C+ S-

Note. C+ and C- refer to more and less complex versions of a video, respectively. S- and S+ refer to the absence and presence of subtitles, respectively.

completion of the study the participants received 6.50 Euro per hour as compensation. Instead of doing a power analysis to a-priori decide on a fixed sample size, the study started with an initial sample size of 50 participants, and sampling continued in batches of 25 until there was sufficient evidence present in the data, as further explained in the next section.

Pre-Registered Analysis Plan

To test the hypothesis, a Bayesian Repeated Measures ANOVA has been performed on the mean test scores with subtitles (yes/no) and visual information (yes/no) as within-subject variables, and English proficiency as between-subject variable (1-5). A Bayesian model comparison has been used to decide on the model with the strongest evidence compared to the other models. This analysis was performed in JASP version 0.7.5 (Love et al., 2015), which uses a default Cauchy prior on effect sizes, centered on 0 with a scaling of 0.707, as argued for by Rouder, Morey, Speckman, and Province (2012). A Bayes Factor of 3 to 10 of one model over another will be interpreted as moderate evidence, 10 to 30 as strong, and above 30 as very strong. This analysis was performed after every batch of participants, and sampling continued until one model had a Bayes Factor of at least 10 compared to every other model. This was the case after 125 participants. All the analyses described below were done using the data from all 125 participants.

Results

First the descriptive statistics will be shown, followed by the confirmatory analysis, and several exploratory analyses. The data and the analysis scripts are available on the Open Science Framework here: https://osf.io/axtgp/?view_only=82dc4b5cd19b4629bca2efa707047b44.

Descriptive Statistics

A total of 125 participants successfully completed the entire study. As is shown in Table 3, the group of participants is well-balanced in terms of gender and age. The language

proficiency is skewed, with 43% of the participants having a high level, but all levels of language proficiency are sufficiently represented in the sample. Note that all participants are non-native English speakers, including the students with the highest proficiency level of C.

Table 3
Descriptive statistics.

Gender	Age	Language Proficiency
67 Male (54%)	Min: 17yr	A1: 14 (11%)
56 Female (46%)	Max: 53yr	A2: 16 (13%)
	Mean: 27.62yr	B1: 23 (18%)
	Sd: 7.24yr	B2: 18 (14%)
		C1-2: 54 (43%)

Notes. Language Proficiency: A1 is the lowest level

In the study, the participants watched four videos, each in a different condition. Table 4 shows the within-subject differences in test scores and self-reported mental effort ratings for each condition pair.

Table 4
Within-subject differences between conditions.

Conditions	Test difference (Sd)	Mental Effort difference (Sd)
Complex: Subs - No Subs	0.10 (2.04)	0.09 (2.06)
Simple: Subs - No Subs	-0.17 (2.10)	-0.06 (1.71)
Subs: Complex - Simple	-0.50 (2.04)	0.32 (2.33)
No Subs: Complex - Simple	-0.77 (2.33)	0.18 (2.00)

These descriptive results give a mixed image. The mean differences between the conditions with the same complexity but subtitles enabled or disabled are the smallest, both for the test scores and mental effort ratings. The differences between conditions with the same setting for subtitles but different levels of complexity are larger, suggesting a main effect of complexity. Furthermore, this difference appears larger when subtitles are disabled, which might mean there is an interaction between complexity and subtitles. Note that Table 4 does not consider a possible main effect or interaction of Language proficiency. The analysis of the full model with all the main effects and interactions is reported below.

Confirmatory Analysis

In accordance with the pre-registered Analysis Plan, a Bayesian Repeated Measures ANOVA was performed on the test scores, with the following predictors: video complex-

ity (high/low), subtitles (yes/no), and the participant's language proficiency (1-5). These results are in the form of model comparison; all the possible combinations of main effects and interactions between the three predictors are compared in terms of how well they can explain the data. Note that in contrast to frequentist ANOVAs, multiple comparions between all models can be performed without the need for corrections. The results of the Bayesian Repeated Measures ANOVA are displayed in Table 5, which shows all the models in descending order of evidence.

Table 5
Bayes Factors of all models relative to the null model.

#	Model	BF(M, 0)	% error	BF(M, M+1)	P(M|D)
1	C + L	99980000.00	1.48%	10.30	0.868
2	C + S + L	9704000.00	1.35%	3.98	0.084
3	C + L + C*L	2439000.00	1.44%	1.17	0.021
4	C + S + L + C*S	2086000.00	2.51%	3.98	0.018
5	C + S + L + S*L	524242.34	1.67%	2.10	0.005
6	C + S + L + C*L	250015.78	1.98%	2.09	0.002
7	C + S + L + C*S + S*L	119737.40	2.77%	1.93	0.001
8	L	62108.53	0.62%	1.17	<0.001
9	C + S + L + C*S + C*L	52917.54	2.37%	3.97	<0.001
10	C + S + L + C*L + S*L	13331.75	2.19%	2.18	<0.001
11	S + L	6125.66	1.06%	2.17	<0.001
12	C + S + L + C*S + C*L + S*L	2818.57	2.33%	1.31	<0.001
13	C	1747.43	8.04%	5.43	<0.001
14	S + L + S*L	321.62	1.64%	1.29	<0.001
15	C + S + L + C*S + C*L + S*L + C*S*L	249.57	2.80%	1.61	<0.001
16	C + S	155.37	1.26%	3.97	<0.001
17	C + S + C*S	39.09	13.77%	39.09	<0.001
0	Null (intercept + subject)	1.00	n/a	10.10	<0.001
18	S	0.10	1.13%	n/a	<0.001

Note. C = Complexity (high/low). S = Subtitles (yes/no). L = Language Proficiency (1-5). BF(M, 0) = Bayes Factor of Model compared to Null Model. BF(M, M+1) = Bayes Factor of Model compared to next model. P(M|D) = Posterior probability of model given the data, if each model had equal probability of being true before this study.

The results show that model 1 has the most evidence, which consists only of the main effects of complexity and language proficiency, no main effect of subtitles, and no interactions

between any of the factors. This model has nearly 10^8 times more evidence than the null model. Importantly, the evidence provided by this study favors the complexity + language proficiency model over complexity + subtitles + language proficiency model (which is the 2nd best model) by a factor of $10.30 : 1$. In other words, there is 10.3 times more evidence for the C + L model than the C + S + L model. Furthermore, every model which does not contain a main effect of subtitles is stronger than its counterpart which includes an effect of Subtitles.

The pre-registered hypothesis was that the data would be best explained by the full three-way interaction model (model 15). While there is more evidence for this model than for a null model, the data favors the simpler C + L model by a factor of $400,000 : 1$.

In the last column of Table 5, the posterior probability of each model is shown. When considering only these models, and having no preference for any model before the study, the P(M|D) gives the probability that the model is true, given the data and priors.

Exploratory Analyses

While the Bayes Factors quantify the amount of relative evidence provided for the models, it does not provide information about estimations of population parameters such as means and effect sizes. Using Markov Chain Monte Carlo (MCMC) methods from the BayesFactor Package in R, we estimated population parameters using all the available data with all the factors and their interactions (Morey & Rouder, 2015; R Core Team, 2016). Chains were constructed with 10^6 iterations; Visual inspection of the chains and auto-correlation plots revealed no quality issues. We put Cauchy priors on the effect size parameters with a scaling factor of $1/2$, as further described in Rouder et al. (2012). A Cauchy with a scaling factor of $1/2$ has half the probability mass between -0.5 and 0.5, and the remaining half on the more extreme values. In other words, we expect effect sizes of around (-)0.5, but the prior is diffuse enough to be sensitive to more extreme effects. Using much wider or more narrow scaling factors (from $1/6$ to 4) does not affect the estimations in a consequential manner. As these priors cover all the effect sizes in the discussed literature, we consider the results to be insensitive to all plausible alternative priors. Complexity and subtitles were entered as factors (yes/no), while language was entered as a continuous variable (1-5). This analysis was done separately for effects on test scores (described in the next section) and on mental effort ratings (described in the subsequent section).

Effects on test scores. A visualization of the the posterior probability densities of the three main effects on test scores is shown in Figure 1. While Figure 1 shows only the main effects, the entire model with all factors and interactions were used to generate these posterior distributions.

The density plot shows the most likely values of each effect size parameter, such that any point in the plot which is twice as high as another point is twice as likely. Note how the effect of subtitles is centered around 0, with higher and lower values becoming increasingly unlikely. In contrast, the effect size of complexity is much stronger, and most likely to be around -0.62 (compared to simple). Language has a positive effect, with an effect size slope of around 0.55. Note that these are unstandardised effect sizes measures in grade points, on a scale of 0 to 10. In other words, while the effect of subtitles is mostly likely to be (close to) zero, both complexity and language have a noticeable effect on test scores. Compared to complexity and subtitles, the effect of language proficiency can be estimated with relatively
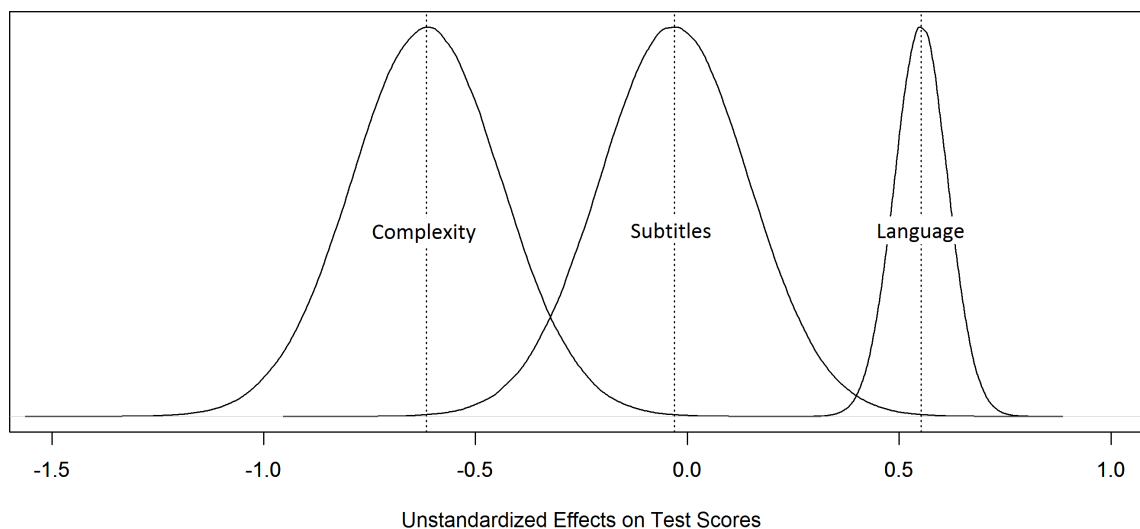
Figure 1. Posterior probability density plots for the effects on test score (1-10)

little uncertainty. The parameter estimations of the main effects and all interactions are shown in Table 6.

As can be seen in Table 6, the difference between two identical videos, which only differ in complexity, is 0.62 grade points (as the difference between low and high complexity is 0.62). Dividing this by the standard deviation results in a Cohen's *d* effect size of 0.31. The slope of language proficiency (measured on a scale of 1 to 5) is 0.55 (Cohen's *d* of 0.27), with a 95% Credible Interval of 0.43 to 0.68. However, the effect of subtitles is 0.04 grade points (Cohen's *d* of 0.02), and we cannot even be certain about the direction of the effect as the credible intervals span both negative and positive values. This means that it is very likely to be (close to) zero. All the interaction effects are similarly centered around 0, with credible intervals which span both negative and positive values. These findings are fully consistent with the Confirmatory Analysis, which suggested that the best model includes only the main effects of complexity and language proficiency, but not the effect of subtitles or any of the interactions. Only complexity and language proficiency have 95% credible intervals which do not include zero, such that we can be confident about the direction of the effect, while the effects of the other factors are close to zero.

Effects on mental effort ratings. In addition to the effects on test scores, we analyzed the effects of complexity, language proficiency, and subtitles on the participants' self-reported mental effort ratings of the videos. This analysis is identical to the previous analysis in every aspect other than the different outcome variable. As described earlier, the participants were asked how much mental effort they had to invest in watching and understanding each video on a 9-point Likert scale, higher meaning more invested effort. A visualization of the the posterior probability densities of the three main effects on mental effort ratings is shown in Figure 2.

As can be seen in Figure 2, the effects of subtitles and language proficiency on the participants' mental effort ratings are both centered near 0. For subtitles, the effect is estimated at a 0.015 difference in mental effort ratings, 95% Credible Interval $[-0.16, 0.19]$. The effect of language proficiency is estimated at 0.017, 95% Credible Interval $[-0.10,$

Table 6

Parameter estimations of intercept and factor effects.

Parameter	Estimation	95% Credible Interval
Intercept	4.81	[4.63, 4.98]
Standard Deviation	2.00	[1.88, 2.13]
C0	0.31	[0.13, 0.48]
C1	-0.31	[-0.48, 0.13]
S0	0.02	[-0.16, 0.19]
S1	-0.02	[-0.19, 0.16]
Lang	0.55	[0.43, 0.68]
C0 * S0	0.06	[-0.11, 0.24]
C0 * S1	-0.06	[-0.24, 0.11]
C1 * S0	-0.06	[-0.24, 0.11]
C1 * S1	0.06	[-0.11, 0.24]
C0 * Lang	0.03	[-0.10, 0.15]
C1 * Lang	-0.03	[-0.15, 0.10]
S0 * Lang	-0.04	[-0.17, 0.08]
S1 * Lang	0.04	[-0.08, 0.17]
C0 * S0 * Lang	-0.05	[-0.18, 0.07]
C0 * S1 * Lang	0.05	[-0.07, 0.18]
C1 * S0 * Lang	0.05	[-0.07, 0.18]
C1 * S1 * Lang	-0.05	[-0.18, 0.07]

Notes. C = Complexity (1 = high, 0 = low).

S = Subtitles (1 = yes, 0 = no).

Lang = Language Proficiency (slope).

0.14]. Of the three effects, complexity is the only one not centered around zero, and is estimated at 0.24, 95% Credible Interval [0.06, 0.41]. When transformed into standardized Cohen's *d* effect sizes, the effect of subtitles is 0.008, for language proficiency it is 0.010, and complexity has an effect of 0.120. The (unstandardized) effects of the interactions are all smaller than 0.02, which - on a scale of 1 to 10 - is so small they will not be further discussed.

Discussion

Open online education plays an important role in the globalization and democratization of education. To ensure not only the *availability*, but also the *accessibility* of open online
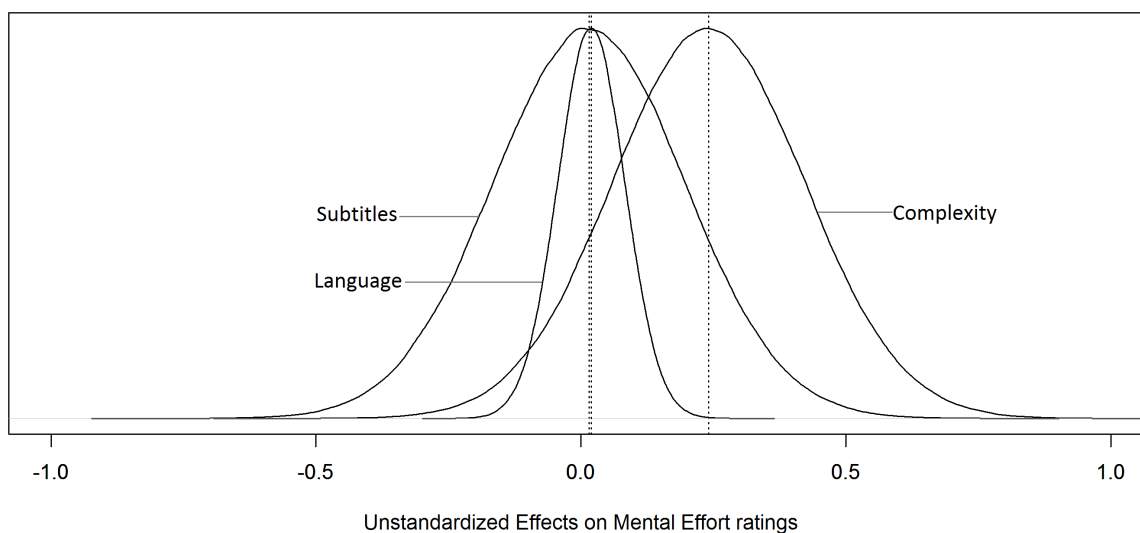
Figure 2. Posterior probability density plots for the effects on mental effort ratings (1-10)

education, it is vital to remove potential obstacles and biases which put certain students at a disadvantage, for example students with lower levels of English proficiency. This is not yet a given, as Massive Open Online Courses (MOOCs) are still provided primarily in English. In this study, we investigated whether the presence of English subtitles has beneficial, or possibly detrimental effects on students' understanding of the content of English videos. Specifically, we tested the hypothesis that the effect of subtitles on learning depends on the English proficiency of the students and the Visual-Textual Information Complexity (VTIC) of the video. Contrary to this hypothesis, we found strong evidence that there is no main effect of subtitles on learning, nor any interaction, but only a main effect of complexity and language proficiency. We will discuss these findings in that order.

No main effect of subtitles

Contrary to a range of previous studies, we found strong evidence that subtitles neither have a beneficial nor detrimental effect on learning from educational videos. In addition, the presence or absence of subtitles also appear to have no effect on self-reported mental effort ratings. This is surprising given an apparent consensus that enabling subtitles increases the general accessibility of online content, as is stated by the Web Content Accessibility Guidelines 2.0 (WCAG, 2008). These null findings contradict two lines of research, one showing beneficial effects of subtitles, the other showing detrimental effects.
Earlier research which has shown beneficial effects of subtitles are primarily studies on second language learning, which show that 2nd language subtitles help students with learning that language (e.g., Baltova, 1999; Chung, 1999; Markham, 1999; Winke et al., 2013). While this appears conflicting with the results of the present study, the important difference is that the current study did not use language learning videos but 'content' videos, and did not measure gains in 2nd language proficiency. Based on the current study it seems that for content videos there is little to no benefit of enabling subtitles, even for students with a low language proficiency and for visually complex videos.

A different body of research has shown detrimental effects of subtitles. This is often labeled the Redundancy Effect, as the reasoning is that because the subtitles are verbatim identical to the narration they are redundant and can only hinder the learning process (e.g., Mayer et al., 2001, 2003). This is in clear contrast with the findings of the current study, which estimates the effect of subtitles to be (close to) zero. Importantly, the English language proficiency of the students did not moderate the effect of subtitles, even though the study included participants with the full range of English proficiency levels. As noted before, it might be possible that the subtitles helped the students with lower proficiency levels to increase their understanding of English, but it did not affect their test performance. With the Bayesian analyses we showed that subtitles do not merely have a non-distinguishable effect (e.g., a non-significant effect in frequentist statistics) but that there is strong evidence for the absence of a subtitle effect on learning and mental effort. While these conclusions are only based on the selection of videos used in the current study, it puts the generalizability of the Redundancy Effect in question by showing that it does not hold for these specific videos, but arguably also for a wider range of similar videos. More research is needed to further establish the potential (lack of) effects of subtitles on learning from videos; both in highly controlled settings as well as in real life educational settings. Specifically, it is essential to study the generalizability of findings like the Redundacy Effect, and establish boundary conditions. Even though the current study used four different videos, each with four different versions, this is not sufficient to be able to generalize to all kinds of educational videos. However, by manipulating the complexity of the videos, we were able to show that the null effect of subtitles cannot be explained by complexity or element interactivity (Paas, Renkl, & Sweller, 2003; Sweller, 1999). Furthermore, we compared the amount of evidence for a wide range of different models and found that every model which does not include a main effect of subtitles is stronger than its respective alternative model which does include subtitles. In addition, the within-subject design of the study severely reduces the plausibility of confounding participant characteristics. Finally, it is noteworthy that the current study only used 2nd language subtitles, meaning that providing subtitles in the native language of students can still have an positive effect on learning and accessibility (Hayati & Mohmedi, 2011; Markham et al., 2001).

Main effect of complexity

The effect of video complexity shows how video design can have a noticeable effect on test performance, either positively or negatively. In this study, the effect was estimated at 0.62 grade points (on a scale of 0-10), which translates to a Cohen's *d* of 0.31. In addition, the self-reported mental effort ratings was 0.24 higher for complex videos (on a scale of 1-10), which is a Cohen's *d* of 0.12. As the quizzes took place immediately after each video, the current study only provides insight in how visual-textual information complexity affects short-term performance on tests. Effects on long-term learning are unknown, but it is plausible that the performance gap remains stable, or even worsens as the test delay increases, as initial (test) performance typically strongly predicts future (test) performance (e.g., Gow et al., 2011; Harackiewicz, Barron, Tauer, Carter, & Elliot, 2000; Karpicke & Roediger, 2007). Furthermore, the current study used individual videos while most online courses have multiple related videos which build on each other. Whether such inter-video dependency strengthens or weakens the effect of visual-textual information complexity is

yet unknown, but warrants further investigation.

In this study, the complexity of the videos was manipulated based on four principles extracted from the literature on multimedia learning. These are the Segmentation Effect, the Signaling effect, the Spatial Contingency Effect, and the Coherence Effect, all of which are further explained and discussed in the Introduction, as well as by Mayer and Moreno (2003). This resulted in two different versions of each video which differ only in the (mainly visual) complexity of the presentation of information. While the mentioned manipulations have each been investigated independently, this is - to the best of our knowledge - the first study which combined all four to experimentally manipulate the complexity of videos. Surprisingly, while the individual manipulations had effect sizes ranging from Cohen's d's of 0.48 to 1.36, the combined effect is estimated at a Cohen's d of 0.31. We note several plausible interpretations for this discrepancy; the effect of the manipulations varies with 1) video characteristics, 2) student characteristics, 3) different implementations, and/or varies with 4) study design. First, while the current study used multiple videos and different versions of each video, a moderating effect of video characteristics cannot be ruled out. For example, the size of the effect might partly depend on characteristics such as the video's length, educational content, or other aspects which were not manipulated in this study. Should this be the case, this would mean that the generalizability of the four effects are limited by these moderating variables. Secondly, characteristics of the students in the different studies might partly explain the discrepancy in effect sizes. While many of the cited studies used the relatively homogeneous sub-population of Psychology students, the current study used participants from various countries, with varying levels of education as well as levels of English proficiency. Given the wider and less selective range of participants, one would typically expect a more accurate estimation of the size and generalizability of the studied effects. Furthermore, given the within-subject design of the current study, it seems unlikely that potentially relevant participant characteristics confounded the results, which would be more likely in a between-subject design. Thirdly, it is important to note that the current study necessarily employed a specific operationalization of the four effects. For example, there are many ways how one could operationalize the Signaling Effect by using attentional cues of different kinds, such as underlining, highlighting, or different kinds of arrows or circles. Given the wide range of possible operationalizations, it is to be expected to see variation in the effects of these manipulations. While this is likely to be of influence, it remains unclear whether it is a sufficient explanation. Finally, the fourth potential explanation of the difference in effect sizes is based in differences in study design and methodology. For example, the current study took place online, and not in an physical location such as a university. Another potential explanation is the way how Cohen's d is calculated, as well as different estimations of the standard deviation, or other choices in statistical procedures which can differ between studies (Baguley, 2009).

In sum, while there are many plausible reasons for the differences in effect sizes, it remains unclear what the exact causes are, and whether these are systematic or due to random variation. This further emphasizes the need to study these instructional design guidelines for videos using a wide range of videos, in different educational contexts, and with a representative sample of participants. While it is unrealistic to expect to be able to accurately predict the effect size of such manipulations with great precision across many different situations, it is paramount to better understand moderating variables and boundary conditions to make

better recommendations of how to make high quality educational videos.

Main effect of language proficiency

Students with a higher English language proficiency scored substantially higher than students with a lower proficiency. The slope of this effect was estimated to be 0.55 grade points. Given the range of language proficiency of 1-5, the grade point difference between the students with the highest and lowest proficiency levels will be over 2.5 grade points. This further signifies the issue that open online courses such as MOOCs are not equally accessible to everyone, as the majority of the courses are provided in English. By extension, this calls for research on investigating interventions or design strategies which might help close this performance gap. However, it is important to mention that the design of the current study does not directly translate to how non-native English speakers engage with online courses. For example, the participants in this study were not allowed to re-watch or pause videos, take notes, or use any other strategy which might be particularly helpful for non-native speakers in online courses. Students who experience trouble with understanding a video might choose to use such strategies to counteract their initial disadvantage. However, it is also plausible that non-native English speakers are put off by the mainly English online courses, and choose to not engage at all, or drop-out early in such courses, which should be prevented.

Summary and consequences for practice

To summarize, the visual-textual information of a video and especially the language ability of the student are both strong predictors of learning from content videos. In contrast, English subtitles neither increased nor decreased the student's ability to learn from the videos. However, this does not lead to the conclusion that English subtitles should not be made available, as they are vital for students with hearing disabilities. Furthermore, students might prefer watching videos with subtitles for other reasons, even though this might not directly affect their learning. The extent to which subtitles in the students' native language might help them cope with lacking English proficiency is as of yet unknown, and remains to be investigated. Another possibility would be to provide dubbed versions of each video to cater to more languages, but this is a costly intervention. Overall, we have shown that both the language proficiency as well as the video's complexity can have a substantial effect on learning from educational videos, which deserves attention in order to increase the quality and accessibility of open online education.

References

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839. doi: 10.1038/nrn1201

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617. doi: 10.1348/000712608X377117

Baltova, I. (1999). Multisensory language teaching in a multidimensional curriculum: The use of authentic bimodal video in core french. *Canadian Modern Language Review*, *56*(1), 31–48. doi: 10.3138/cmlr.56.1.31

Boucheix, J.-M., & Lowe, R. K. (2010). An eye tracking comparison of external pointing cues and internal continuous cues in learning with complex animations. *Learning and Instruction*, *20*(2), 123–135. doi: 10.1016/j.learninstruc.2009.02.015

Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology*, *98*(1), 182–197. doi: 0.1037/0022-0663.98.1.182

Campbell, J., Gibbs, A. L., Najafi, H., & Severinski, C. (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), 235–262. Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/1854

Chapman, P. R., & Underwood, G. (1998). Visual search of driving situations: Danger and experience. *Perception*, *27*(8), 951–964. doi: 10.1068/p270951

Chung, J. M. (1999). The effects of using video texts supported with advance organizers and captions on chinese college students' listening comprehension: An empirical study. *Foreign Language Annals*, *32*(3), 295–308. doi: 10.1111/j.1944-9720.1999.tb01342.x

Council of Europe. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment.* Cambridge: Cambridge University Press.

Danan, M. (2004). Captioning and subtitling: Undervalued language learning strategies. *Meta*, *49*(1), 67-77. doi: 10.7202/009021ar

Diana, R. A., Reder, L. M., Arndt, J., & Park, H. (2006). Models of recognition: A review of arguments in favor of a dual-process account. *Psychonomic Bulletin & Review*, *13*(1), 1–21. doi: 10.3758/bf03193807

Garza, T. J. (1991). Evaluating the use of captioned video materials in advanced foreign language learning. *Foreign Language Annals*, *24*(3), 239–258. doi: 10.1111/j.1944-9720.1991.tb00469.x

Ginns, P. (2006). Integrating information: A meta-analysis of the spatial contiguity and temporal contiguity effects. *Learning and Instruction*, *16*(6), 511–525. doi: 10.1016/j.learninstruc.2006.10.001

Gow, A. J., Johnson, W., Pattie, A., Brett, C. E., Roberts, B., Starr, J. M., & Deary, I. J. (2011). Stability and change in intelligence from age 11 to ages 70, 79, and 87: the lothian birth cohorts of 1921 and 1936. *Psychology and Aging*, *26*(1), 232. doi: 10.1037/a0021072

Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement: An empirical study of mooc videos. In *Proceedings of the first acm conference on learning scale conference* (pp. 41–50). doi: 10.1145/2556325.2566239

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., Carter, S. M., & Elliot, A. J. (2000). Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *Journal of Educational Psychology*, *92*(2), 316. doi: 10.1O37//0022-O663.92.2.316

Harskamp, E. G., Mayer, R. E., & Suhre, C. (2007). Does the modality principle for multimedia learning apply to science classrooms? *Learning and Instruction*, *17*(5), 465–477. doi: 10.1016/j.learninstruc.2007.09.010

Hayati, A., & Mohmedi, F. (2011). The effect of films with and without subtitles on listening comprehension of EFL learners. *British Journal of Educational Technology*, *42*(1), 181–192. doi: 10.1111/j.1467-8535.2009.01004.x

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology*, *13*(4), 351–371. doi: 10.1002/(SICI)1099-0720(199908)13:4<351::AID-ACP589>3.0.CO;2-6

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151–162. doi: 10.1016/j.jml.2006.09.004

Kruger, J., Hefer, E., & Matthew, G. (2014). Attention distribution and cognitive load in a subtitled academic lecture: L1 vs l2. *Journal of Eye Movement Research*, *7*(5), 1-15. doi: 10.16910/jemr.7.5.4

Krupinski, E. A., Tillack, A. A., Richter, L., Henderson, J. T., Bhattacharyya, A. K., Scott, K. M., . . . Weinstein, R. S. (2006). Eye-movement study and human performance using telepathology virtual slides. implications for medical education and differences with experience. *Human Pathology*, *37*(12), 1543–1556. doi: 10.1016/j.humpath.2006.08.024

Liu, Y., Liu, M., Kang, J., Cao, M., Lim, M., Ko, Y., . . . others (2013). Educational paradigm shift in the 21st century e-learning. In *E-learn: World conference on e-learning in corporate, government, healthcare, and higher education* (Vol. 2013, pp. 373–379).

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., . . . Wagenmakers, E. J. (2015). *JASP.* https://jasp-stats.org/.

Markham, P. (1999). Captioned videotapes and second-language listening word recognition. *Foreign Language Annals*, *32*(3), 321–328. doi: 10.1111/j.1944-9720.1999.tb01344.x

Markham, P., Peter, L. A., McCarthy, T. J., et al. (2001). The effects of native language vs. target language captions on foreign language students' dvd video comprehension. *Foreign Language Annals*, *34*(5), 439–445. doi: 10.1111/j.1944-9720.2001.tb02083.x

Mayer, R. E. (2003). The promise of multimedia learning: using the same instructional design methods across different media. *Learning and Instruction*, *13*(2), 125–139. doi: 10.1016/s0959-4752(02)00016-6

Mayer, R. E. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *American Psychologist*, *63*(8), 760–769. doi: 10.1037/0003-066x.63.8.760

Mayer, R. E., Dow, G. T., & Mayer, S. (2003). Multimedia learning in an interactive self-explaining environment: What works in the design of agent-based microworlds? *Journal of Educational Psychology*, *95*(4), 806–812. doi: 10.1037/0022-0663.95.4.806

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology*, *93*(1), 187–198. doi: 10.1037/0022-0663.93.1.187

Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology*, *90*(2), 312–320. doi: 10.1037/0022-0663.90.2.312

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 43–52. doi: 10.1207/S15326985EP3801$_6$

Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology*, *91*(2), 358–368. doi: 10.1037/0022-0663.91.2.358

Moreno, R., & Mayer, R. E. (2000). A coherence effect in multimedia learning: The case for minimizing irrelevant sounds in the design of multimedia instructional messages. *Journal of Educational Psychology*, *92*(1), 117. doi: 0.1037//0022-0663.92.1.117

Moreno, R., & Mayer, R. E. (2002a). Learning science in virtual reality multimedia environments: Role of methods and media. *Journal of Educational Psychology*, *94*(3), 598–610. doi: 10.1037/0022-0663.94.3.598

Moreno, R., & Mayer, R. E. (2002b). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology*, *94*(1), 156–163. doi: 10.1037/0022-0663.94.1.156

Morey, R. D., & Rouder, J. N. (2015). Bayesfactor: Computation of bayes factors for common designs [Computer software manual]. doi: 10.1016/j.jmp.2012.08.001

Nesterko, S. O., Dotsenko, S., Han, Q., Seaton, D., Reich, J., Chuang, I., & Ho, A. (2013). Evaluating the geographic data in moocs. In *Neural information processing systems.* Retrieved from `http://nesterko.com/files/papers/nips2013-nesterko.pdf`

Ozcelik, E., Arslan-Ari, I., & Cagiltay, K. (2010). Why does signaling enhance multimedia learning? evidence from eye movements. *Computers in Human Behavior*, *26*(1), 110–117. doi: 10.1016/j.chb.2009.09.001

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1–4. doi: $10.1207/s15326985ep3801_1$

Perez, M. M., Noortgate, W. V. D., & Desmet, P. (2013). Captioned video for l2 listening and vocabulary learning: A meta-analysis. *System*, *41*(3), 720–739. doi: 10.1016/j.system.2013.07.013

Prolific Academic. (2015). *Prolific academic.* Retrieved 2016-01-05, from `https://prolificacademic.co.uk`

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Reingold, E. M., Charness, N., Pomplun, M., & Stampe, D. M. (2001). Visual span in expert chess players: Evidence from eye movements. *Psychological Science*, *12*(1), 48–55. doi: 10.1111/1467-9280.00309

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. doi: 10.1016/j.jmp.2012.08.001

Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010). A closer look at split visual attention in system- and self-paced instruction in multimedia learning. *Learning and Instruction*, *20*(2), 100–110. doi: 10.1016/j.learninstruc.2009.02.011

Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P., & Pritchard, D. E. (2014). Who does what in a massive open online course? *Communications of the ACM*, *57*(4), 58–65. doi: 10.1145/2500876

Sweller, J. (1999). Instructional design. In *Australian educational review.*

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*(2), 123–138. doi: 10.1007/s10648-010-9128-5

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. doi:

10.1023/A:1022193728205

TrackTest. (2016). *Tracktest english.* Retrieved 2016-01-05, from `http://tracktest.eu/`

Vanderplank, R. (1988). The value of teletext sub-titles in language learning. *ELT Journal*, *42*(4), 272–281. doi: 10.1093/elt/42.4.272

WCAG. (2008). *WCAG 2.0 web content accessibility guidelines (wcag) 2.0.* Retrieved 2016-01-05, from `http://www.w3.org/WAI/WCAG20/glance/`

Winke, P., Gass, S., & Sydorenko, T. (2013). Factors influencing the use of captions by foreign language learners: An eye-tracking study. *The Modern Language Journal*, *97*(1), 254–275. doi: 10.1111/j.1540-4781.2013.01432.x