Continuous Evaluation of Video Lectures from Real-Time Difficulty Self-Report

Namrata Srivastava
The University of Melbourne
Melbourne, VIC
srivastavan@student.unimelb.edu.au

Eduardo Velloso
The University of Melbourne
Melbourne, VIC
eduardo.velloso@unimelb.edu.au

Jason M. Lodge
The University of Queensland
Brisbane, QLD
jason.lodge@uq.edu.au

Sarah Erfani
The University of Melbourne
Melbourne, VIC
sarah.erfani@unimelb.edu.au

James Bailey
The University of Melbourne
Melbourne, VIC
baileyj@unimelb.edu.au

ABSTRACT

With the increased reach and impact of video lectures, it is crucial to understand how they are experienced. Whereas previous studies typically present questionnaires at the end of the lecture, they fail to capture students' experience in enough granularity. In this paper we propose recording the lecture difficulty in real-time with a physical slider, enabling continuous and fine-grained analysis of the learning experience. We evaluated our approach in a study with 100 participants viewing two variants of two short lectures. We demonstrate that our approach helps us paint a more complete picture of the learning experience. Our analysis has design implications for instructors, providing them with a method that helps them compare their expectations with students' beliefs about the lectures and to better understand the specific effects of different instructional design decisions.

CCS CONCEPTS

• **Applied computing** → **E-learning**; *Interactive learning environments.*

KEYWORDS

Audio-visual instruction, e-Learning, Video Lectures, Education

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300816

ACM Reference Format:
Namrata Srivastava, Eduardo Velloso, Jason M. Lodge, Sarah Erfani, and James Bailey. 2019. Continuous Evaluation of Video Lectures from Real-Time Difficulty Self-Report. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300816

1 INTRODUCTION

Online videos have become an important medium for the delivery of video lectures in higher education [3]. They are the main information delivery tool in online learning environment such as Massive Open Online Courses (MOOCs) [22], and also play a complementary role in traditional, blended, and flipped classrooms [21, 37]. Existing research has shown that video lectures are effective educational tools [8, 14, 27], but video production decisions such as length, speaking rate, video type and production style substantially affect students' engagement and learning experiences [10, 13, 16]. Therefore, devising methods and tools that are able to tease out these effects in detail is an important driver in e-learning research.

Standard measures for video assessment are mostly focused on the use of questionnaires *after* the lecture for capturing learners' subjective feedback on affect, mental-effort, perceived learning, and user preferences [10, 16, 17, 23]. Alternatives include recent attempts that measured these effects at periodic intervals during interactive tutorials [2, 24]. Though these instruments are effective in capturing overall sentiments and reactions, they do not provide enough granularity to conduct detailed analyses on how specific parts of the lecture affect the learning experience.

To address the limitations of traditional one-off and periodical self-reports, we propose capturing continuous subjective reactions in real-time throughout the lecture with the use of a physical slider. The idea is inspired by Tangen et al. [36] who measured learners' interest with a 5-point scale that could be used to rate the segment of the lecture whenever their interest on the content changed. This approach has

been widely used in market research to gauge an audience's reactions to visual stimuli over time. Other approaches for extracting continuous signals from video experiences build a time series based on collective interactions with the video player [9, 11]. However, whereas these approaches build a continuous signal by aggregating the number of times a particular segment is viewed and reviewed across all viewers, we collect an individual signal for each participant.

Real-time continuous self-report measures have been successfully used to measure human emotions with sliders [19] or digital reports [4]). However, to the best of our knowledge no work has explored the use of this tool in the educational domain. Further, there is little advice on how to analyse such continuous data in order to obtain actionable insights regarding the instructional and production design of the lecture. We aim to address this gap by collecting a large dataset with different lecture topics and instructional designs continuously rated by participants and by demonstrating how to use time series analysis techniques to generate insights. The primary focus of the study is to demonstrate the usefulness of the slider tool. The issues of media designs are the secondary focus of our study.

We evaluated our approach in an experiment with 100 participants who watched two variants of two short video lectures on different topics while continuously reporting the difficulty level of the current point in time using a physical slider. Additionally, participants were also asked to report their experience after the lecture using standard evaluation instruments and to answer pre- and post-tests.

We demonstrate the usefulness of continuous analysis of video lectures based on real-time subjective ratings in five ways:

(1) we analyse peaks and troughs in the signal to identify points of difficulty in the lecture
(2) we demonstrate how local ratings of particular segments relate to the performance in the post-hoc test questions corresponding to those segments
(3) we compare instructors' continuous ratings to learners' ratings to identify misalignments of expectations
(4) we compare synchronised lecture variations to identify points of divergence
(5) we compare non-synchronised lecture variations to assess the effect of differences in instructional design techniques

In the remainder of the paper, we discuss related work and our experiment design and method. We then report results from the *within-video* and *between-video* time-series analysis to demonstrate the effectiveness of our approach. Finally, we present design implications for instructors for designing effective learning tools, and conclude with limitations and future work.

2 RELATED WORK

Existing research involving video analysis mainly focused on the sporadic use of video lectures and investigation of macro-level video activity features such as number of videos watched [1], amount of time spend on a video and problem attempts [5, 13], students' dwelling time [38] as well as navigation styles [13]. For instance, Guo et al. [13] compared 6.9 million video watching sessions across four courses to measure effects of video production decisions such as length, speaking rate, video type and production type on students' engagement. Similarly, in another study, Lau et al. [18] use trends in view count such as total percentage of video viewed and audience retention (percentage of viewers watching at a time point compared to the initial total) for comprehensive evaluation of ten medical-science lectures. The problem we have with these macro-level techniques is that they are rather rough judgements about students' learning experiences as they contain data about entire videos, not *within* videos.

An alternative approach is to scale the analytics down to the click-level by closely examine how student interacts with each video lecture. Compared to macro-level video activity features, click-level-in-video analysis allows the instructors to identify which specific parts students are watching, skipping, re-watching or speeding. Such models were then used for predicting students' dropout [15], identifying difficult segments in the lecture [20], measuring students' attitudes [12] and improving user navigation experience [9]. This also does not help to evaluate videos at a fine-grained level as click-level interaction doesn't necessarily represent learners' real intent (e.g., play a video but not watch) and can be dependent on other pedagogical methods such as problem sets and exams.

Some approaches use human intervention to identify where the important segments are located in the video lecture. Previous research used user rating data [28] or lecture annotations tools [31, 33].For instance, Olsen and Moon [28] summarised sports videos to identify which plays were most interesting based on user rating of the plays, whereas CLAS [31] is a lecture video annotation tool where students click a button when they find a lecture segment important. Although, the lecture annotation tools (CLAS) provides a collaborative tool to identify the important concepts in the lecture, however, they fail to answer the reasoning why the segments were important.

To counterbalance the advantages and disadvantages of both the micro-level analysis and human intervention for video assessment, in this paper we propose to combine the two by recording students' subjective ratings about the lecture difficulty throughout the session using a slider. Most

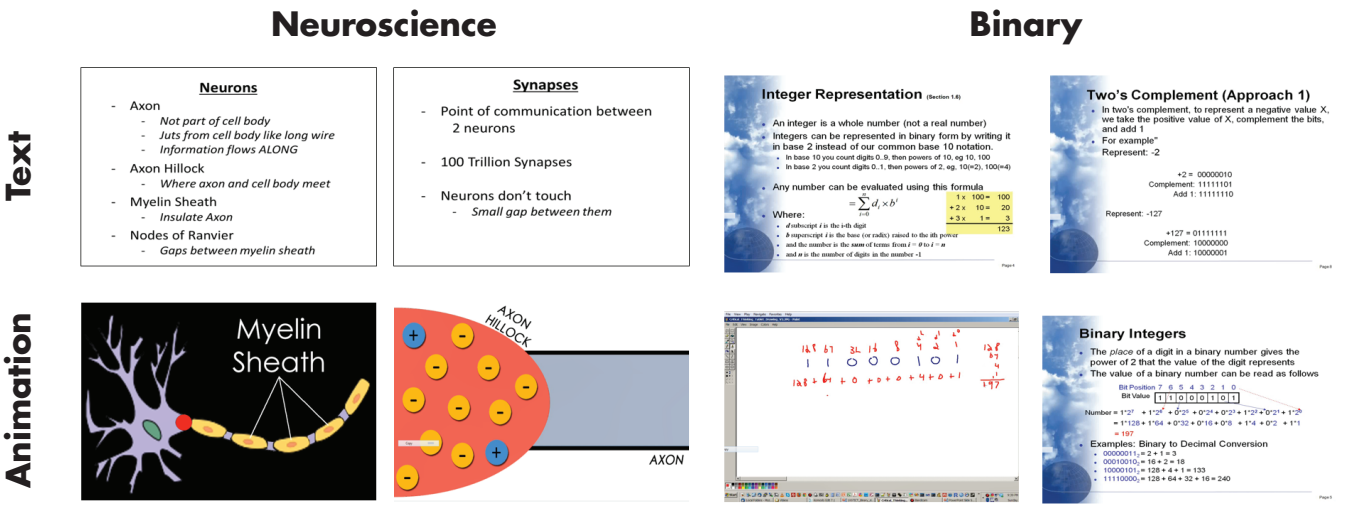


Figure 1: Sample frames from each version of the lectures. Both versions of the Neuroscience lecture had the same audio track. The Binary lecture combined slides and digital ink, and had different audio tracks.

of the previous works have focused on assessing video lectures by using self-reported questionnaires after the lecture, measuring students' mental effort, affect, perceived learning and preferences. [10, 16, 17, 23]. A few recent attempts were found in the literature of intelligent tutoring systems and interactive tutorials, where researchers record students' interaction at periodic interval of times [2, 24, 32]. For instance, Lodge and Kennedy [24] used periodic measurements of level of confidence and challenge to evaluate the effect of prior knowledge on students' confidence and perceived difficulty in a digital learning environment. In another study, Shin et. al. [34] explores instructor and learner perceptions of in-video prompting where learners answer reflective questions while watching videos. These traditional one-off and periodic self-reports although capture students' overall sentiments and affects, they do not provide enough granularity to conduct micro-level analysis of lecture segments.

Our work is inspired from Tangen et al. [36] where they investigated role of interest in different slideware presentation conditions by measuring the students' interest level throughout the video presentation. The participants were asked to indicate any changes in their interest level by pressing the up and down arrows on the keyboard, such that each stroked toggled a 5-point scroll-bar on the screen. However, instead of using a key-press for recording students' interactions, we choose to use a slider in order to extract continuous time-series signals. Previous work has shown the usefulness of slider in measuring human emotions [19], however to the best of our knowledge, no work has explored the tool in educational domain.

3 EXPERIMENT

With the goal of exploring the use of real-time self-reported ratings for enabling continuous analysis of video lectures, we conducted an experiment. Our experiment is framed under the overarching theme of instructional design, but aimed at comparing two data collection and analysis methods. We collected data on two types of instructional design (between-participants)—text and animation—in two lecture topics (within-participants)—Neuroscience and Maths (see Figure 1). By collecting both *post-hoc* and *real-time* subjective ratings (within-participants) we were able to not only derive findings about instructional design, but also about these methods. As such, we conduct the following analyses on the continuous data:

(1) How do the real-time ratings relate to the post-hoc ratings?
(2) How can we identify points in the lecture where students experience difficulty?
(3) How do the difficulty expectations from the instructor relate to the difficulty experienced by learners?
(4) How can we compare two synchronised instructional designs?
(5) How can we compare two non-synchronised instructional designs?

Participants

We recruited 100 participants aged between 18 and 42 (M = 25.04, SD = 4.66) from the same university. Participants were evenly distributed in terms of gender (52 Female, 48 Male), educational level (38 undergraduate, 30 masters, 32 PhD), and experience with the subject matter (21 had previous

Figure 2: Experimental setup.

knowledge of only Neuroscience, 35 had previous knowledge of only binary system, and 22 had previous knowledge of both topics). All the participants were rewarded with a gift card for their time and contribution. The experiment took place at our institution's usability lab (see Figure 2).

Materials

Video Lectures.
In selecting our video material, we drew from the literature on Educational Psychology by reusing four video lectures designed to evaluate instructional design techniques. In a larger study, Lodge et al. [23] compared different video types (text, static image, image and presenter face and full animation) and video design elements (static slides, digital ink, Socratic dialogue), developing video lectures corresponding to each of these conditions. We selected a subset of four of these lectures for our study—two about the basic workings of a neuron in the human brain (Neuroscience) and two about the conversion of base-ten numbers to binary (Binary). The Neuroscience lectures had the same audio track, but the Binary lectures did not. We deliberately chose these two combinations to allow us to devise methods for both types of comparison—synchronised and non-synchronised. The lectures on each topic had the same content, but different instructional designs:

(1) **Text:** The lectures consisted of a slide deck with a voice-over. Each slide contained mostly text and equations structured as 3-5 bullet points, with no illustrations.
(2) **Animation:** The lectures were presented in a dynamic style with little to no text. The Neuroscience lecture included the exact same voice-over as the Text version, but with a slide deck containing basically annotated images. The Binary lecture consisted of the same slide deck as the Text version, but interspersed with examples recorded with digital ink and a Socratic dialogue between the teacher and a novice.

All lectures were manually transcribed and annotated using Anvil[1].

Slider.
To collect the real-time perceptions of difficulty, we used a physical linear slider, which participants held and manipulated throughout the whole lecture. The slider was one of the faders available in the Numark Mixtrack PRO Midi Controller, which outputs a number between 0 and 127. The slider was placed to the right of the participant and we offered a desk support for their elbows, so as not to tire them. The physical nature of the device meant that participants did not have to look away from the lecture to set the difficulty level. Further, the kinesthetic feedback of holding the cap of the slider reminded participants to continuously provide their ratings.

Sensors.
This study sits within a wider research project about developing adaptive e-learning tools that respond to changes in biometric signals. As such, as well as the instruments described in this procedure, our experimental setup also included a set of sensors that unobtrusively monitored participants' biometric signals (see Figure 2). Specifically, during the study, we recorded participants facial expressions with a Logitech webcam, their eye movements with a Tobii Pro X2-30 eye-tracker, and their facial temperature with an Optris PI-400 thermal camera. The analysis of the sensor data is outside the scope of this paper and, therefore, is not reported here.

Pre-test and Post-test.
In order to test previous knowledge and the understanding of the material, participants completed two separate tests, one before and one after the lecture. The tests were also drawn from the same previous work as the lectures [23]. Each test contained 9 multiple choice questions with 4 answers and one "I don't know" (IDK) option. To counteract correct answers from random guesses, we instructed participants to select IDK if they were not sure which was the right answer. The tests were marked by awarding one point for each correct answer, while the incorrect, unselected or IDK options received zero points.

In addition, before each test, participants were asked to rate their confidence in the lecture topic. For example, before the Neuroscience lecture they were asked the following question - "Please indicate how confident you are that you can correctly answer questions about the structure and function of neurons, actions potentials, and synapses (0 being not confident, 100 being very confident)". The same question was asked again after they watched the lecture. In the Educational Psychology literature, these ratings are known as Judgements of Learning (JOLs) [35], which pertains to

[1]http://www.anvil-software.org/

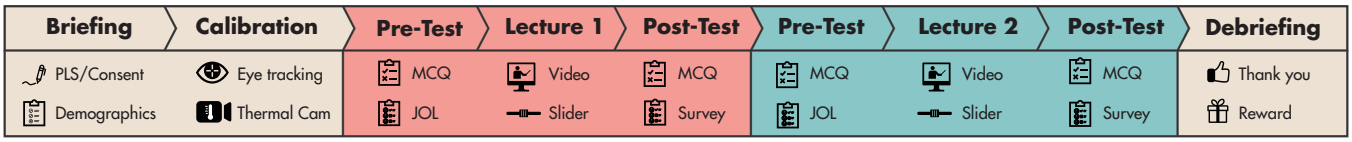


Figure 3: A figure showing the study procedure.

knowing what one knows and has an important implications for understanding how people learn and use memories.

Lecture Feedback Questionnaire.
After the lecture, participants were asked to fill in a survey asking about the feedback of the lecture in terms of 'mental-effort', 'challenge', 'confidence' and 'interest' on a 10-point bipolar scale. For instance, the participants were asked to rate "How much mental effort [they invested] while watching the video lecture" on a 10-point bipolar scale ranging from "Not much mental effort" (1) to "Lot of mental effort"(10). We also asked whether participants had been exposed to the content of the lecture prior to the experiment, and asked them to rate the lecture according to five factors on a Likert agreement scale ranging from "Strongly disagree"(1) to "Neutral (3)" to "Strongly agree"(5):

(1) **Organization**: "The video was well-organized"
(2) **Clarity** : "The video was clear"
(3) **Engagement**: "The video engaged me in learning"
(4) **Helped to Learn**: "The teaching in this video helped me to learn"
(5) **Satisfaction**: "Overall I am satisfied with video"

The lecture feedback questionnaire items were also adopted from the previous work as the lectures [23].

Procedure

Figure 3 shows a summary of our procedure. Upon arrival, participants were asked to read a plain language statement describing the purpose of the study and the procedure, to sign a consent form and to fill in a short demographics questionnaire asking about their age, gender, department and degree of study at the University. We then calibrated the eye tracker using the manfacturer's default 9-point procedure. We also verbally explained the purpose of the experiment and how they should operate the slider during the experiment. The experimenter was present in the same room during the course of the whole experiment, but outside the field-of-view of the participant.

The procedure was split into two learning tasks, where each task corresponds to each of the lecture topics. The procedure for each learning task is shown in Figure 3. First, participants completed the pre-test with 9 multiple-choice questions and rated their confidence-level about the lecture topic (JOLs). Then, participants watched the corresponding video lecture without pauses and were not allowed to take notes. While they watched the lecture, they continuously rated the difficulty level of the lecture using the slider. After the lecture, participants were first asked to fill out the questionnaire asking about their engagement and opinion of each lecture. Finally, they answered a Post-Test questionnaire which tested their retention and understanding of the material. The same procedure was then repeated with the second topic. Throughout the procedure, participants were monitored using the contactless sensors (webcam, thermal camera, and eye tracker). The order of the topics and the instructional designs were counter-balanced between participants, but each participant watched one lecture with each instructional design. For example, if the first lecture was the Animated version of the Neuroscience lecture, then the second lecture was the Text version of the Binary lecture. The complete session lasted approximately 60 minutes.

4 RESULTS

We present our results in terms of the research goals of the study, namely, in terms of analysing the time series formed by the continuous, real-time subjective ratings (1) in relation to the effects of instructional design, (2) on their own to identify points of difficulty, (3) in comparison to the expert's time series, (4) in comparison to a synchronised time series with the same audio but a different visual design, and (5) in comparison to a non-synchronised time series with different audio and visual design.

Effect of instructional design on Difficulty Ratings

During the lecture, participants were asked to move the slider to indicate the lecture difficulty on a scale from 0 (Low) to 127 (High). We first computed the mean difficulty rating for each condition. Consistent with the previous work around cognitive load [26], dual channels [6] and cognitive theory of multimedia [29, 30], it was found that mean difficulty ratings were higher for animation-based lectures as compared to text-based lectures. An independent-samples t-test was conducted to compare the effect of instructional design on mean difficulty scores of the participants. There was a statistical difference in the mean-difficulty rating for both the Neuroscience lecture – animation-based (*mean* = 46.18, *SD*= 29.84) and text-based (*mean* = 61.01, *SD* = 31.11); t(97)=-2.43, p<0.05 and Binary lecture – animation-based (*mean* = 46.72, *SD*=

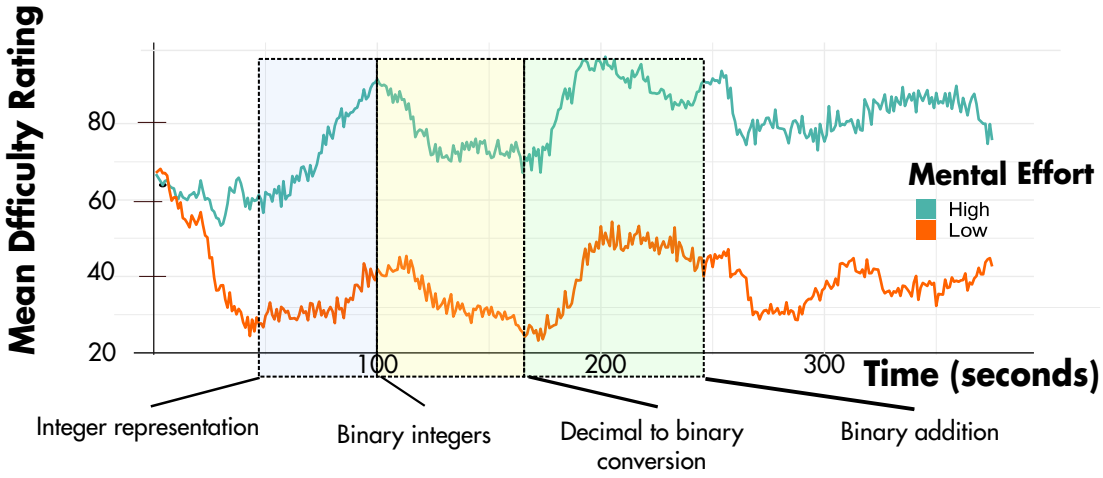


Figure 4: Segment by segment analysis of a Binary (text-based) lecture. The plot shows change in mean -difficulty ratings for the two groups of participants reporting high and low mental effort. The different colours in the plot correspond to different slides of the lecture as shown at the bottom.

28.05) and text-based (*mean* = 62.51, *SD*= 33.23); $t(97)$=-2.57, $p < 0.05$.

Identifying points of difficulty

The first advantage of collecting continuous subjective ratings from learners is that it enables instructors and educational designers to identify specific points of the lecture with which learners seem to be struggling with. Works that aggregate viewing behaviours perform this kind of analysis by identifying segments that are watched repeatedly. However, this kind of analysis requires that learners are able to rewind and rewatch segments of the video. By measuring difficulty in real-time, we can perform this analysis even if each student only watches the lecture once. Although difficulty is a subjective experience [25], recording real time continuous ratings helped us to understand the reasons which might contribute in the individual differences of difficulty ratings between the participants.

During the study, the participants self-reported the 'amount of mental effort they invested' on a 10-point bipolar scale after watching the video lectures. The mean mental effort score of all participant was used as the cut-off point to split the participants into two subgroups—low and high mental effort groups. Figure 4 presents the difficulty ratings of these two groups during the text version of the Binary lecture. Immediately, we can see the difficulty varied concomitantly, with peaks and troughs at the same points in the lecture. However, the absolute value of the ratings was considerably higher amongst the high mental effort group than in the low mental effort group, (i.e. *High Mental-Effort* : *mean*=78.7, *SD*=10.5, *Low Mental-Effort* :*mean*=38.3, *SD*=8.7;$t(45)$=5.335, p<0.001).

To explore the connection between visual transition and the peaks and troughs in the difficulty time-series, we manually annotated each state transition in all videos using Anvil. The first trough in the signals happened at $t = 48s$. This is exactly at the start of a slide that included a mathematical equation containing a summation of products (see Slide 1 in Figure 4). The difficulty steadily increased until $t = 100s$, which corresponded to a slide change. This new slide contained a practical example that clarified the equation in the previous slide (see Slide 2 in Figure 4). We can see that the difficulty decreased as the concepts became clearer with the explanation. We see a new trough followed by an increase in difficulty when a new slide introducing more equations appears at $t = 170s$ (see Slide 3 in Figure 4). At $t = 196s$, the instructor says "as an example", and proceeds to give a step-by-step example of the operation, causing the difficulty ratings begin to drop once again.

These findings are consistent with previous studies [15], where the authors demonstrated that visual transitions in video are often associated with a peak. However, we also found that formulas in the slides if not well understood by the participants may also result in subsequent peaks.

The same behaviours were found for the Neuroscience lectures, with troughs at the introduction of new terms and concepts, and peaks at the explanations. It is important to note that whereas in analyses based on aggregate numbers of visits to a video segment, a peak is indicative of a point of difficulty, but in our analysis, the opposite happens. Rather, a trough suggests the beginning of a segment that learners

experienced difficulty and a peak indicates the end of this segment.

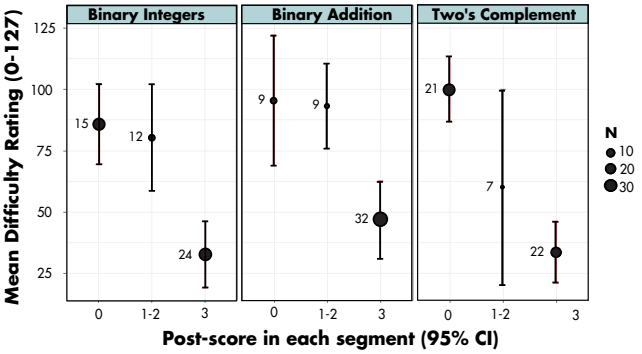


Figure 5: A 95% Confidence Interval plot showing mean-difficulty ratings and post-test scores of the participants in each segment of a Binary (text-based) lecture, where N=number of students.

Predicting test performance from segment ratings

A second advantage of analysing continuous data is the ability to relate the difficulty experienced in specific segments to the performance in the questions in the test corresponding to those segments. The multiple-choice questionnaires we used in our post-test contained 9 questions, which formed three groups. Each group of three questions mapped to three segments of each lecture. For example, in the Binary lecture, they mapped to the discussion of binary integers, binary addition, and two's complement.

To analyse this data, we aggregated participants' average ratings within the segments based on their performance on the corresponding group of questions—according to whether they answered all three questions correctly, incorrectly, or made 1-2 mistakes. Figure 5 shows that the participants' difficulty ratings for a given segment significantly predicted their performance on the questions corresponding to that segment.

A 95% confidence interval of the plot showing the relationship between mean difficulty ratings of the participants in each segment and their calculated post-score is shown in Figure 5. We found that in each corresponding segment (such as Binary Integers, Binary Addition and Two's Complement), the participants who scored high marks ($Post - score = 3$) reported low mean difficulty. Similarly, the participants who were not able to answer any of the questions correctly ($Post - score = 0$) reported comparatively high difficulty. Moreover, an interesting pattern was observed for the participants who scored one or two points in the post-score in a section. We found that that although, they scored high marks, they faced the same level of difficulty as the participants with zero-points in Binary Integers and Binary Addition segment. This shows that difficulty can be both incremental as well as detrimental to learning performance.

Assessing instructors' expectations

When designing a lecture, the instructor will have some idea of the relative difficulty of each segment in the lecture. A well-designed lecture balances this difficulty to ensure an optimal level of cognitive load. However, given the differences in experience with the content material, the instructor's expectations might not necessarily match the difficulty experienced by the students. We now demonstrate that by collecting continuous signals from the instructor and students, we are able to identify the points in the lecture where these expectations are misaligned.

To explore this idea, we invited the instructor of the Neuroscience lecture to participate in our study. The instructor watched the text version of the Neuroscience lecture and followed the same procedure as our participants, but rating the difficulty of the lecture according to his expectations of the level of difficulty that students should be experiencing at segment.

To assess the instructor's expectations, we averaged all participants' ratings of the corresponding video to generate a single time series. We then compared it to the instructor's time series by computing the correlation between the two signals with a 60-second sliding and 10-second step size. Figure 6 shows the two signals and the corresponding correlation curve.

The troughs in the correlation curve suggest segments where the expectations did not match the experience of the learners. The first of these happened at the 81-91st points in the correlation curve, which corresponded to $t = 121 - 141s$ in the lecture. Upon further inspection, we found that during this segment, 4 new terms were introduced in the lecture slides—*axon, axon hillock, myelin sheath, and nodes of Ranvier*. As this was the first time that a large proportion of the students were presented with these terms, and there were no pictures on the slides to support their learning in this version of the lecture, this naturally caused an increase in the difficulty ratings. However, given the instructor's familiarity of the term and the notion that this was simply a definitional slide, his own rating was low. The alignment resumed when both the instructor and the participants rated as highly difficult the segment that described more complex structures, such as the synapse, pre-synaptic terminal, synaptic cleft and post-synaptic terminal ($t = 180s$).

Overall, during the first half of the lecture, the instructor's expectations were well-aligned, but this did not happen in the second half. To investigate the reasons behind this, we analysed the transcript of the lecture corresponding to those segments. We found that from $t = 431s$ to $t = 531s$, the

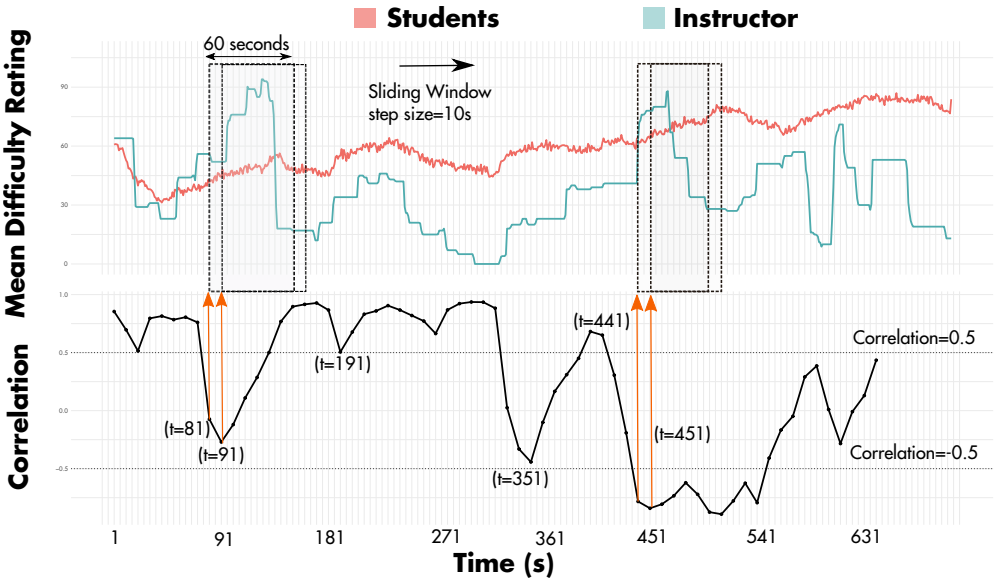


Figure 6: A graph showing mean difficulty ratings of the instructors and the participants (top) for the Neuroscience (text-based) lecture and a correlation plot between the two ratings (bottom) using a sliding window of 60 seconds with 10 seconds step-size.

lecturer said 251 words but only two slides were shown. The problem seemed to lie on the fact that the spoken words did not match the slides. Many unfamiliar terms (e.g. *myelin sheath*) were mentioned, but not spelled out, described, or explained in the slide. This lack of familiarity with these terms, combined with the fact that many participants in our pool were ESL students from disciplines other than biological sciences, led to high difficulty ratings. The instructor on the other hand, was familiar with the terms, and therefore reported lower difficulty. This suggests that lexical complexity is an important factor in the subjective ratings—particularly regarding the misalignment of expectation—and points towards a promising direction for future work.

Comparing synchronised lectures

Because the Neuroscience lectures contained the same audio track, only varying the video content, we could compare the two designs—text vs animation—point-by-point. By sharing a common audio track, we can isolate any differences in ratings to the video design and examine them in detail through the continuous time series. Figure 7 shows the difficulty ratings for each condition.

This comparison allowed us to identify three interesting points. First, we see that both groups experienced similar levels of difficulty in the first minute of the lecture, but began to diverge at $t = 74s$. Upon inspection of the video material, we found that whereas the animated version of the video presented the content one concept at a time, the text version presented a slide containing the definitions for all the terms presented in the next minute at once. This led to an increase in the ratings for the text version while the ratings for the animated version levelled off.

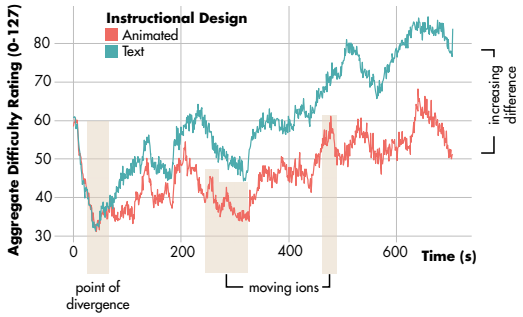
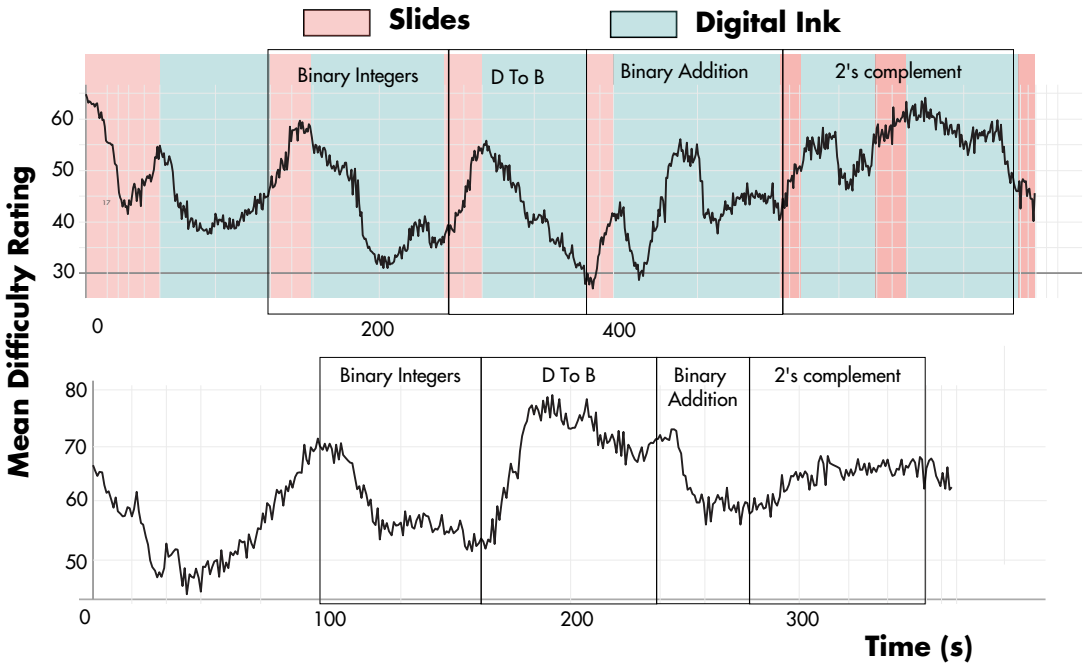


Figure 7: Comparison of two versions of Neuroscience lecture, both with the same audio track.

Second, we notice that after beginning to diverge, the difference in difficulty caused by the difference in presentation appeared to be amplified as the lecture progressed. We computed the difference between the two curves and built a linear regression model, which explained 76% of the variance (adjusted R^2, p<.001). For every second, the difference between the two curves increased by 0.033, reaching a maximum of 30 points at the end (which corresponds to approximately 25% of the slider's range). Though this trend should have a ceiling at some point due to the maximum value in the



		Binary Integers	D To B	Binary Addition	2's complement	
Binary(text-based)		58.9	69.57	62.8	63.4	65.4
Binary (animation)	Slides	55.6	44.65	32.44	44.89	57.80
	Digital Ink	41.92	41.65	41.05	50.38	58.30

Mean Difficulty Ratings for both the lectures for each section

Figure 8: A figure showing mean difficulty rating for all the participants for Binary (animation-based) lecture (top), Binary (text-based) lecture (middle) and a table showing the mean difficulty ratings across four major sections of both the lecture (bottom).

slider, it was maintained until the end of the lecture. This suggests an interesting potential for future work to explore unbounded controls such as rotational knobs.

The third point is that by comparing the relative shapes of the time series, we see that even though they match overall, the animated curve contains two spikes starting at $t = 246s$ and at $t = 520s$, which corresponded to segments of the corresponding lecture where ions flowed through synapses. Participants reported that keeping up with what was being said and watching the animation caused an increase in the difficulty of the lecture. Because the text version did not contain such animation, no spike is present.

Comparing non-synchronised lectures

To push the limits of the method, as well as comparing two lectures with the same audio track (as in the Neuroscience example), we compared two lectures that presented the same content, but with different audio and visual material (as in the Binary example). Due to differences in the length of the time-series for these lectures, a direct comparison was not a good approach. To make these comparable, we coded the transcript of both lectures and mapped the corresponding segments to each other. We then computed the mean difficulty rating in each segment for the groups of participants watching each version of the lecture. The results are presented in Figure 8. We found that participants who watched the Binary (animation-based) lecture reported less difficulty as compared to the participants who watched the text-based variant of the same lecture in each major section of the lecture.($t_{binary-integers}(93) = 1.788, p > 0.05$; $t_{D-To-B}(93) = 3.695, p < 0.001$; $t_{binary-addition}(93) = 2.73, p < 0.001$; $t_{2's-complement-m1}(92) = 1.57, p > 0.05$, $t_{2's-complement-m2}(90) = 0.86, p > 0.05$)

In both versions, we see a pattern of increasing difficulty with the introduction of a new topic, followed by a decrease

as soon as the lecturer begins to exemplify what was just described. However, the examples given with digital ink in the animated version of the lecture cause sharper declines than the examples given in the text slides.

5 DISCUSSION

The results from our 100 participants dataset demonstrated the advantages of using the slider as compared to traditional methods for assessing students' mental workload. Not only the results of the slides are consistent with the self-report ratings at the end of the lecture, they provide a lot more fine-grained information, which enabled us to inspect in detail specific characteristics of their instructional design.

In our study, to understand the differences of difficulty ratings across video lectures, we used two versions of two instructional videos—text-based and animation-based. We found that participants consistently rated the text-based lectures as more difficult than the animation-based ones. These findings are consistent with prior research around cognitive load [26], dual channels [6] and cognitive theory of multimedia [29, 30]. Additionally, we found the reason for the distinction by performing a more-fine grained analysis - *between design analysis* and *within design analysis*. In between design analysis, we compared the two different instructional designs which consisted of both synchronous or asynchronous video lecture designs. For the synchronised versions of the lecture, we compared the mean difficulty ratings across all the participants for the different designs, and identified points of divergence between the two. In contrast, for the lecture designs that were not in sync, we manually coded them according to the content and compared the individual segments. We found that pictorial representation of fundamental concepts helps in reducing the difficulty level of the lecture, such as picture of a neuron or an example illustrated using digital ink.

Similar analysis was done in within-design analysis, where we compared the difficulty ratings of the participants with instructors' expectations and found that for a complex topic such as Neuroscience, it is important to have a symmetry between the audio-visual modes of communication. We found that when new information is introduced but is not present on the slides, students tend to report higher difficulty levels. Moreover, for conceptual and formula-based topics such as conversion of Binary numbers, both groups of participants who reported high and low mental effort showed similar difficulty peaks and troughs in the time-series data. Further analysis revealed that introduction of new formulae and their explanation were the reasons for the peaks and troughs. If the explanation was well understood by the participants, they reduced the difficulty ratings, otherwise the difficulty ratings was kept high until a new topic is being introduced in the lecture. We also highlight the expected effect of examples in reducing the lecture difficulty. In all instructional designs, the introduction of new content led to an increase in difficulty level, whereas the clarification through examples reduced it. The reduction was even more pronounced for examples given with digital ink.

We argue that these insights would not have been revealed by looking at a single difficulty rating after the lecture. The continuous measurement of difficulty helped us to minutely identify differences between text-based and animation-based lectures, and perform a more fine-grained time-series analysis of the data.

Design Implications for Instructors

The micro-level analysis of participants' continuous self-reported difficulty ratings introduced in this paper can help the instructors to answer the following questions-

- **What is the impact of different instructional design on students' learning and performances?** As shown in the study, the students faced less difficulty while watching the animation-based variants of the two lectures, although there was no difference in their learning outcomes. Therefore, we can conclude animation-based slides including either pictorial representation of complex phenomenon such as working of human brain or detailed explanation of formula based concepts using examples and hand-writing tool can help to reduce the amount of difficulty faced by the students.
- **Which part of the lecture content was difficult for most of the students?** Difficulty plays a crucial part in learning. However, difficulty experiences during learning and resulting confusion can be either productive or unproductive [25]. Therefore, it's important for instructors to compare their expected difficulty ratings of the content with the students' belief about the lecture difficulty. As evident by the results, the continuous measurement of difficulty ratings gave an insight about the learning experiences of the students and incurred difficulties. For example, we found that results for change in difficulty depends on three factors- pictorial representation of complex terms, difference between the amount of verbal and visual information presented per slide and number of words spoken by the instructor per second.

Limitations and Future Work

Through this work, we offered a first look at how to collect and analyse real-time difficulty data for the evaluation of video lectures. Though we demonstrate the usefulness of the method for achieving new insights about instructional

design, asking students to actively rate the lecture as they watch it might incur extraneous cognitive load. Though Tangen et al. suggest that this type of data collection should not incur additional cognitive load, we do acknowledge this possibility. Having said that, all conditions in our study included this component, so any additional workload would be distributed across conditions.

We found a trade-off in terms of the size of the sample and its ecological validity, because if we were to recruit only students that are learning the specific material, our sample would be much smaller (as not everybody in the class would volunteer), and we would have to run the study in a much shorter timespan (as after some time the previous experience across the sample would vary). In this study, we chose to sacrifice ecological validity in order to have a large sample size ($N = 100$).

Even though the number of participants in our study was large for HCI standards [7], we only examined a small number of topics and instructional designs. Moreover, we believe the videos are slightly decontextualized from the learning design, however the main contribution of the study is regarding the kinds of analysis that can be done based on the continuous data.

As described earlier, this study is limited in its ecological validity. The data collection took place in a lab setting and we did not allow participants to exhibit otherwise natural behaviours, such as pausing and rewinding the video or taking notes. Further, participants were asked to continuously manipulate the slider, which is not a natural behaviour either. Given the usefulness of continuous data for the analysis of video lectures, but the difficulty in collecting this data, in future work we will attempt to automatically predict the continuous signal that participants self-reported through the multimodal set of sensor data that we also collected during the study. If successful, this could pave the way for adaptive learning experiences that are able to measure learners' difficulty in real-time.

6 CONCLUSION

In this paper, we demonstrated the use of slider as compared to traditional methods for assessing student's mental workload. Our results on a 100 participant dataset helped us to analyse the time-series formed by the continuous real-time subjective ratings to (1) identify points of difficulty in lecture (2) predict test performance from segment ratings (3) assess instructor's expectations and (4) compare different instructional designs. We believe our approach has design implications for the instructors by enabling them to understand the impact of different instructional designs on students' learning experiences and also provides a quantitative framework to compare their expectations with students' belief of the lectures.

ACKNOWLEDGMENTS

Dr Eduardo Velloso is the recipient of an Australian Research Council Discovery Early Career Award (Project Number.: DE180100315) funded by the Australian Government. Namrata Srivastava is supported by Melbourne Research Scholarship.

REFERENCES

[1] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2014. Engaging with massive online courses. In *Proceedings of the 23rd international conference on World wide web*. ACM, 687–698.

[2] Amaël Arguel, Jason Lodge, Mariya Pachman, and Paula de Barba. 2016. Confidence drives exploration strategies in interactive simulations. ASCILITE.

[3] S Adams Becker, Michele Cummins, Adam Davis, Alex Freeman, C Glesinger Hall, and Vanish Ananthanarayanan. 2017. *NMC horizon report: 2017 higher education edition*. Technical Report. The New Media Consortium.

[4] Alberto Betella and Paul FMJ Verschure. 2016. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one* 11, 2 (2016), e0148037.

[5] 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), 13–25.

[6] Roland Brünken, Jan L Plass, and Detlev Leutner. 2004. Assessment of cognitive load in multimedia learning with dual-task methodology: Auditory load and modality effects. *Instructional Science* 32, 1-2 (2004), 115–132.

[7] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[8] Hee Jun Choi and Scott D Johnson. 2005. The effect of context-based video instruction on learning and motivation in online courses. *The American Journal of Distance Education* 19, 4 (2005), 215–227.

[9] Konstantinos Chorianopoulos. 2013. Collective intelligence within web video. *Human-centric Computing and Information Sciences* 3, 1 (15 Jun 2013), 10. https://doi.org/10.1186/2192-1962-3-10

[10] Andrew Cross, Mydhili Bayyapunedi, Edward Cutrell, Anant Agarwal, and William Thies. 2013. TypeRighting: combining the benefits of handwriting and typeface in online educational videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 793–796.

[11] Michael N. Giannakos, Konstantinos Chorianopoulos, and Nikos Chrisochoides. 2014. Collecting and making sense of video learning analytics. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*. 1–7. https://doi.org/10.1109/FIE.2014.7044485

[12] Michail N Giannakos, Konstantinos Chorianopoulos, and Nikos Chrisochoides. 2014. Collecting and making sense of video learning analytics. In *Frontiers in Education Conference (FIE), 2014 IEEE*. IEEE, 1–7.

[13] 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.

[14] Päivi Karppinen. 2005. Meaningful learning with digital and online videos: Theoretical perspectives. *AACE Journal* 13, 3 (2005), 233–250.

[15] 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.

[16] 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.

[17] René F Kizilcec, Kathryn Papadopoulos, and Lalida Sritanyaratana. 2014. Showing face in video instruction: effects on information retention, visual attention, and affect. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2095–2102.

[18] KH Vincent Lau, Pue Farooque, Gary Leydon, Michael L Schwartz, R Mark Sadler, and Jeremy J Moeller. 2018. Using learning analytics to evaluate a video-based lecture series. *Medical teacher* 40, 1 (2018), 91–98.

[19] Gaël Laurans, Pieter MA Desmet, and Paul Hekkert. 2009. The emotion slider: A self-report device for the continuous measurement of emotion. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII 2009)*. Ieee, 1–6.

[20] Nan Li, Lukasz Kidzinski, Patrick Jermann, and Pierre Dillenbourg. 2015. How Do In-video Interactions Reflect Perceived Video Difficulty? *Proceedings Papers* (2015), 112.

[21] Shih-Yin Lin, John M Aiken, Daniel T Seaton, Scott S Douglas, Edwin F Greco, Brian D Thoms, and Michael F Schatz. 2016. Exploring University Students' Engagement with Online Video Lectures in a Blended Introductory Mechanics Course. *arXiv preprint arXiv:1603.03348* (2016).

[22] Tharindu Rekha Liyanagunawardena, Andrew Alexandar Adams, and Shirley Ann Williams. 2013. MOOCs: A systematic study of the published literature 2008-2012. *The International Review of Research in Open and Distributed Learning* 14, 3 (2013), 202–227.

[23] Jason Lodge, Jared Cooney Horvath, Alex Horton, Gregor Kennedy, Sven Venema, and Shane Dawson. 2017. Designing videos for learning: Separating the good from the bad and the ugly. (01 2017).

[24] Jason Lodge and Gregor Kennedy. 2015. Prior knowledge, confidence and understanding in interactive tutorials and simulations. ASCILITE.

[25] Jason M Lodge, Gregor Kennedy, Lori Lockyer, Amael Arguel, and Mariya Pachman. 2018. Understanding difficulties and resulting confusion in learning: An integrative review. In *Frontiers in Education*, Vol. 3. Frontiers, 49.

[26] Richard E Mayer and Roxana Moreno. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist* 38, 1 (2003), 43–52.

[27] Barbara Mitra, Jenny Lewin-Jones, Heather Barrett, and Stella Williamson. 2010. The use of video to enable deep learning. *Research in Post-Compulsory Education* 15, 4 (2010), 405–414.

[28] Dan R Olsen and Brandon Moon. 2011. Video summarization based on user interaction. In *Proceedings of the 9th European Conference on Interactive TV and Video*. ACM, 115–122.

[29] Fred Paas, Juhani E Tuovinen, Huib Tabbers, and Pascal WM Van Gerven. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist* 38, 1 (2003), 63–71.

[30] Fred GWC Paas and Jeroen JG Van Merriënboer. 1993. The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human factors* 35, 4 (1993), 737–743.

[31] Evan F Risko, Tom Foulsham, Shane Dawson, and Alan Kingstone. 2013. The collaborative lecture annotation system (CLAS): A new TOOL for distributed learning. *IEEE Transactions on Learning Technologies* 6, 1 (2013), 4–13.

[32] Jennifer Sabourin, Bradford Mott, and James C Lester. 2011. Modeling learner affect with theoretically grounded dynamic Bayesian networks. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 286–295.

[33] Ryan Shaw and Marc Davis. 2005. Toward emergent representations for video. In *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 431–434.

[34] Hyungyu Shin, Eun-Young Ko, Joseph Jay Williams, and Juho Kim. 2018. Understanding the Effect of In-Video Prompting on Learners and Instructors. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 319.

[35] Nicholas C Soderstrom, Carole L Yue, and Elizabeth Ligon Bjork. 2016. 11 Metamemory and Education. *The Oxford Handbook of Metamemory* (2016), 197.

[36] Jason M Tangen, Merryn D Constable, Eric Durrant, Chris Teeter, Brett R Beston, and Joseph A Kim. 2011. The role of interest and images in slideware presentations. *Computers & Education* 56, 3 (2011), 865–872.

[37] Peter Tiernan. 2015. An inquiry into the current and future uses of digital video in University teaching. *Education and Information Technologies* 20, 1 (2015), 75–90.

[38] 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.