The effect of captions and written text on viewing behavior in educational videos

Jonas Rolf Persson[1], Eirik Wattengård[2] and Magnus Borstad Lilledahl[1]

[1] Department of Physics, Norwegian University of Science and Technology, Norway
[2] Multimedia Centre, Norwegian University of Science and Technology, N-7491 Trondheim, Norway

The use of videos as learning objects has increased together with an increased variation in the designs of these educational videos. However, to create effective learning objects it is important to have detailed information about how users perceive and interact with the different parts of the multimedia design. In this paper we study, using eye-tracker technology, how fast and for how long viewers focus on captions and written text in a video. An educational video on thermodynamics was created where captions were used to highlight important concepts. Screen recordings of written text from a tablet were used to illustrate mathematical notations and calculations. The results show that there is a significant delay of about 2 seconds before viewers focus on graphical objects that appear, both for captions and for written text. For captions, the viewers focus on the element for 2-3 seconds, whereas for written text blocks, it is strongly dependent on the amount and quality of the presented information. These temporal aspects of the viewers' attention will be important for the proper design of educational videos to achieve appropriate synchronization between graphical objects and narration and thereby supporting learning.

Article details

LUMAT General Issue
Vol 7 No 1 (2019), 124–147

Received 29 January 2018
Accepted 17 September 2019
Published 3 October 2019

Pages: 24
References: 28

Contact:
jonas.persson@ntnu.no

https://doi.org/10.31129/LUMAT.7.1.328

Keywords: video, cognitive science, eye-tracker, viewing behaviour, physical science, cueing

1 Introduction

The use of video in higher education is increasing. Universities feel pressure from several sources to use videos or lecture recordings, including students, funding agencies, and other governmental agencies (O'Callaghan et al. 2017; Shah et al. 2013). Most instruction is currently provided in traditional lectures that are usually instructor-centered with the lecturer controlling both pace and content. This may cause some students to miss out on important aspects if the instruction goes too fast while other students are not able to observe and understand the content properly. Videos are seen as a solution to this, providing instruction that is perceived as student-controlled and self-paced. Another advantageous feature of videos compared to lectures is the opportunity for multiple videos on the same theme, addressing students at different levels of understanding. O'Callaghan et al. (2017) report in their review a range of benefits of lecture recordings, for example, accessibility, self-paced learning,

and reuse. There exist studies that report positive student perceptions. However, there are far fewer studies demonstrating an effect of videos on student academic performance and there are no conclusive results in either direction for this aspect as concluded by O'Callaghan et al. (2017), why they recommended further research.

The inconclusive results stem partially from the difficulty in assessing the effect of video on learning separate from other effects, especially when there is no appropriate control group, which is often difficult to establish in an educational setting. As an example, Bos et al. (2016) divided students into four groups based on their usage of recorded and live lectures. The students who attended lectures and used videos as a supplement did significantly better in an assessment compared to those only watching recorded lectures. However, this result is just as reasonable to attribute to the group using both lecture and video representing the more diligent and hard-working part of the student population rather than the specific learning effect of watching videos. Due to this inherent difficulty in studying videos as a generic learning object, it seems pertinent to rather focus on the details of a specific video recording and how it directly affects student behaviour and learning. This is the approach taken in this work.

The purpose of producing videos differs, depending on how the videos fit the overall course design. In many cases, the videos are no more than a recording of the lectures, with little or no editing. In other cases, the videos are intended to be watched by the students in advance, to prepare them for the coming lecture, and designed accordingly. It is important to design the video according to its intended use to optimise the desired learning process. To make these design choices it is useful to have a model of the learning process when interacting with a video. We have used the Cognitive theory of multimedia learning (CTML), described below, as the premise for our video design.

1.1 The cognitive theory of multimedia learning

The CTML put forward by Richard E. Mayer (Mayer, 2006; Sorden, 2012), is a model where multimedia instruction refers to presentations involving words and pictures that are intended to foster learning. The CTML is based on three assumptions: 1) The dual-channel assumption, 2) the limited capacity assumption, and 3) the active processing assumption. The dual-channel assumption is based on the works of Badderley (1986) and Pavio (1986) on working memory with auditory and visual channels with dual coding. The limited capacity assumption is based on the cognitive

load theory (CLT) (Sweller,1988; 1994) which states that each subsystem of the working memory has a limited capacity. The active processing assumption states that people construct knowledge in meaningful ways when they actively pay attention to relevant information and organise it into a coherent structure in their mind.

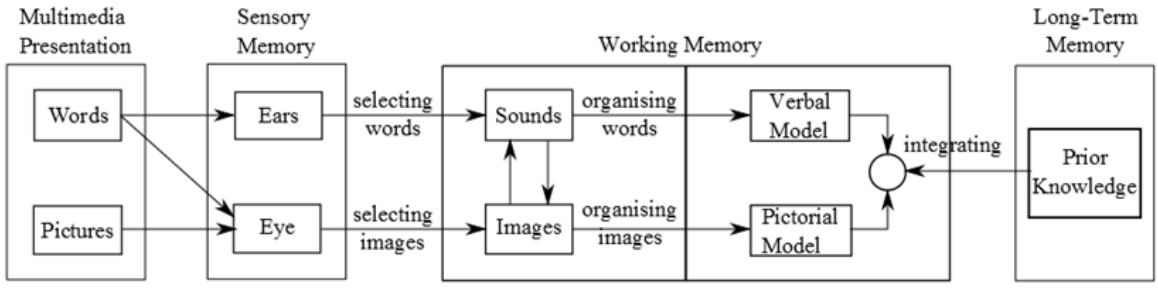

Figure 1. A graphical representation of the information processing in the CTML.

Figure 1 graphically presents the CTML. We have three different memory stores, the sensory memory, the working memory and the long-term memory (Mayer, 2006). The information in the multimedia presentation (such as words and pictures) enters the sensory memory through the eyes and ears. The information is then divided into visual (text and pictures) and auditory (spoken words) information. The sensory memory can only hold the information for a brief period. The main processing of information takes place within the working memory where it is organised into models, constructed with the information available. The dual channels (auditory and visual) exchange information in the working memory (e.g. when you read the word “cat”, a visual image of a cat may appear in your mind, while an image of a cat can invoke a sound image of the word cat). The long-term memory holds the learners’ prior knowledge. To actively use this knowledge, it must be brought into the working memory and integrated with the new information.

The CTML results in three important aspects of multimedia design that should be addressed in educational video design.

1. It is important to exclude extraneous information as the amount of information that can be processed is limited.
2. The integration and organization of new information with prior knowledge are essential and the content should, therefore, be presented in an organised form, preferably with references to previous knowledge and different applications.

3. As the information is acquired through the senses and the selected content is transmitted to the working memory, it is important to direct the attention of the viewers to the appropriate content in the video. Using various cues in the video to direct the attention of the viewers to the appropriate location or concept is the main topic of this paper. This is described further in the next section.

As a side note, it is important to be aware of the social cues in videos, which might be advantageous to learning (Mayer et al, 2004). The presentation should be appealing and inspiring to students in a social context, including eye contact, language and body language.

The CTML model has been extensively studied and validated in several studies (Sorden, 2012; Mayer, 2014). Still, there are critics of the theory and the experimental studies. Ballantyne (2008) has criticised some studies for being quite narrow in scope and discussed if the principles tested in an experimental situation can be applied in a realistic situation. Most studies tend to focus on the understanding of physical systems, which raises questions about how applicable the results are in other contexts.

1.2 Using CTML and the cinematic language in educational video production

The rules for how to communicate through moving images have developed rapidly over the past century. Since the one-shot precursors to the narrative film, like "Workers Leaving the Factory" from 1895, shot by the Lumière Brothers, and early fiction short films like Georges Méliès' "Cinderella" from 1899, his first work with more than one shot (Thompson, Bordwell, 2010, p. 14), the medium today produces complex narrative works created from and available on multiple digital and analogue platforms, e.g. Peter Jackson's Lord of the Rings trilogy from 2001-2003 or HBO's TV series Game of Thrones, running continuously since 2011. Digital media have allowed the moving visual narrative language to be available to almost anybody almost anywhere (Thompson, Bordwell, 2010, p. 730). Thus, the language of cinema–we use this term loosely to include video with film and television–in this regard is relatively new, and its details are still somewhat incomplete, compared to other forms of communication and to the conventions in other visual arts. Like linguistic forms of communication, the language of cinema is still dynamic and ever-developing, albeit the basic conventions are by now quite well established, at least in part by their accessibility to wide audiences through the multiple platforms of new media

(Manovich, 2001), which makes the foundation for communication through moving images clear to the vast majority of contemporary viewers.

The main target audience for educational videos are students attending an institution for learning–a school, college or university–most of whom are young enough to be a part of the post-MTV media consciousness. This audience became media savvy at a young age and is comfortable relating to audio-visual presentations (Brooks, et al., 2005). Hence, the established visual languages speak clearly to them, and it is natural to use the rules and conventions of contemporary cinema (Thompson, Bordwell, 2010) and television as basic principles when communicating to them with the use of video.

These basics of the cinematic language include methods of montage (in our context "montage" means editing–the juxtaposition of different moving images), like the continuity editing method which combines shots that are related in content in order to generate a logical progression of the narrative where one thing leads to another (Rosenberg, 2010), or the Eisensteinian method of combining shots not directly related to each other in order to create new ideas in the viewers, e.g. in a montage sequence of a series of non-successive shots where the juxtaposition of these unrelated shots generates associations and ideas in the mind of the audience (Eisenstein, 1929). When observing educational videos one can see that continuity editing is often used to explain specific concepts and skills. Other general rules for filmmaking, like not crossing the axis–staying on one side of the line generated by subjects' sightlines or direction of movement in order to keep the viewers oriented on subjects' positions and orientation in a scene–and using high or low angles to establish a subject's position of power or authority, should be adhered to in educational videos the same way they are in other narrative works of moving images.

When integrating CTML's set of pointers for how to approach the production of an educational video with the use of the basic principles of the cinematic language, it should be possible to enhance the learning effects of videos in education.

1.3 Viewers' attention and visual focus

Based on CTML, Mayer (2006) developed 12 principles of multimedia design. Two of these principles are especially important for our study:

- **The signalling principle**: People learn better when cues that highlight the organization of the essential material are added.

- **The temporal contiguity principle**: People learn better when corresponding words and pictures are presented simultaneously rather than successively.

Taken together, these two principles indicate that the viewers' attention should be directed to the graphical object under discussion (temporal contiguity) and that these should be cued (signalling principle) to direct the attention of the viewers to the important content. The topic of this paper is how different cues, or absence thereof, affect the attention of the viewers.

Eye movements and where a person looks gives an indication of the visual information intake. As the information must be processed, the intake and cognitive processing are connected. Just & Carpender (1984) found this in the case of reading texts. The working hypothesis is that there is a strong correlation between where one is looking and what one is thinking about, the so-called eye-mind hypothesis (Just & Carpender 1984).

Eye-tracking technology is used to observe such eye-movements in an educational context (Holmqvist et al., 2011; Mayer, 2010). The recording of the eye-movement using eye-tracking equipment provide a dynamic trace of where a viewers' attention is directed in relation to a visual display. Measuring different aspects of eye movement, such as time to first fixation, duration and sequence of fixations, can be used as indicators of processing of information, according to the eye-mind hypothesis (Rayner, 1998). It has been suggested by Rayner (1998) that eye-movement parameters such as the number of fixations, fixation duration, and total inspection time are relevant to the learning process. During the last years, studies using eye-tracking technology in an educational context have been performed (for example Lai et al., 2013; Scheiter & Eitel, 2016). The results demonstrate that the use of eye-tracking gives important insights into how students learn.

The study of visual focus as a proxy for attention has been studied in a wide variety of fields including reading, medical diagnosis, web interfaces and video games (See Holmqvist et al., 2011, Liversedge et al., 2011). However, for educational videos, the use of eye-tracking technology for assessing viewers' attention is rather new. In this setting, graphical objects (text or pictures) are often included together with the narration with the goal of increasing the transfer of information. The time to first fixation of a new graphical object is then important to determine when the processing

of the information by the viewers can start while the total fixation time is related to the total amount of time required to process the information.

The use of text as a conveyor of information in videos differs depending on the presentation. Some use a slide-based presentation where all the information is visible at once, thus causing the viewers to try to read all the text at once. The viewers might, therefore, miss what is said during the time it takes to read the text, as indicated by the CLT. Writing the text live in the video will lessen the cognitive load as the viewers cannot read in advance. The use of smaller text blocks, instead of presenting all text at once, might also lessen the cognitive load by controlling the pace of information accessible to the viewers.

A recent study by Fiorella and Mayer (2016), found that an instructor drawing a diagram had positive effects on the performance for low prior knowledge learners, with no or negative effect on high prior knowledge learners. They also observed positive effects when the instructor or the instructor's hand was visible while drawing. The use of animated pedagogical agents has proved positive for learning (Wang et al., 2017), especially when gesturing. In this case, the time it takes for the viewers to find the text is important, especially if narration and text are synced. The time to first fixation will then give a measure of a possible delay that is important to take into account.

Cueing, in the form of, for example, salient features like colour, movements and gestures, can be used to direct the viewers' attention to important information. The preceding content can also work as a cue, e.g. when writing text on a line. Cueing has been studied in different contexts, such as animation (see review by de Konig et al, 2009) and to a lesser extent in static diagrams (Canham & Hegarty, 2010). The effects of cueing in these cases are mixed, from being positive in the case of static diagrams, but inconclusive for animations.

The reaction time to cueing was studied by Jonides & Yantis (1988) where the case of abrupt visual onset gave the fastest response (about 0.5-0.6 s). However, the context in this study is different and the results of Jonides & Yantis (1988) can only be used as a baseline. To our knowledge, there does not seem to exist any published studies on the time to first fixation on cues in videos.

In videos where the presenter is visible, the saliency of the human face will provide a strong attractor for attention (Buswell 1935, Yarbus 1967) and thus the viewers will focus on it for different reasons. Nonverbal communication has been found to be an

important aspect in human communication (Burgoon et al, 2016) and could be one explanation for the viewer's focus. For example, by looking for emotional (facial expressions) or attentional (gaze) cues on the importance of different topics. The design in this video makes the use of the presenter's gaze as a cue. If the scene changes the viewers gaze will linger on the last fixation and this will be the starting point for scanning the new scene or changes within the scene. Recurrent starting points will be preferred in the scanning process.

In this work, we are interested in four main research questions:

1. "When a caption is presented to emphasize an important concept, will the viewers' attention be attracted to this caption without significant delay?"
2. "When the scene changes from showing the presenter to a screen recording where text appears as it is written, will the viewers be able to focus on this un-cued content without significant delay?"
3. "Will the following text blocks (AoI) show a shorter delay in focusing their attention to it?"
4. "How long should captions and text be presented in order to make sure viewers can read the information when an attention delay is taken into account?"

The answers to these questions will be important for an optimal design of instructional videos.

2 Video design

An educational video on thermodynamics, more specifically, equations of state and state variables was recorded in a TV studio (In Norwegian, the language used in the video, the title was "*Termodynamikk: Tilstandsvariabler og tilstandslikninger*" ["Thermodynamics: State variables and equations of state]"). The lecturer used a SmartTechnologies Podium tablet for writing, from which we recorded the screen image separately. The video was edited from the medium close-up shot of the lecturer and the screen recording from the tablet according to the narrative. For this video we used continuity montage, staying on the axis, and shot the lecturer straight on in an eye-level angle to establish him not as a person of authority (done with a low-angle shot) or a servant (high-angle shot) but as an honest and trustworthy communicator. We also ensured that the technical quality of the video was as good as possible, with sharp images and good sound.

We chose to have the presenter on-camera, communicating verbally, directly to the viewers, as opposed to producing a screen recording of the tablet only with the presenter's voice in the audio. This choice was made due to the social cue provided by the presenter, a face for the viewers to relate to. An absence of the visual presence of the presenter in the video could create a distractor. The gaze of the presenter may also be used for cueing. In the present design with both presenter and screencast, there will exist a mismatch in the cues between the two modes as the presenter's face is a very strong cue and the initial tablet screencast without any graphics has no cues.

Clarity in the presentation was ensured by a didactic quality assurance of the script before shooting and by directing the presenter's on-camera presentation to follow a strictly written script yet stay informal enough to avoid the risk of alienating the viewers.[1]

The presenter wrote and drew on the tablet in real time according to the planned script, and when he used the tablet the video cut to show the screen recording. During the lecture, terms, such as s*tate variables* and *equation of state* were introduced. In accordance with Mayer's temporal contiguity principle–that elements are perceived most efficiently when presented simultaneously –, we introduced an on-screen caption with the terms in question synchronized with the lecturer's voice. Adding visual captions synchronized with the auditory information utilizes the dual (visual and auditory) channels leading to the sensory memory.

2.1 Captions

In this study, we were interested in the viewers' focus on different objects as they appear in the video. When viewing a video, the gaze is normally fixed on objects of interest or the face of the presenter. The motion or appearance of an object might draw attention to it, but the delay of this focus will be context dependent and affect the role of the object in supporting the cognitive processes.

When an object, such as a caption or subtitles is expected to appear, the viewers scan that area continuously and will, therefore, focus more quickly on the object. The

[1] All these production elements were utilized to focus the video presentation on the necessary content, as per Mayer's personalization, voice and image principles.

captions in the video were placed in the lower half (figure 2).[2] The white and yellow captions used resulted in high visibility and strong signalling with the background used. The figure also shows the *areas of interest* (AOIs) that were used in the analysis of the viewer's gaze, further described in section 2.2. Captions were shown 4 frames before the word was spoken as is common practice for subtitles in movies.

To study the effect of different presentations of the captions we produced two separate videos presenting the same content, but with a slightly different presentation of the captions. The master recordings–the medium shot of the presenter and the separate screen recording–were the same. The editing of the two recordings were also very similar. In version A, the terms *tilstandsvariabler* (state variables) and *tilstandslikninger* (equations of state) are initially introduced as separate captions, one following the other; in version B the same terms are presented simultaneously as the terms are introduced and then presented as separate captions as each term is discussed. In the summary of the video, the terms are shown simultaneously with the one being explained highlighted in yellow as opposed to the other in white. In version B the terms are initially presented simultaneously and then presented as separate captions synchronized to the presenter's voice.

Figure 2. Lecturer and caption, with defined AOIs in red and green, respectively. The AOIs are not visible while viewing the video.

[2] The placement of the captions just below the presenter's face was determined based on common practice and Mayer's principle on spatial contiguity, that (new) visual information should appear near the current point of focus (i.e. the presenter's face in this case)

2.2 Written text

The written text in the video was a screencast from a tablet. In order to make the transition as smooth as possible, the video was designed so that when the lecturer shifted his gaze down to the tablet in the studio, the video cut from the live action video shot to the screencast. The screencast from the tablet presented three different panels as illustrated in figures 3-5. The first panel introduced the main physical variables under discussion (figure 3). The second introduced typical values (1 L of gas under atmospheric pressure and room temperature) for these variables (figure 4). The third panel introduced a simple calculation using the introduced equations (figure 5).

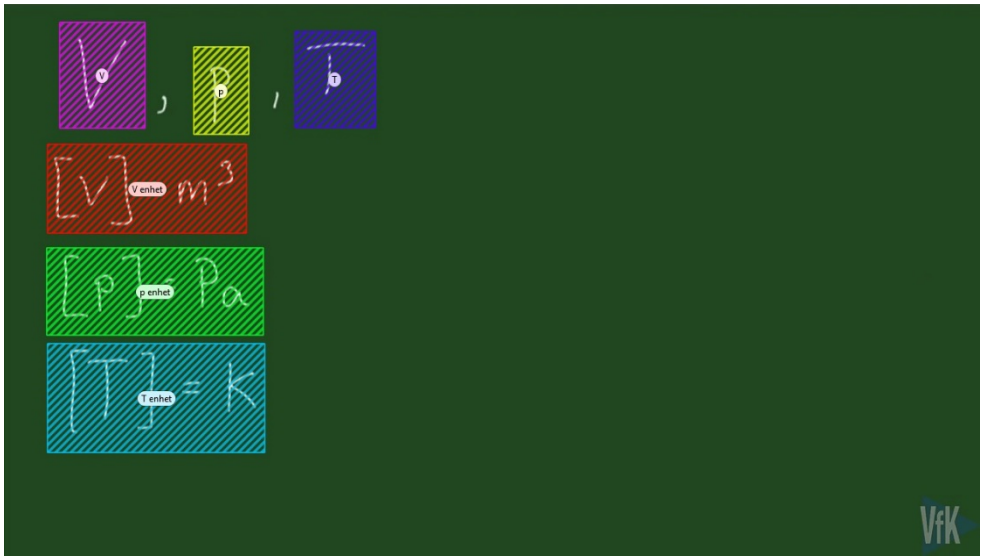

Figure 3. First panel from the screencast recording, defining important thermodynamic variables and their units. The different AOIs used in the analysis are illustrated by rectangular areas of different colours.

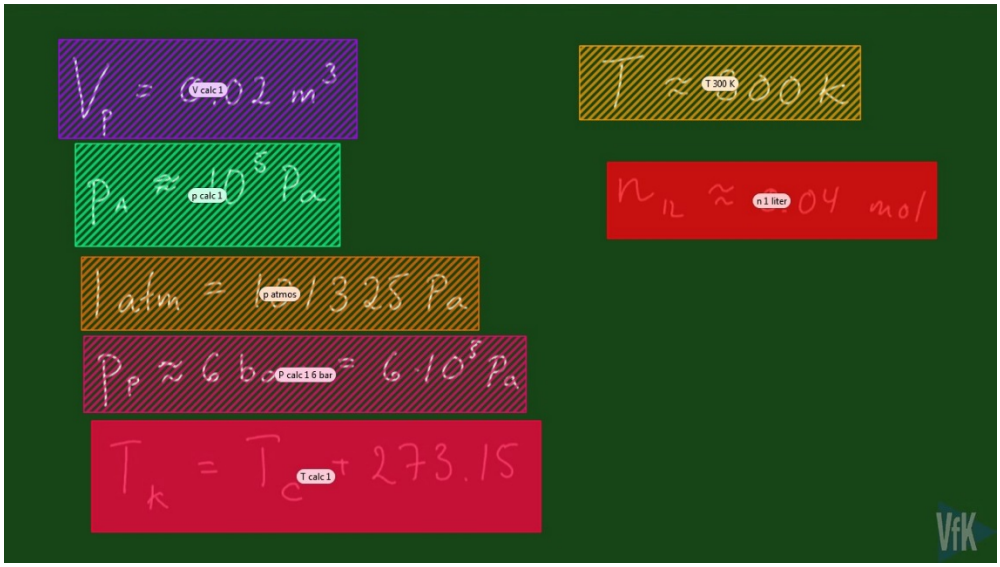

Figure 4. The second panel from the screencast recording, illustrating common values for the quantities under discussion. The different AOIs used in the analysis are illustrated by rectangular areas of different colours.

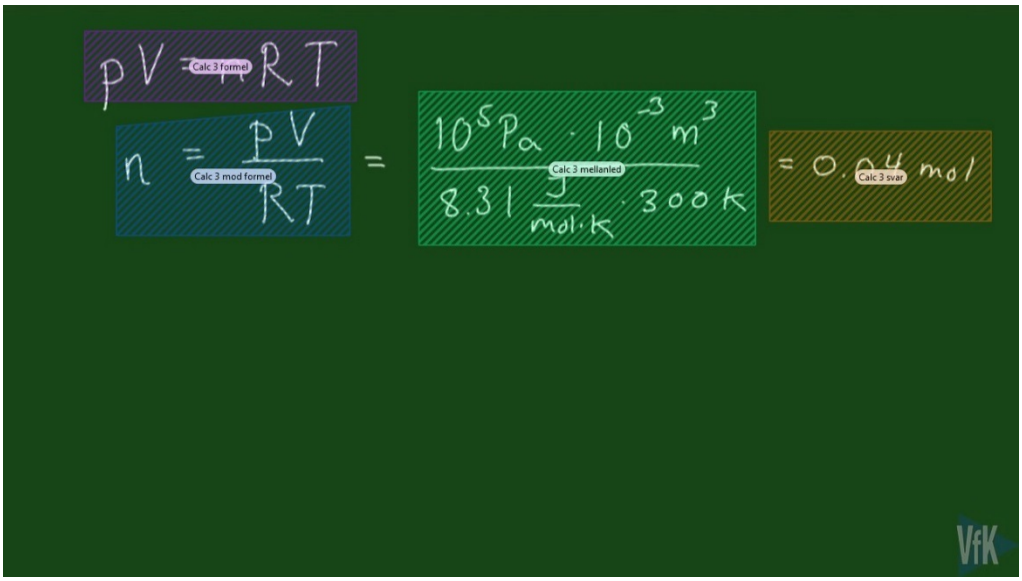

Figure 5. The third panel from the screencast recording, showing a simple calculation of the quantities introduced. The different AOIs used in the analysis are illustrated by rectangular areas of different colours.

3. Experimental setup

3.1 Eye-tracker setup

Participants' eye movements while watching the videos were recorded with a Tobii X2-60 eye-tracker (Tobii, 2016) below a 17" display. The eye-tracker data were collected and analysed using the Tobii Studio software (Tobii, 2016). The Tobii X2-60 tracks eye movements with a sampling frequency of 60 Hz and angular resolution of 0.25°. With the viewers placed about 60 cm from the screen, this provided an accuracy of 2.6 +/- 0.5 mm on the screen. Due to limitations in the eye-tracking software, it was not possible for the participants to pause or rewind the video. This will give rise to an unnatural situation for students that normally pause or actively search for information in the video.

Areas around the different captions and written text were designated as *areas of interest* (AOI) and used to determine if the viewers focus on these objects. See figure 2 for how these were defined for the captions and figures 3-5 for the written text.

In the video, a tablet was used to write the presented text. The text will then emerge in the video without a hand or other form of cueing. The defined AOIs cover the whole area of the finished text segments (see figures 3-5). In the narration, there is a natural short pause between subsequent segments, making it possible to define several AOIs on one line.

3.2 Data analysis

Two parameters were extracted for each AOI from the eye-tracking data, *time to first fixation* (TTFF) and *total fixation time* (TFT). TTFF is the time it takes from an AOI becomes visible until the viewers gaze rests on the object while TFT is the total time the viewers look at the AOI. As the content appears gradually for the written text, we started the timing for TTFF and TFT when part of the first letter in every text segment was clearly visible.

TTFF was treated as a Poisson process, reflecting the search and find process. TFT was treated as a Normal distribution and both mean and standard deviation is given. Satistical significance was analysed using Student's t-test

3.3 Participant sample

The participants in the study were voluntary students attending the course of the lecturer in which the video was intended to be used. The students had not yet been exposed to the topic of the video in lectures. 17 students (10 males and 7 female) watched the videos. The two versions of the video, A and B, were randomly assigned, 8 (3 female) and 9 (4 female) students, respectively.

3.4 Procedure

Each participant was tested individually according to the following procedure:

1. Receiving an introduction to the test procedure and the eye-tracker technique.
2. Calibration of the Tobii X2-60 to the participant's gaze.
3. Answering a questionnaire on demographics and study habits.
4. Watching the educational video.
5. Answering a questionnaire on how the participant experienced the video.
6. Answering specific questions on the topic of the video.

Participants were given the option to review the video with eye-tracker markings. The study was conducted at the Norwegian University of Science and Technology in Trondheim in March 2015.

4 Results

4.1 Captions

For the captions, the mean of TTFF was typically between 1.5-3 s for the different captions. The mean of TFT for all the captions was 2.3 s with a range of 1.5-3 s.

It should also be noted that a few (typical 2-4) participants did not fix their gaze on the different captions at all.

The different highlighting of the captions in the two versions, A and B, did not give any statistically significant differences in TTFF or TFT.

4.2 Written text

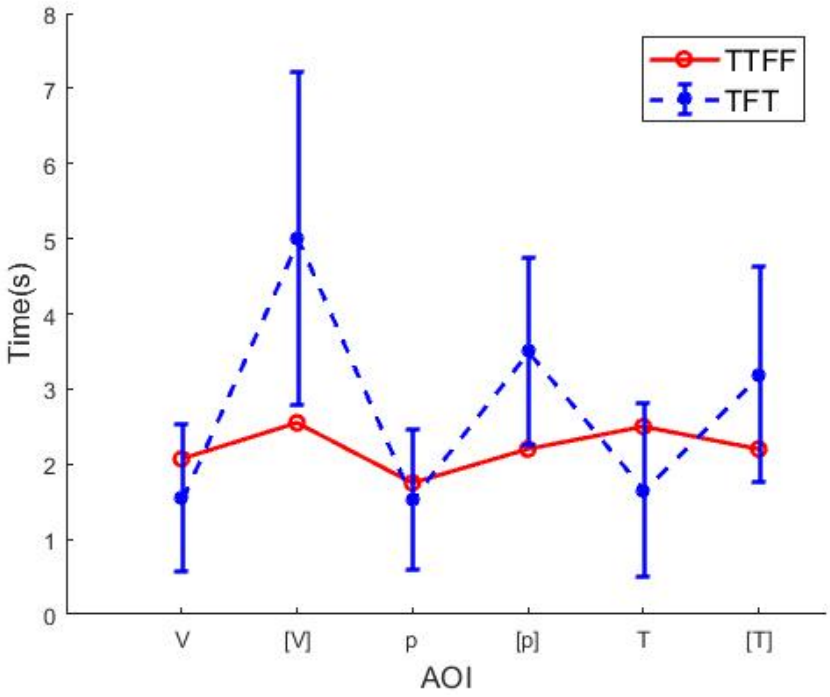

Figure 6. TTFF and TFT for the first panel (figure 3). The markers describe the sample median and the error bars the sample standard deviation.

Figure 6 shows TTFF and TFT for the different text segments in the first panel (figure 3). A median of about 2.1 seconds were found for the viewers to focus on the first text, the variable "V", their gaze scanning the scene before that. For the other text blocks, the mean median time from the first appearance to fixation was 2.0 seconds. The fastest fixation time of any viewer was about 1 second for all the different AOIs.

TFT for the different AOIs depends on both the amount of information and the time it takes to write the text. A short text, such as the variables, takes a short time to both read and write, giving almost the same fixation duration (around 1.5 s). More information (i.e. presenting the unit) resulted in a larger TFT, around 3 to 5 s.

The results for the next two panels, shown in figure 4 and 5, are presented in figures 7 and 8. TTFF is now more constant than in the first panel. In the second panel, the mean of the median was 2.2 s for all the AOIs, while it was 1.9 s for the third panel. There was a larger variation in TFT, with the duration seemingly correlated with the complexity of the mathematical notation and the amount of information. In panel 2 (figures 4 and 7) it was longest for AOIs 2, 3 and 5 (P_A, P_p and T_K) which represented more complex mathematical notation of the equations (exponentials and

summations). For panel 3 (figures 5 and figure 8), TFT was naturally much longer for AOI 3 (numbers) which contained a lot more information.

One participant had an anomalously long time to first fixation, several seconds delay compared to the second slowest participant's first fixation and was excluded from the analysis. This participant was the only clear outlier, while the second to last TTFF varied between the other participants. For the remaining data, the spread of the data can be roughly inferred from the sample standard deviation and ranges which are presented in figures 6 through figure 8.

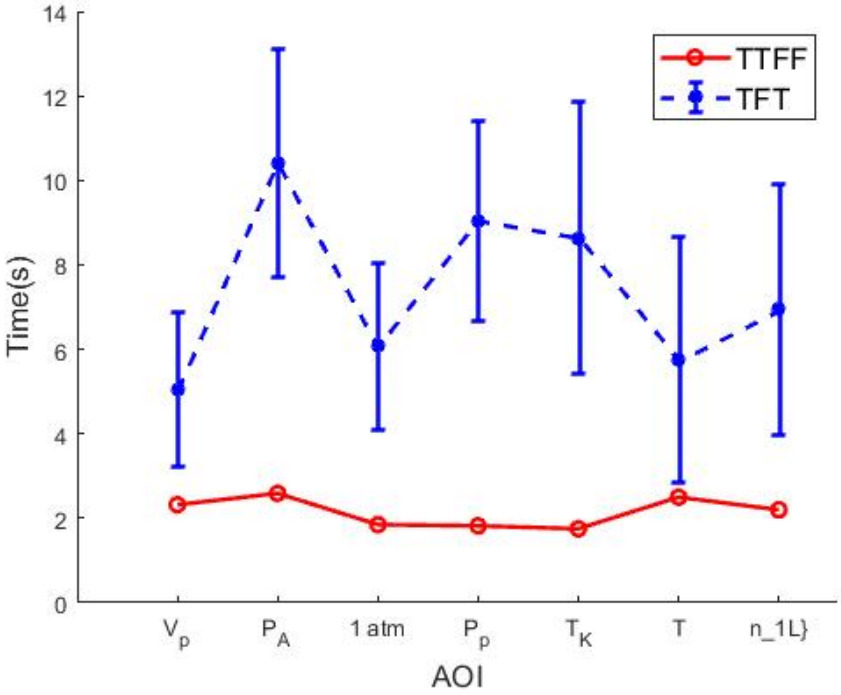

Figure 7. TTFF and TFT for the second panel (figure 4). The markers describe the sample median and the error bars the sample standard deviation.

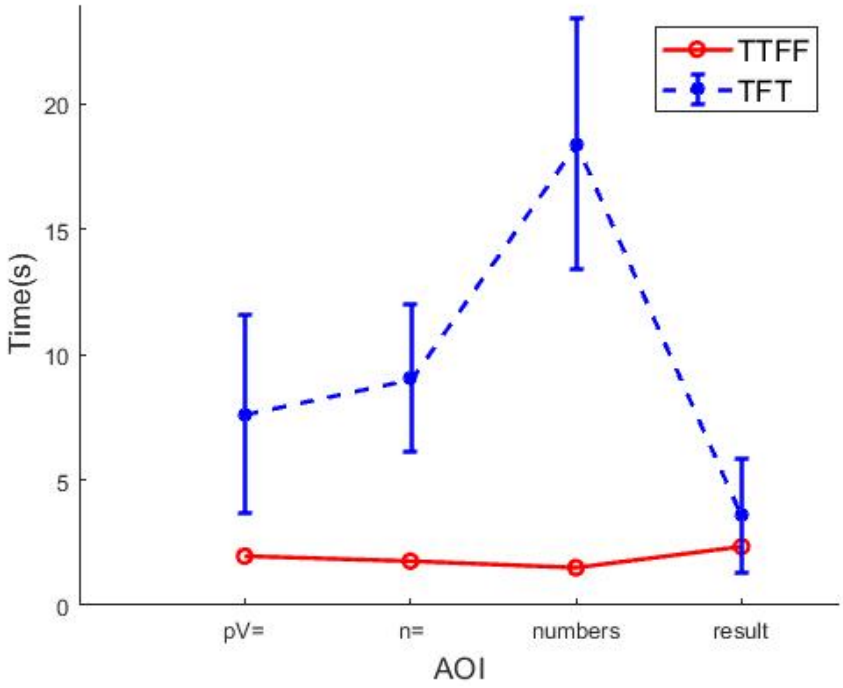

Figure 8. TTFF and TFT for the first panel (figure 5). The markers describe the sample median and the error bars the sample standard deviation.

4.3 Questionnaire

To get a picture of the volunteering students, as well as subjective opinions on the video and their learning experience, participants were asked to complete a general pre-questionnaire before watching the video and a post-questionnaire after. The intention was to get a picture of their attendance during lectures and their use of videos in their studies. All 16 students answered the questionnaire.

Participation in lectures where reported to be high, 13 reported over 75% attendance. Most (11 of 16) found attending lectures to be useful, while the others were neutral. A majority (14 of 16) reported that they took notes and that the main medium was paper (12 of 14, 2 used tablet). Nine reported not to prepare in advance for lectures, for example by reading in the textbook about the subject of the next lecture, while 12 reported to review their notes or read the textbook after lectures. 15 of the students had accessed videos (mainly lecture recordings from previous years or videos on YouTube), but 15 of 16 (which hade accessed videos) indicated the videos to be of

quite low (very little or some) value for them. When asked what type of video they would like, short thematic videos were preferred by 13.

After the video, participants were asked about their subjective impression of the video. The participants' opinions were in general positive. The content was judged as good by 13 and OK by the rest. Almost everybody (14 of 16) also considered the duration and difficulty as good. One participant found it too long and another too short. Two reported that they had difficulties following the presentation. The participants' opinion on value and perceived learning was quite positive, 12 found the video useful.

To check if the participants understood the content in the video, five questions on the content were included. The result indicates that most of the information got through, with less than 25% wrong answers to the first four questions (0%, 19%, 6%,25%). The fifth question was conceptually more difficult, with 38% correct, 44% partially correct and 18% wrong answers. The participants had to consider several statements where two were correct. Most participants only found one statement to be true and got this right but missed the other correct statement.

The questionnaire for the pre- and post-test, as well as the conceptual questions, are included in the supplementary material.

5 Discussion

We have applied the eye-tracker technique to study viewers' behaviour when reading captions and following the written text in a narrated educational video. Captions are important according to Mayer's signalling principle, and the temporal relationship between the time viewers gaze on the written text (as opposed to the time it is written) and the narration is important according to the temporal contiguity principle.

We observed a delay in the first fixation of un-cued written text and captions of about 2 seconds. As we use the captions to enforce the importance of certain concepts in the narration, this delay is intriguing. To capitalize on Mayer's signalling principle to enhance transfer of information to the working memory, it will be important to follow Meyer's temporal contiguity principle and keep this delay in mind when the caption is placed relative to the content and pauses in the narration. One could present the caption in advance (more than 4 frames) to the spoken word. However, this might be context dependent, if the viewers expect captions to appear or not, and should be

studied further before giving a more define recommendation, as the risk of introducing a distractor may be high. As TFT for the captions varied in the range 1-3 s it seems likely that the captions should at least be present for 4 s (for the amount of information presented in this video) to have the desired signalling effect. If the caption is presented for a shorter time, it might easily turn into a distractor. It would be an interesting avenue for future studies to see how this fixation time scales with increasing content, especially considering the amount of information that is often presented on (PowerPoint) viewgraphs in video designs.

The two versions of the videos were the captions appeared together or separate did not yield different results. This indicates that as long as the saliency of the captions is sufficiently strong this point is of minor importance. However, too high saliency might become a distractor.

In the case of writing text, the results suggest that without any cue there is a significant delay before the viewers focus on the important information. This is especially true for the first AOI of the written text (figure 6) where the viewers have received no information or cue as to where to look. For the subsequent AOIs and panels, the viewers know where to look and the time to focus is shorter. It should be noted that the video studied is a mixture of the presenter on camera and screenshot, where there is a shift in the centre of attention, from the presenter's face to the point where the text appears. This indicates that it is advantageous when designing a video to include some sort of cue for where the text is to appear, especially if there is a shift in the scene (from presenter to tablet) to aid the viewers in focusing on the important information. A hand or a coloured spot could be used initially and then faded out or removed as to avoid distracting the viewers.

The study by Fiorella and Mayer (2016) on presenting a ready-made diagram versus drawing a diagram found positive effects on the performance for low prior knowledge learners in the case of drawing a diagram, with no or negative effect on high prior knowledge learners. They also observed positive effects (size effect of 0.35) when the instructor or the instructor's hand were visible while drawing. This supports our findings on the value of the use of cues and the need to consider the temporal aspect of the viewers' attention, perhaps especially for viewers with low prior knowledge that need more time to process information.

The answers from the questionnaire indicated that the students were in general quite happy with the video and that the majority preferred short thematic videos and

that the length of the video was appropriate. Especially, in light of most educational videos being unedited recordings of standard lectures, investigating how different types of videos are used will be important for future investigations. However, with such a small group of students (about 15% of all students attending the course), we cannot assume these to be a representative group. As for the perceived learning, one must remember that the group was heterogeneous and that some students may have studied the subject before, explaining why one participant indicated very little learning, while still being very positive to the video.

5.1 Limitations

The study was performed with students watching a video in a laboratory context. This is not a natural situation, as students normally watch videos in a more active way, with, for example, pausing or rewinding. As this was not possible due to present limitations in the eye-tracking software, the results might differ from a natural situation. The experimental situation might also cause the participants to make an extra effort to focus on the video. It is therefore expected that the result can be considered as a best-case scenario. The sample size and selection (voluntary) of participants is also a source of uncertainty. The students who volunteered might be high achievers with a good study technique and therefore less sensitive for design effects (individual differences principle, Mayer, 2014).

5.2 Future work

We intend to repeat this study using a video with cued written text to investigate the effect of cuing. The design and production of these videos are underway. In addition, we also plan to study the effect of using text-blocks instead of written text to investigate if this affects the viewers' focus. As we have two information processing channels, the effect of timing between text and speech is particularly interesting when it comes to fixation. The questions on the efficiency of special design on and eye-movements of high-knowledge viewers, the individual differences principle and prior knowledge principle (Mayer, 2014) remain an interesting question, which should be addressed in a study with more participants.

6 Conclusion

The main conclusion of this paper is that there is a significant delay between when an object appears in a video and when the viewers focus on this object. When designing educational videos this delay will be important to keep in mind. In addition, when there is a significant change in the scene the delay increases, which indicates that in such cases cueing might be important to guide the viewers' attention or to allow viewers to catch up before presenting new information.

References

Baddeley, A. D. (1986). *Working memory*. Oxford, England: Oxford University Press.

Ballantyne, N. (2008). Multimedia learning and social work education. *Social Work Education,* 27(6), 613–622.

Bos, N., Groeneveld, C., Bruggen, J., & Brand-Gruwel, S. (2016). The use of recorded lectures in education and the impact on lecture attendance and exam performance. *British Journal of Educational Technology*, *47*(5), 906–917.

Brooks, G., Hughes, J., Ritchie, L., Roberts, S., Wright, K. (2005). *Digital Beginnings: Young children's use of popular culture, media and new technologies*. The University of Sheffield.

Burgoon, J. K., Guerrero, L. K., & Floyd, K. (2016). *Nonverbal communication*. Routledge.

Buswell, G. T. (1935). *How people look at pictures: a study of the psychology and perception in art*. Oxford, England: Univ. Chicago Press.

Canham, M., & Hegarty, M. (2010). Effects of knowledge and display design on comprehension of complex graphics. *Learning and instruction*, *20*(2), 155–166.

De Koning, B. B., Tabbers, H. K., Rikers, R. M., & Paas, F. (2009). Towards a framework for attention cueing in instructional animations: Guidelines for research and design. *Educational Psychology Review*, *21*(2), 113–140.

Eisenstein, S. (1929). *"The Dramaturgy of Film Form (The Dialectical Approach to Film Form)"*, Taylor, R., ed., The Eisenstein Reader. The British Film Institute.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Fiorella, L., & Mayer, R. E. (2016). Effects of observing the instructor draw diagrams on learning from multimedia messages. *Journal of Educational Psychology*, *108*(4), 528.

Jonides, J., & Yantis, S. (1988). Uniqueness of abrupt visual onset in capturing attention. *Perception & psychophysics*, *43*(4), 346–354.

Just, M. A., & Carpenter, P. A. (1984). Using eye fixations to study reading comprehension. *New methods in reading comprehension research*, 151–182.

Lai, M. L., Tsai, M. J., Yang, F. Y., Hsu, C. Y., Liu, T. C., Lee, S. W. Y., ... & Tsai, C. C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational research review*, *10*, 90–115.

Liversedge, S., Gilchrist, I., & Everling, S. (Eds.). (2011). *The Oxford handbook of eye movements*. Oxford University Press.

Manovich, L. (2001). *The language of new media*. MIT Press.

Mayer, R. E., Fennell, S., Farmer, L., & Campbell, J. (2004). A personalization effect in multimedia learning: Students learn better when words are in conversational style rather than formal style. *Journal of Educational Psychology, 96*(2), 389.

Mayer R.E. (2006), *Multimedia Learning*, Cambridge University Press, ISBN: 978-0-521-51412-5

Mayer R.E. (2010), *Unique contribution of eye-tracking research to the study of learning with graphics, Learning and Instruction* Vol. 20, p167–171.

Mayer R.E. (ed.) (2014), *Cambridge Handbook of Multimedia Learning*, Cambridge University Press, ISBN: 978-1-107-03520-1

NTNUOpenvideo (2015) http://video.adm.ntnu.no/pres/550824c785d35 (Accessed 2016-07-07)

O'Callaghan, F. V., Neumann, D. L., Jones, L., & Creed, P. A. (2017). The use of lecture recordings in higher education: A review of institutional, student, and lecturer issues. *Education and Information Technologies, 22(*1), 399–415.

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford, England: Oxford UniversityPress.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372.

Rosenberg, J. (2010). *The Healthy Edit: Creative Editing Techniques for Perfecting Your Movie*. Focal Press.

Scheiter, K., & Eitel, A. (2016). The use of eye tracking as a research and instructional tool in multimedia learning. *Eye-tracking technology applications in educational research*, 143–164.

Shah, M., Sid Nair, C., & Bennett, L. (2013). Factors influencing student choice to study at private higher education institutions. *Quality Assurance in Education*, *21*(4), 402–416.

Sorden, S. D. (2012). The cognitive theory of multimedia learning. *Handbook of educational theories*, 155 –168.

Sweller, J. (1988). Cognitive load during problem-solving: Effects on learning. *Cognitive Science, 12*, 257–285.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4,* 295–312.

Thompson, K., Bordwell, D. (2010). *Film History: An Introduction. 3rd Ed.* McGraw-Hill.

Tobii, (2016) http://www.tobiipro.com/ (Accessed 2016-07-07)

Wang, F., Li, W., Mayer, R. E., & Liu, H. (2018). Animated pedagogical agents as aids in multimedia learning: Effects on eye-fixations during learning and learning outcomes. *Journal of Educational Psychology*, *110*(2), 250.

Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum.

Supplementary material – Questionnaire

Pre-questions and answer alternatives

Sex: Male, Female

How many lectures do you attend? – <25%, 26-49%, 50-74%, 75-99%, 100%

How useful are lectures for you?–Not at all, Little, Somewhat, Useful, Very useful

Do you normally take notes?–Yes, No

If you take notes, on which medium do you use? – Own papers, on handouts, on PC, on Tablet

Do you prepare for lectures? –For example, reading in the Textbook on the subject of the lecture. Yes, No

Do you review your notes/ read the Textbook after the lecture? – Yes, No

Have you accessed educational videos? – Yes, No

If you have accessed educational videos, did you find them useful? – Yes, all of them, Yes some, No

Do you want to be able to access educational videos at our university? – Yes, No

In which format do you want the videos? – Whole recorded lectures, Thematic videos, Other

Post-questions:

What was your impression of the video? – Very bad, Bad, OK, Good, Very Good

How was the duration? – Too short, fine, Too long

Was the video difficult to understand? – Yes, No

Were you able to focus on the presentation? –Yes, No

Was the presentation clear? –Yes, No

How was the presenter? – Very bad, Bad, OK, Good, Very good.

How valuable was the video for you? – Very little, Little, OK, Much, Very much.

Did you learn anything new? – Very little, Little, Some, Much, Very much.

Questions on content-attention

We have two statements:

A: The number of molecules in the same volume with the same temperature and pressure, are the same for all gases.

B: The number of molecules in a mole is 6,02205 * 10^{23} for all gases.

Which of these answers are correct? – Both A and B; Only A; Only B; Both A and B are wrong.

Gas in a closed container is heated. What happens to the pressure in the container?

– Pressure increases;Pressure decreases; Pressure is contant; The question can not be answered with given data.

Which of the following systems contains most molecules? A: p = 1 atm, V = 1l, T = 300K; B: p = 2 atm, V = 1l, T = 300K; C: p = 1 atm, V = 1l, T = 600K; D: p = 2 atm, V = 1l, T = 600K

Which of the following are not a state variable?

– Pressure; Volume; Temperature; the number of moles; Number of molecules; Molar mass.

What is an equation of state? An equation which describes.

– how the state varies with time;
– the temperature at a specific moment of time;
– the relations between state variables in a system;
– the state.