# Are Captions in Video Tutorials a Bad Idea?

**Yihe Wang[1] and Chris Evans[2]**

[1]University College London (UCL), UK

[2]UCL Interaction Centre, University College London (UCL), UK

## Abstract.

The use of video tutorials is becoming increasingly prevalent in education. To make them accessible to a wider audience, it has also become common practice to produce the videos with captions that duplicate the words being spoken. When added, captions are presented visually as on-screen text simultaneously with audio information (narration). This is known as redundancy because the same information is communicated twice. Current research in multimedia learning suggests that this redundancy may result in cognitive overload, causing a decrease in memory and understanding. The current study considers evidence for the redundancy principle for captioning which states that for video tutorials, people learn better with captioning turned off. In an experiment, participants were randomly divided into a caption group and a no-caption group. The caption group watched an instructional video with a narrated visual demonstration and captions turned on. The no-caption group watched the same video with captions turned off. After watching the video, learning was assessed through three types of tests: retention, matching and transfer. The results showed evidence that the use of captions resulted in a decrease in understanding for the caption group in at least some circumstances. The study concludes by discussing the circumstances in which captioning may have negative impact and also considers those circumstances in which it may have a positive impact on learning.

**Keywords:** accessibility, cognition, multimedia learning, redundancy effect, subtitles

## 1. Introduction

Screen-recorded video tutorials have gained increasing popularity in education. By recording the screen whilst narrating a task, educators can create simple videos to teach concepts or explain procedures. The software and equipment to do this is now readily available and with sites such as YouTube™ (Google, 2020) and Vimeo™ (Vimeo, Inc., 2020) it is easy to publish and distribute video tutorials to a large audience. To increase accessibility, it has also become common practice to add captions that duplicate the words being spoken. These may take the form of *closed captions* that the learner can toggle on or off; *open captions* which are permanently added to the video; or *subtitles* in which the narration is translated into another language. With developments in speech recognition these captions can often be auto-generated.

Video tutorials usually combine two modes of presentation: words (spoken by the narrator) and pictures (of the screen in the case of screen-recording). Learning from video tutorials is therefore known as *multimodal* or more commonly *multimedia* learning. A lot of research has already been conducted into multimedia learning. For example, Mayer (2009) has established

principles such as the multimedia principle and the modality principle. One principle, the redundancy principle, has particular relevance to the use of captions. The redundancy principle states that people learn better from pictures and narration than from pictures, narration and written words. The basis for this principle is explained in the next section. Applied to the special case of captions in video tutorials, this principle states that people learn better from pictures and narration than from pictures, narration and captions. In other words, the use of captions should result in a decrease in learning rather than an increase. The research question that this study considers, therefore, is "Are captions in video tutorials a bad idea?".

The structure of this paper is as follows. It begins by reviewing research into multimedia learning with particular reference to cognitive theory. An empirical study is presented in which participants watch video tutorials with and without captions. This is followed by a discussion of the extent to which the results of the study support the redundancy principle for captioning, which states that for video tutorials people learn better with captioning turned off. The conclusion discusses the types of circumstances in which captioning has a negative impact and also consider types of circumstances in which it may have a positive impact on learning.

## 2. Cognitive Theory

### 2.1 Multimedia Learning

It is well known that human beings process information in working memory with limited capacity (Baddeley, 1986). There are two historically significant theories about the processing of multimedia information, i.e. information combining words and pictures. Baddley (1992) takes a sensory-modality perspective in which visual information (from the eyes) is processed separately from audio information (from the ears). Paivio's (1991) dual-coding theory takes a presentation-mode perspective in which visual and non-verbal audio information is processed separately from verbal information. More recently, Mayer (2009) has proposed a compromised model in which auditory/verbal information is processed separately from visual/pictorial information. All three theories posit that words are processed separately from pictures. This is known as the *dual channel assumption* (Mayer, 2009).

According to Mayer (2009), in combination with the working memory assumption derived from Baddeley (1986), this dual channel assumption gives rise to a number of empirically supported learning principles. Three of them are of particular relevance to video tutorials. The first, the *multimedia principle*, states that people learn better when information is presented as words and pictures, rather than words alone or pictures alone. The theoretical basis for this principle is that when presented with both words and pictures learners are able to build both a verbal and a visual mental model and build connections between them. Mayer (2009) reports eleven out of eleven experiments yielding statistically reliable results in support of this principle. Applied to video tutorials, this principle tells us that narrated videos are better than audio narration alone.

The second relevant principle is known as the modality principle. The modality principle states that people learn better from pictures and spoken words than from pictures and written words. The theoretical basis for this principle is that written words as well as pictures are partially processed by the visual/pictorial channel, causing the channel to become overloaded. Mayer (2009) reports seventeen out of seventeen experiments in support of the principle. Applied to video tutorials, this principle tells us that videos with spoken text are better than videos with just written text, i.e. captions.

The third and most relevant principle to this study is the redundancy principle. The redundancy principle states that people learn better from spoken words and pictures than from spoken words, pictures and written text. The theoretical basis for this principle is that spoken words and written words are both processed by the auditory/verbal channel, causing it to become overloaded. In general, redundancy in which duplicate information is processed by a single channel results in overloading. Mayer (2009) reports five out of five experiments in which this principle is supported. Applied to video tutorials, the redundancy principle for captioning states that for video tutorials people learn better with captioning turned off. These principles support a view of learning known as knowledge construction.

### 2.2 Information Delivery versus Knowledge Construction

Mayer (2009) distinguishes two opposing views of multimedia learning. In the first, information delivery, the process of teaching is the process of delivering information to the learner. By the same token, the process of learning is the process of acquiring information. This view is sometimes known as the empty vessel view or the blank tablet view as the learner is seen as a vessel into which information is to be poured or a blank tablet (page) onto which information is written. The teacher, under this view, is sometimes characterised as the sage on the stage (King, 1993) because their role is to provide the expert information to be delivered. The second contrasting view (favoured by Mayer) is known as knowledge construction. In this, the process of teaching is the process of trying to help the learner make sense of the material. Learning is the active process of selecting, organising and integrating information. The teacher is then characterised by King (1993) as the guide on the side.

Under the information delivery view, the use of captions in addition to spoken narration should either have no effect because the information content of the two media is the same, or it should increase learning because there is greater opportunity to acquire the information. By contrast, the knowledge construction (constructivist) view, coupled with the dual channel assumption, predicts that learning will decrease due to the need to process redundant information.

The constructivist view suggests a redundancy principle for captions which predicts that students who watch a video tutorial without captions should learn more than students who watch the same video with captions. This study aims to put this principle to the test.

## 3. Methodology

The research hypothesis for the study is that learners watching video tutorials without captions will outperform learners watching the same video with captions in tests of their learning. The study involved a between-subjects design in which participants were randomly divided into two groups. Each group watched a short video and was then given identical tests of their knowledge and understanding of the information presented in the video. The chosen video tutorial introduced a specific game development platform (details below). In terms of the mean test scores for the no-caption group $m_{NC}$ and caption group $m_C$, the experimental hypothesis was thus:

$$H_1: m_{NC} > m_C$$

This hypothesis is consistent with the knowledge construction view of learning and inconsistent with the information delivery view. By contrast, the null hypothesis is that students

in the group with captions should outperform the group with captions in tests or there should be no difference:

$$H_0: m_C \geq m_{NC}$$

### 3.1 Participants

Thirty-three participants were recruited via online postings. Eight were male, twenty-four were female, and one preferred not to say. Ages ranged from 20 to 35 years with a median age of 23. The participants majored in a range of disciplines including Accounting and Finance, Architecture, Arts, Business, Chemical Engineering, Commerce, Computer Science, Design, Engineering, Human-Computer Interaction, Information Science, Material Science, Pharmacy, Sociology, and Sustainable Development. All participants were at or above average proficiency in English. Approximately 18% (6 out of 33) of the participants were native English speakers. Five of these were in the no-captions group and one in the captions group. All participants indicated that they had no prior experience or knowledge in using the chosen software package, as established through a self-report participant questionnaire. Approval for the study was granted by the UCLIC Research Ethics Committee, reference UCLIC/1718/011/Staff Evans.

### 3.2 Design

Using the presence or absence of captions as a factor, a between-subjects experimental design was used to investigate the redundancy effect for captions. The independent variable was therefore instructional format (video tutorial with captions or video tutorial without captions). The dependent variables were test performance scores in three types of test question (retention, matching and transfer) as well as total completion time (the time each learner spent on completing all three tests). Data were collected over a two-week period.

### 3.3 Materials

Each person participated remotely in the study through a link from his or her own computing device (desktop or laptop), in his or her own environment, with computer-enforced rules but without any human supervision. The online experiment was established on Gorilla™ (Cauldron Science, 2020), an online experiment builder. For each participant, the computer-based materials consisted of a participant questionnaire, a video tutorial, a retention test, a matching test, and a transfer test. The video tutorial introduced the software package Unity™ (Unity Technology, 2018), a game development engine.

#### 3.3.1 Participant questionnaire

The participant questionnaire solicited information about gender, age, and field of study. In addition, a five-item scale was used for participants to rate their knowledge of Unity: very little, between very little and average, average, between average and very much, and very much. The scale was designed to ensure that none of the participants had previous exposure to Unity and thus had low prior knowledge of the domain taught by the lesson. In addition, participants' preference towards captions was asked in the final question.

#### 3.3.2 Screen-capture video tutorial

A 13-minute video was used in this experiment. The video was an edited version of one of the most popular Unity tutorials on YouTube. The topic of the video was "How to build a 2D game". The instructor presented the tutorial by narrating the performance of a real task. The

video was edited to reduce its length and to focus on the Unity game engine. Two versions of the video were created: one with open captions and one without.

### 3.3.3 Post-tests

For both groups of students, a set of questions was designed to collect quantitative data of their knowledge and understanding of the video. This consisted of ten multiple-choice questions, each with 4 options, related to Unity 2D game development. The questions were designed to reflect relevant content that users might encounter when operating the software in real life. Three types of questions were developed: a retention test (7 questions), a matching test (1 question), and a transfer test (2 questions).

The retention test contained seven questions that required remembering information presented in the video. For example, one of the questions was "When you create a new project for your 2D game, the most important thing you should make sure is that...". The options were (a) "Change the project name to '2D Game example'", (b) "Change the Unity version to '2018.2.0b7'", (c) "Save the project to your projects folder", and (d) "Select 2D Template". This question required participants to retrieve the author's explanation when demonstrating the creation of a new project.

The matching test contained one question requiring participants to recall the name of a visual image. The matching test presented one element from the software demonstration along with the following question: "What is this called?", followed by a list of four items: (a) "Tile workflow," (b) "Tile set", (c) "Tile map", and (d) "Tile palette".

The transfer test was designed to test higher-order learning by asking learners to consider a situation they had not met before. It consisted of the following two questions: (T1) "The coin is completely overlaid by a platform. Which of the following guesses is impossible?" and (T2) "In Super Mario Bros, we can see the secret bricks (the one with a question mark). What game object physics might be applied?" Both questions were illustrated by accompanying graphical representations.

The tests were also implemented in Gorilla. The responses to each question and the time it took the learners to complete the tests were recorded electronically by the system. Each question was worth one point, and all answers were calculated using Microsoft Excel™, providing each learner with a calculated final score out of ten.

## 3.4 Procedure

The study was conducted entirely online. Participants independently received a link to the online experiment and accessed it on a computer (desktop or laptop). The tutorial system needed to be loaded in the Chrome web browser and was presented in approximately 45% screen mode.

The experiment consisted of three phases: a questionnaire phase, an instruction phase, and a test phase. After providing informed consent, participants received an instruction and a brief timeline for the experiment. They were told that for the main task they would watch a thirteen-minute Unity tutorial and they were going to complete ten Unity related multiple-choice questions afterwards.

In the questionnaire phase, participants completed a short demographic questionnaire and a participant survey asking about their knowledge of Unity. Then participants were randomly assigned into one of two groups (no-caption group and caption group).

In the instruction phase, participants watched the video tutorial. The viewer control was turned off during the whole process, which means participants were not able to fast-forward, pause or replay the video.

In the test phase, participants were given ten multiple-choice questions to complete. Participants did so at their own pace and were not able to turn back to the video tutorial page during the experiment. They were also instructed not to refer to any other information resources during the tests. The entire process lasted approximately 30 minutes.

## 4. Results and Analysis

Two outliers were eliminated from the data set based on scatter point analysis. This resulted in 31 participants in the final analysis.

Table 1 summarises the mean scores and standard deviations for the two groups on each of the three tests. For each dependent measure (retention, transfer, and matching) the data were subjected to an unpaired Student's *t*-test with the between-subjects factor being captions or no-captions.

*Table 1: Test scores*

| **Group** | **Retention (out of 7)** | | **Matching (out of 1)** | | **Transfer (out of 2)** | |
|---|---|---|---|---|---|---|
| | Mean | Std. Deviation | Mean | Std. Deviation | Mean | Std. Deviation |
| **No Captions (NC) n=16** | 4.125 | 1.500 | 0.438 | 0.512 | 1.125 | 0.806 |
| **Captions (C) n=15** | 3.533 | 1.598 | 0.533 | 0.516 | 0.733 | 0.799 |

### 4.1 Post-test scores

Whilst the captions group (C) scored lower than the no-captions group (NC) in two of the three tests, these differences were not statistically reliable. Parametric analyses revealed no significant difference in the retention test results (Student's *t*-test, $t(29)=1.064$, one-tailed, p=.296), or transfer test results (Student's *t*-test, $t(29)=1.358$, one- tailed, p=.185).

Table 2 presents the results for individual questions.

*Table 2: Individual question scores*

| **GroupP** | **R1** | | **R2** | | **R3** | | **R4** | | **R5** | | **R6** | | **R7** | | **M1** | | **T1** | | **T2** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M | STD | M | STD | M | STD | M | STD | M | STD | M | STD | M | STD | M | STD | M | STD | M | STD |
| **NC** | 0.438 | 0.512 | 0.938 | 0.250 | 0.313 | 0.479 | 0.563 | 0.512 | 0.625 | 0.500 | 0.688 | 0.479 | 0.563 | 0.512 | 0.438 | 0.512 | 0.438 | 0.512 | 0.688* | 0.479 |
| **C** | 0.600 | 0.507 | 0.800 | 0.414 | 0.133 | 0.352 | 0.400 | 0.507 | 0.600 | 0.507 | 0.667 | 0.488 | 0.333 | 0.488 | 0.533 | 0.516 | 0.400 | 0.507 | 0.333 | 0.488 |

***Note. NC = No-captions group, C = Captions group, M = Mean, and STD = Standard Deviation.***
***R1 to R7 are retention questions, M1 is a matching question, and T1 and T2 are transfer questions.***
**** p < .05.***

It can be seen from Table 2 that there was a significant difference between scores for the second transfer question T2 (unpaired Student's *t*-test, $t(29) = 2.039$, one-tailed, $p = .025$). This has a medium effect size of d = 0.733. This provides some support for the experimental hypothesis $H_1$ and thus partial evidence for the redundancy principle for captioning.

## 5. Discussion

The aim of this study was to determine whether there was evidence for a redundancy principle for captions. This principle states that students who watch a video tutorial without captions should learn more than students who watch the same video with captions. The study tested this hypothesis using a short video tutorial watched by 31 participants divided into two groups. Learning was assessed using a set of ten multiple-choice questions assessing knowledge, matching and transfer (deep learning). One of the two transfer questions showed a statistically significant increase in scores for the group with no captions. This shows partial evidence for the redundancy principle for captions. However, there was no overall effect on scores for the no-captions group. Further studies are needed to see if stronger evidence can be found.

Mayer's original studies into the redundancy principle involved video tutorials in which either (a) animations of lightning formation were narrated (Mayer, Heiser & Lonn, 2001; Moreno & Mayer, 2002a), or (b) animations of plant growth were narrated (Moreno & Mayer 2002b). In both scenarios the visual element consists of a series of evolving static images. By contrast, the present study used screen-recording to show real-life use of a software package. It is possible that animations have a greater redundancy effect than videos because their visual elements are more distinct. Further research is needed to determine whether there is any difference between animations and video.

There are two other possible explanations for the redundancy effect being weak. The second possible reason is that the instructor in the videos that were used in the current experiment usually made the corresponding demonstration before giving a detailed explanation. When the demonstration and redundant textual information (narration and captions) are presented in a sequential manner, the positive impacts of redundant explanations may remain (Moreno & Mayer 2002a). Although on-screen text and pictures are both processed in the visual/pictorial channel, the sequential presentation avoids visual working memory overload (Moreno & Mayer, 1999). This is also observed in the study by Moreno and Mayer (2002a). While the instructional materials used in the present experiment were not clearly split up into chunks of explanation and demonstration, the instructional strategy used by the author of the tutorial was to firstly demonstrate the operation and then explain the reason or concept. The crucial information was rarely presented at the same time as the author explained things. This opens up the interesting possibility that the redundancy effect for captions is strongest when captions are presented simultaneously with the narration.

A third possible explanation for the weakness of the effect is that some of the test questions were not sensitive enough to assess learning gains. All learners scored high on one retention test (see Table 2), which might cause a ceiling effect (Moreno & Mayer, 1999; Moreno & Mayer 2002).

Are there circumstances in which captions might have a positive effect? The redundancy effect assumes that learners can learn from both words and pictures. There are two obvious circumstances in which this assumption might be weakened. The first is in the case of visually-impaired learners, the study of whom falls outside the scope of this research. The second is the case of learners who are not native English speakers. For these learners it is possible that having both spoken and written English text might not have the redundancy effect that can be observed for native English speakers. However, a study by Diao and Sweller (2007) suggests that the redundancy effect still applies to non-native groups of learners. In their research they considered reading comprehension for English as a Foreign Language (EFL) students. They compared groups with written presentation and written plus spoken presentation. They found that the

written-only presentation group performed better in translation scores, subjective mental load ratings, and free recall performance. In our own study, six of the participants were non-native English speakers with average or above-average English proficiency. The study did not control for the number or distribution of these participants. Future studies are planned to conduct controlled experiments comparing learning from video tutorials for native and non-native English speakers.

## 6. Conclusion

This study considered whether, contrary to common practice, adding captions to video tutorials might be a bad idea. An experiment was conducted in which two groups of participants watched the same video with and without captions. The group without captions performed better in one of two questions designed to test the ability to transfer the knowledge they had learned to a situation they had not met before. This kind of test is usually regarded as a test of understanding (deep learning) as opposed to rote memory. This result provides partial evidence in support of the redundancy principle for captions which states that for video tutorials people learn better with captioning turned off. It also provides support for the view of learning as knowledge construction rather than information delivery. The results are weak in the sense that they were not apparent for other questions in the test. However, they are statistically reliable and have a medium effect size. Further studies are needed to investigate whether this effect is different for non-native English speakers and is affected by the timing and format of captions.

**Acknowledgments**

The experimental procedures were approved by the University College London Interaction Centre (UCLIC) ethics committee. Both authors have reviewed the contents of the manuscript and approve of its contents and validate the accuracy of the data.

**References**

Baddeley, A. D. (1986). *Working memory*. Clarendon.

Baddeley, A. D. (1992). Working memory. *Science*, 255, 556-559.Clarendon.

Cauldron Science. (2020). *Gorilla Experiment Builder*. https://gorilla.sc/

Diao, Y., & Sweller, J. (2007). Redundancy in foreign language reading comprehension instruction: Concurrent written and spoken presentations. *Learning and Instruction*, *17*(1), 78–88.

Google. (2020). *YouTube.* https://www.youtube.com/

King, A. (1993). From sage on the stage to guide on the side. *College Teaching*, *41*(1), 30-35.

Mayer, R. E. (2009). *Multimedia learning*, Second Edition. Cambridge University Press.

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive Constraints on Multimedia Learning: When Presenting More Material Results in Less Understanding. *Journal of Educational Psychology*, *93*(1), 187–198.

Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. Journal of educational psychology, 91(2), 358.

Moreno, R., & Mayer, R. E. (2002a). Verbal Redundancy in Multimedia Learning: When Reading Helps Listening. *Journal of Educational Psychology*, *94*(1), 156–163.

Moreno, R., & Mayer, R. E. (2002b). Learning Science in Virtual Reality Multimedia Environments: Role of Methods and Media. *Journal of Educational Psychology*, *94*(3), 598–610.

Paivio, A. (1991). Dual Coding Theory: Retrospect and Current Status. Canadian Journal of Psychology, 45(3), 255–287.

Unity Technology. (2018) *Unity Real-Time Development Platform* (Version 2018.1.1f1) [Computer software]. https://unity.com/

Vimeo, Inc. (2020). Vimeo. https://vimeo.com/