Learning and Instruction 77 (2022) 101542

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

# Does adding video and subtitles to an audio lesson facilitate its comprehension?

Yueyuan Zheng [a], Xinchen Ye [a], Janet H. Hsiao [a,b,*]

[a] Department of Psychology, University of Hong Kong, Hong Kong Special Administrative Region
[b] The State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong, Hong Kong Special Administrative Region

ARTICLE INFO

*Keywords:*
Multimedia learning
Eye movement
Hidden markov model
EMHMM

ABSTRACT

We examined whether adding video and subtitles to an audio lesson facilitates its comprehension and whether the comprehension depends on participants' cognitive abilities, including working memory and executive functions, and where they looked during video viewing. Participants received lessons consisting of statements of facts under four conditions: audio-only, audio with verbatim subtitles, audio with relevant video, and audio with both subtitles and video. Comprehension was assessed as the accuracy in answering multiple-choice questions for content memory. We found that subtitles facilitated comprehension whereas video did not. In addition, comprehension of audio lessons with video depended on participants' cognitive abilities and eye movement pattern: a more centralized (looking mainly at the screen center) eye movement pattern predicted better comprehension as opposed to a distributed pattern (with distributed regions of interest). Thus, whether video facilitates comprehension of audio lessons depends on both learners' cognitive abilities and where they look during video viewing.

## 1. Introduction

### 1.1. Literature review

With technology advances, the use of multimedia has become a popular means for learning. Multimedia learning refers to knowledge construction from both verbal and pictorial information, with the verbal form including spoken words or printed texts, and the pictorial form including pictures or videos (Mayer, 2014a). According to the modality principle, it is generally beneficial to receive visual and audio information than visual and on-screen text information during learning (Low & Sweller, 2014). Indeed, in academic learning, participants learning through multimedia displayed better knowledge acquisition and improved content comprehension than those who learned through text reading or traditional lectures (Starbek et al., 2010). The cognitive theory of multimedia learning further posits that learning would be facilitated if different sources of information are received from independent channels, but would be undermined if multiple sources of information are received from the same perceptual channel (Mayer, 2014b). For example, students viewing computer animations showed better comprehension when concurrently listening to narration than when concurrently viewing on-screen text (Moreno & Mayer, 1999).

Performance during multimedia learning also depends on the level of redundancy between two independent channels. For example, Moreno and Mayer (2002) compared learning from auditory explanations with and without additional on-screen text to examine the effect of verbal redundancy and found that reading enhances listening comprehension. A recent meta-analysis study on verbal redundancy showed that students with spoken-written presentation learned better than those with spoken-only or written-only presentation (Adesope & Nesbit, 2012). Similarly, redundancy between the visual and audio information was found to enhance learning as compared with receiving either one alone (Moreno & Mayer, 2002), as noted in the multimedia principle (Fletcher & Tobias, 2005). Mayer and Anderson (1992) compared animations with narration against animation-only and narration-only conditions to test multimedia principle and found that it is beneficial to receive both information in learning. Note however that when multimedia learning involves more than two sources of information, such as presenting visual animation and audio narration with concurrent on-screen text, redundancy between the narration and the text may hurt rather than help learning, as on-screen text may create competition with visual information from the animation. For example, Mayer et al. (2001) reported

* Corresponding author. Department of Psychology, University of Hong Kong, Room 623, 6/F Jockey Club Tower, Pokfulam Road, Hong Kong Special Administrative Region.
*E-mail address:* jhsiao@hku.hk (J.H. Hsiao).

https://doi.org/10.1016/j.learninstruc.2021.101542
Received 29 September 2019; Received in revised form 2 August 2021; Accepted 6 September 2021
Available online 20 September 2021

that students viewing an educational animation with concurrent narration had better comprehension without subtitles than with subtitles. This effect has been referred to as the redundancy principle (Mayer et al., 2001; Hoffman, 2006). In situations where learners have sufficient opportunities to process important visual information, verbal redundancy between audio narration and on-screen text may still be helpful. For example, watching recorded lecture videos with subtitles was associated with better comprehension performance (Kruger & Steyn, 2014). Mayer and Johnson (2008) showed that when learning scientific knowledge from narrated slide presentations, students who received short redundant on-screen phrases outperformed those who did not on knowledge retention but not on transfer (see also McCrudden et al., 2014). Consistent with this finding, same-language subtitles in video advertisements were shown to enhance the viewers' memory of the brand and slogan (Brasel & Gips, 2014). Subtitled videos were reported to create a lower cognitive load, as indicated by pupil diameter change, than unsubtitled versions (Kruger et al., 2013). These findings suggest the facilitation of verbal redundancy (subtitles and audio narration) in the presence of video information.

However, whether redundancy facilitates or impairs learning remains controversial, especially when multimedia materials with redundant information demand the learners to allocate their attention to various sources of information (Sweller, 2005; Florax & Ploetzner, 2010). For example, the multimedia principle, i.e., the facilitation effect of combining word and picture information, has been shown to depend on multiple factors, including the content to be learned and the learner's general learning ability (Fletcher & Tobias, 2005). According to the cognitive load theory, the effect of redundant information on learning depends on the cognitive resources it requires to extract and integrate the information (Adesope & Nesbit, 2012). Thus, learners' comprehension may depend on whether they have adequate cognitive abilities to coordinate among multiple sources of information. More specifically, individuals with better abilities to flexibly switch attention among various sources of information and to focus on important information while inhibiting unimportant information may benefit more from multimedia learning (Miyake et al., 2000). Indeed, research has reported that readers with high working memory capacity (Schnotz, 2005; Fenesi et al., 2016) and better task switching ability (Baadte et al., 2015) achieved better comprehension in multimedia learning.

Another possible factor is individual differences in attention allocation strategy, which can be revealed through eye tracking (Hyona, 2010; Alemdag & Cagiltay, 2018; Van Gog & Scheiter, 2010). For example, through eye tracking, Kruger and Steyn (2014) reported an association between subtitle reading and comprehension performance when participants were learning from recorded lecture videos. Previous research using eye tracking to understand multimedia processing typically only focused on group level comparisons (e.g., D'Ydewalle & De Bruycker, 2007). However, recent studies have reported significant individual differences in eye movement patterns that can reflect differences in cognitive strategy (e.g., Chan et al., 2018; Chan, Barry, et al., 2020; Chan, Jackson, et al., 2020; Chan et al., 2021; Chan, Suen, et al., 2020; Chuk et al., 2020; Hsiao, Lan, et al., 2021; Zhang et al., 2019). Thus, participants adopting different eye movement patterns during multimedia learning may differ in their comprehension performance.

### 1.2. Objectives

Here we aimed to examine, in multimedia learning of audio lessons consisting of statements of facts and thus the important information for comprehension is in the audio, how verbatim subtitles (i.e., on-screen text that are the same as the audio content) and video showing relevant images influences participants' comprehension of audio lessons. In addition, we aimed to examine whether the comprehension depends on participants' attention allocation during video viewing as reflected in their eye movement behavior after individual differences in cognitive abilities, including working memory capacity, switching ability, and executive planning, were accounted for. Subtitles provide high-level redundant information due to the exact match to the content of the audio lessons semantically and phonologically, whereas video content is relevant to the audio lessons semantically but does not provide verbatim information as subtitles. In Experiment 1, we created four conditions according to whether subtitles were presented and whether video was presented with the audio lessons. In the audio-only condition, participants were presented with audio lessons only; in the audio + text condition, they were presented with audio lessons and the corresponding subtitles; in the audio + video condition, audio lessons were presented with relevant video; in the audio + text + video condition, audio lessons were presented with both corresponding subtitles and relevant video. We then tested (1) the verbal redundancy effect, i.e., whether concurrent presentation of verbatim subtitles with audio lessons facilitated comprehension, by comparing the audio + text with the audio-only condition; (2) the multimedia principle, i.e., whether concurrent presentation of relevant graphics/videos with audio lessons facilitated comprehension, by comparing the audio + video with the audio-only condition; (3) the redundancy principle, i.e., whether adding verbatim subtitles to narrated video (that is, audio lessons with relevant video) facilitated comprehension, by comparing the audio + text + video condition with the audio + video condition. We also examined in the conditions where video was presented (i.e., audio + video and audio + video + text), whether participants' comprehension could be predicted by their attention allocation as reflected in eye movement behavior and cognitive abilities. In Experiment 2, we aimed to replicate the findings related to the redundancy principle and how well eye movement and cognitive ability measures account for individual difference in the comprehension of audio lessons in the audio + video and audio + text + video conditions.

We recruited native speakers of the language used in the audio lessons to control for language experience. The Eye Movement analysis with Hidden Markov Models (EMHMM, Chuk et al., 2014), a novel machine-learning based analysis method, was used to analyze eye movement data because of its ability to provide a quantitative measure of a person's eye movement pattern in a visual task, taking both regions gazed and the *order* of the regions gazed into account. This quantitative measure of eye movement pattern allowed us to examine whether eye movement pattern could account for comprehension performance difference in multimedia learning.

### 1.3. Theory and predictions

According to the verbal redundancy effect reported in the literature (e.g., Adesope & Nesbit, 2012), we hypothesized that subtitles are helpful in the comprehension of audio lessons because of the exact match to the content of the audio lessons. Thus, we predicted that participants would have better comprehension in the audio + text than the audio-only condition. In contrast, when video is used, since the multimedia principle is shown to depend on individual differences in learning ability (e.g., Fletcher & Tobias, 2005), we hypothesized that comprehension of audio lessons with video may require more attention coordination between auditory and visual information, and thus may depend on individual differences in cognitive abilities and attention allocation behavior as reflected in eye movement pattern. Accordingly, we predicted that adding video to the audio lessons (the audio + video condition) would not lead to significantly better comprehension than the audio-only condition in the participants, and individual difference in comprehension performance could be predicted by both cognitive abilities and eye movement pattern. Specifically, people with better cognitive abilities may show better comprehension; also, people with more explorative eye movement patterns during video viewing may be more distracted by unimportant video content, leading to worse comprehension performance. As for the redundancy principle, which predicts poorer comprehension of narrated videos with subtitles than without subtitles due to competition between visual information from

videos and subtitles (Hoffman, 2006; Mayer et al., 2001), since we used audio lesson stimuli, the important information for learning is mainly in the audio instead of the video. Thus, there may be reduced competition between visual information from videos and subtitles. In contrast to the redundancy principle, we speculated that adding subtitles may help maintain the viewers' attention on the subtitles instead of unimportant video content and enhance comprehension through verbal redundancy (e.g., Kruger & Steyn, 2014). Accordingly, we predicted that participants would have better comprehension in the audio + text + video condition than the audio + video condition.

## 2. Experiment 1

### 2.1. Method

#### 2.1.1. Participants

In a meta-analysis of multimedia learning (Adesope & Nesbit, 2012), the average number of participants of the studies included was 60.59 (SD = 41.066). Also, previous studies using the EMHMM method to examine behavioral differences between two participant groups with different eye movement patterns typically recruited 48 to 68 participants (e.g., Chan et al., 2018; Chuk, Chan, & Hsiao, 2017; Chuk, Crookes, et al., 2017). Accordingly, here we recruited 60 native Mandarin speakers (40 females, 18–30 years old, M = 21.07, SD = 3.32) from a local university[1] to learn from audio lessons in Mandarin. Participants were from different majors except for ecology, astronomy, geography and chemistry, which were the topics of the audio lessons used here. All participants reported normal or corrected-to-normal vision.

#### 2.1.2. Materials

The materials consisted of 16 audio lessons in ecology, astronomy, geography, and chemistry, with 4 lessons in each topic. The length of each lesson was 75 s. The audio lessons were accompanied by relevant videos. The lessons had subtitles that were the same as the spoken words in the audio. The resolution of the videos was 1920 x 1280 pixels. All materials were produced by China Central Television (CCTV) and Shanghai Education Television (SETV) and were accessible to the general public, and thus the use of technical terms was minimal. Since our participants were native Mandarin speakers and simplified Chinese readers, to ensure the understandability of the lessons to native Mandarin speakers and to avoid possible linguistic biases, all lessons were in Mandarin with simplified Chinese subtitles, were produced as statements of facts, and were not translated from foreign languages. Video images were relevant to the audio lessons but not required for the understanding of the audio lessons (Fig. 1A. See Appendix 1 for more examples). The materials were unfamiliar to the participants.

Lessons were presented in 4 different conditions: a) *Audio-only* condition (i.e., audio lessons alone without video or subtitles, with a static icon presented at the centre of a black screen); b) *Audio + text* condition (i.e., audio lessons played with subtitles that were the same as the spoken words); c) *Audio + video* condition (i.e., audio lessons played with full-screen video, with a fixed title masking the subtitles in the video clips); and d) *Audio + text + video* condition (i.e., audio lessons played with both video and subtitles; Fig. 1). Each lesson was shown to each participant only once in one of the four presentation conditions. Among the 16 original video clips, half of them had subtitles located at the bottom left of the screen, while the other half set the subtitles at the bottom centre.

The materials also included audio files of 96 multiple-choice questions (MCQs) as comprehension questions to test participants' memory of the content of the lessons, with 6 MCQs for each lesson. The comprehension questions were based on the audio lesson content but not video content (Fig. 1A; see Appendix 1 and 2 for more examples). Thus, although video content was relevant to the audio lesson content, it was not required for the understanding of the audio lessons. The voice of the audio files was synthesized by the online Baidu voice producer.

#### 2.1.3. Design

Here we examined how adding video and subtitles affected comprehension of audio lessons. The independent variables were video (with vs. without) and subtitles (with vs. without), resulting in four experimental conditions: audio-only, audio + text, audio + video, and audio + text + video. The subtitles matched the spoken words verbatim. A within-subject design was used to minimize the effect of inter-participant variability. Each participant was presented with 16 lessons in total with four lessons in each condition (one from each topic). The lessons used in the four conditions were counterbalanced across participants. More specifically, participants were randomly and equally assigned to one of the four presentation condition groups shown in Table 1. For each participant, the presentation order of the four lessons in each block in Table 1 and the presentation order of the four blocks were both randomized. The dependent variable was comprehension performance measured as the accuracy in answering the comprehension questions related to the lessons. Repeated measures ANOVA was used for the data analysis.

In a separate analysis, we examined whether eye movement pattern used in the audio + video and audio + text + video conditions could predict comprehension performance. Participants' eye movement pattern was quantitatively assessed using the EMHMM approach (Chuk et al., 2014). Please see the Eye Movement Data Analysis section below for details.

We also examined whether participants' cognitive abilities could predict comprehension performance. Participants completed a two-back task for testing working memory capacity (Lau et al., 2010), a Tower of London task for assessing executive function and planning ability (Phillips et al., 2001), and a multitasking task for testing task switching ability (Stoet et al., 2013). We then performed a hierarchical analysis to examine whether eye movement pattern could still predict comprehension after variation due to cognitive abilities was controlled.

#### 2.1.4. Procedure

**Comprehension of Audio Lessons.** During the entire experiment, participants' eye movements were recorded by an eye tracker, Eyelink 1000 with desktop mount (SR Research). For eye movement data, only those during the audio + video and audio + text + video conditions were analysed. The tracking mode was pupil and corneal reflection with a sampling rate of 1000 Hz. Stimuli were displayed on a 22" CRT monitor with a resolution of 1920 by 1440 pixels and 150 Hz frame rate. The viewing distance was 60 cm. A chinrest was used to reduce head movement. The standard nine-point calibration procedure was carried out before the experiment and whenever the drift correction error was larger than 1° of visual angle. Each trial started with a white solid dot appearing at the center of the screen. Participants were asked to look at the dot whenever it appeared for drift correction. Afterwards, an audio lesson was played in one of the four presentation conditions as shown in Fig. 1B. Videos were presented in full screen. After each lesson, participants were asked to answer 6 aurally-presented MCQs presented one at a time binaurally. Participants could replay each question unlimited times before their response. Participants performed 4 blocks of the task, with four lessons in each block presented in the four different presentation conditions respectively. The lesson order in each block and the block order were both randomised. They proceeded to the cognitive ability tests described below after the comprehension task.

Cognitive Ability Tests.

[1] According to a power analysis, for within-factor repeated measures ANOVA with 4 measurements, assuming a medium effect size f = 0.25, α = 0.05, and power = 0.8, the required sample size is 24. For linear multiple regression with four independent predictors, assuming each with 0.25 correlation with outcome (effect size $f^2 = 0.33$), α = 0.05, and power = 0.80, the required sample size is 41.

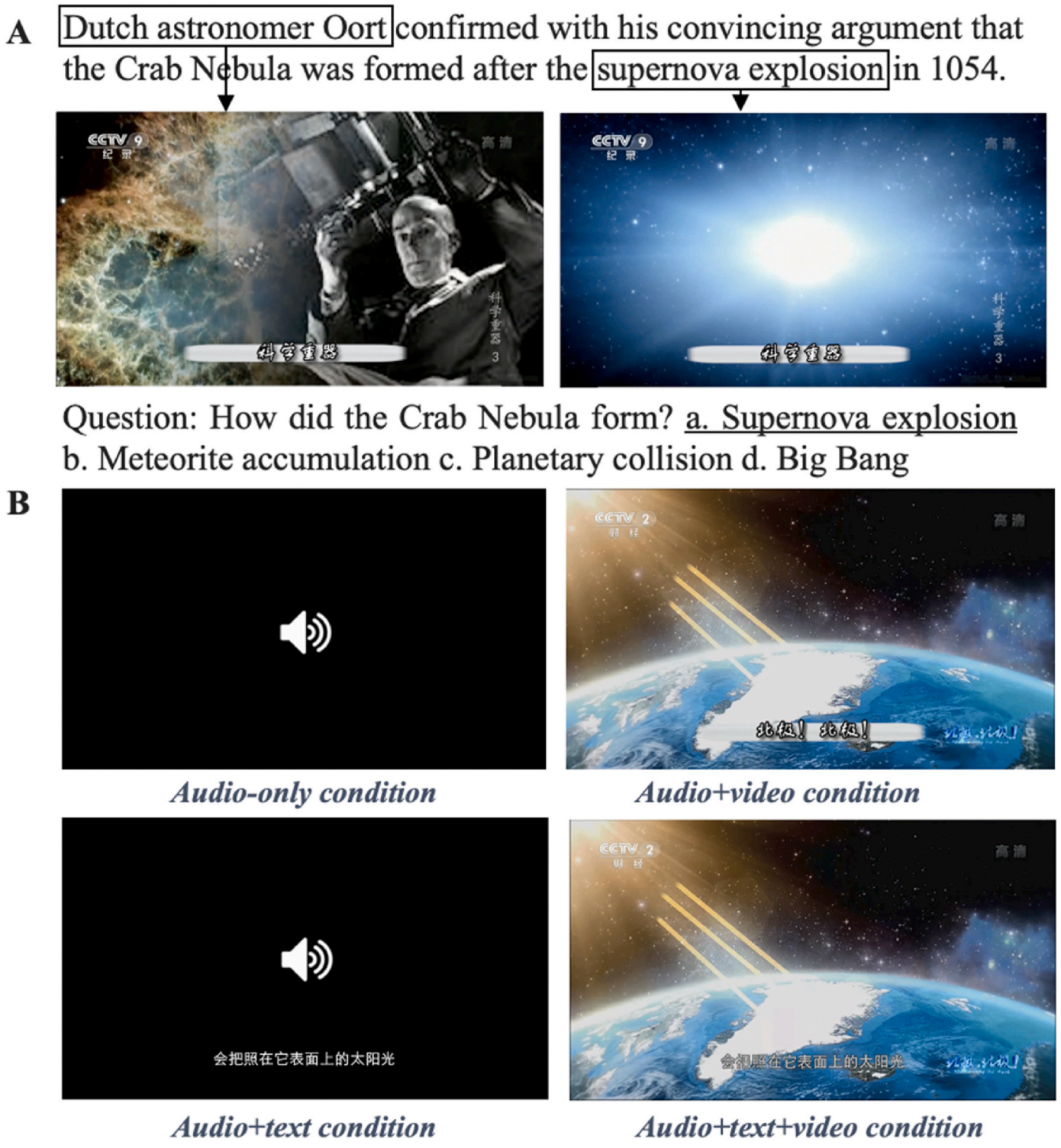



**Fig. 1.** (A) Example lesson scripts translated in English, video images, and a corresponding comprehension question. (B) Four presentation conditions of the lessons.

1. **Two-back Test**: Three types of stimuli, including visual English letters, spoken numbers, and irregular shapes (Fig. 2A) were used to test visual and verbal working memory in three separate blocks. The visual English letters were capital letters presented at the center of the screen (Nystrom et al., 2000). The spoken numbers contained single-digit numbers (0–9) presented binaurally (Crottaz-Herbette et al., 2004). The irregular shapes consisted of computer-generated abstract shapes (Attneave and Arnoult structures; Fig. 2A) presented at the center of the screen (Attneave & Arnoult, 1956; Nystrom et al., 2000). For each type of stimuli, participants were presented with 30 items one at a time, each for 2.5 s with a 0.5 s interval (Lau et al., 2010), and asked to judge whether the item presented in a trial matched the one that appeared 2 trials back. The accuracy and response time (RT) were measured.
2. **Tower of London Test**: Participants were asked to move 3 color discs one at a time from an initial position to match a goal position with the minimum number of moves, and to plan the moves in mind before execution (Fig. 2B; Phillips et al., 2001). Participants

**Table 1**
Presentation condition groups adopted in the current study to counterbalance the lessons used in the four different experimental conditions (audio-only, audio + text, audio + video, and audio + text + video).

| Block no. | Lessons | Presentation condition groups | | | |
|---|---|---|---|---|---|
| | | Group 1 | Group 2 | Group 3 | Group 4 |
| **Block 1** | Ecology 1 – Bamboo worms | Audio-only | Audio + text + video | Audio + text | Audio + video |
| | Astronomy 1 – LAMOST | Audio + text + video | Audio + text | Audio + video | Audio-only |
| | Geography 1 – Arctic adventure | Audio + text | Audio + video | Audio-only | Audio + text + video |
| | Chemistry 1 – Phosphorus | Audio + video | Audio-only | Audio + text + video | Audio + text |
| **Block 2** | Ecology 2 – Termites | Audio + text + video | Audio + text | Audio + video | Audio-only |
| | Astronomy 2 – Spectrum | Audio + text | Audio + video | Audio-only | Audio + text + video |
| | Geography 2 – Glacier melting | Audio + video | Audio-only | Audio + text + video | Audio + text |
| | Chemistry 2 – Oxygen | Audio-only | Audio + text + video | Audio + text | Audio + video |
| **Block 3** | Ecology 3 – Lizard | Audio + text | Audio + video | Audio-only | Audio + text + video |
| | Astronomy 3 – Mebula | Audio + video | Audio-only | Audio + text + video | Audio + text |
| | Geography 3 – Icebreaker | Audio-only | Audio + text + video | Audio + text | Audio + video |
| | Chemistry 3 – Fabric | Audio + text + video | Audio + text | Audio + video | Audio-only |
| **Block 4** | Ecology 4 – Leech | Audio + video | Audio-only | Audio + text + video | Audio + text |
| | Astronomy 4 – Telescope | Audio-only | Audio + text + video | Audio + text | Audio + video |
| | Geography 4 – Fishing | Audio + text + video | Audio + text | Audio + video | Audio-only |
| | Chemistry 4 – Liquid crystal | Audio + text | Audio + video | Audio-only | Audio + text + video |

completed 10 trials. The accuracy, total number of moves, total execution time, and total preplanning time before executing the first move were measured. Five practice trials were provided.

3. **Multitasking Test**: Four types of figures with different combinations of shapes and fillings were used as the stimuli (Fig. 2C). The stimuli were presented one at a time in either the top or the bottom half of a box at the center of the screen (Fig. 2C, left). Participants were asked to perform a dual task where they judged the shape of the figure (the shape task) as fast and correctly as possible when the figure was presented in the top half of the box, and judged the number of dots in the filling of the figure (the filling task) when the figure was presented in the bottom half of the box (Stoet et al., 2013). The figure was presented for 2500 ms, followed by a 500 ms blank screen. Participants were asked to respond by the end of the 3-s trial. A shape-only and a filling-only task were tested sequentially before the dual-task to measure participants' baseline behavior where no task switching was involved. Their task switching ability was measured as the RT in the dual task minus the average RT during the two no-switching tasks.

#### 2.1.5. Eye Movement Data Analysis

We used the EMHMM method (Chuk et al., 2014) to obtain a quantitative measure of a participant's eye movement pattern in the audio + video and the audio + text + video conditions separately. The EMHMM method is based on the assumption that in a visual task, the current eye fixation location is conditioned upon the previous fixation location. Thus, eye movement behavior in a visual task may be understood as a Markovian stochastic process, which can be better captured using an hidden Markov model (HMM, a type of time-series statistical model in machine learning). In the EMHMM method, each participant's eye movement pattern was summarized using an HMM in terms of personalized regions of interest (ROIs) and transition probabilities among the ROIs, with the hidden states of the HMM corresponded to the ROIs (Fig. 3A). The parameters of the HMM were estimated directly from the participant's data using the Variational Bayesian Expectation Maximization (VBEM) algorithm (Bishop, 2006), with the number of ROI automatically determined from a pre-set range through the variational Bayesian approach. All participants' HMMs then could be clustered according to their similarities to discover representative eye movement patterns/HMMs among participants using the variational hierarchical expectation maximization (VHEM) algorithm (Coviello et al., 2014, Fig. 3B). Each participant's eye movement pattern then could be quantitatively assessed as the likelihood of the participant's eye movement data being generated by a representative HMM: the higher the likelihood, the higher the similarity to the representative HMM (please refer to Chuk et al., 2014, for more details).

In the current study, for each participant, we trained one HMM to summarize the participant's eye movement pattern in the audio + video condition, and another HMM for the audio + text + video condition. For each HMM, when using the variational Bayesian method to determine the optimal number of ROIs, we ran each HMM with a different number of ROIs (ranging from 1 to 6) 300 times with a random initialization each time and selected the model with the largest log-likelihood given the data. For the audio + video and the audio + text + video condition separately, individual HMMs were clustered into two groups, which were referred to as Group 1 and Group 2. The representative HMMs of the discovered clusters were generated with the number of ROIs set to the median number of ROIs of the individual HMMs (following previous studies, e.g., Chan et al., 2018. Note that individual HMMs may have different numbers of ROIs since the optimal number of ROIs was determined automatically using the variational Bayesian approach), and were referred to as Group 1 HMM and Group 2 HMM. The clustering algorithm was run for 300 times based on different initializations, and the group HMMs with the highest data log-likelihood were used for the analysis. A participant's eye movement pattern was then assessed as the log-likelihood of the participant's eye movement data being generated by Group 1 HMM and Group 2 HMM separately. These two log-likelihood measures reflected the similarity of the participant's eye movement data to the Group 1 and Group 2 representative pattern respectively. Following previous studies (e.g., An & Hsiao, 2021; Chan et al., 2018; Hsiao, An, et al., 2021), we quantified a participant's eye movement pattern along the dimension of the contrast between Group 1 and Group 2 patterns by defining Group 1–2 scale as:

$$Group\ 1-2\ scale = \frac{Group\ 1\ \log-likelihood - Group\ 2\ \log-likelihood}{|Group\ 1\ \log-likelihood| + |Group\ 2\ \log-likelihood|}$$

Where Group 1 log-likelihood is the log-likelihood of the participant's eye movement data being generated by Group 1 HMM, and Group 2 log-likelihood is the log-likelihood of the participant's data being generated by Group 2 HMM. A more positive value in Group 1–2 scale indicated higher similarity to Group 1 pattern, whereas a more negative value indicated higher similarity to Group 2 pattern. This Group 1–2

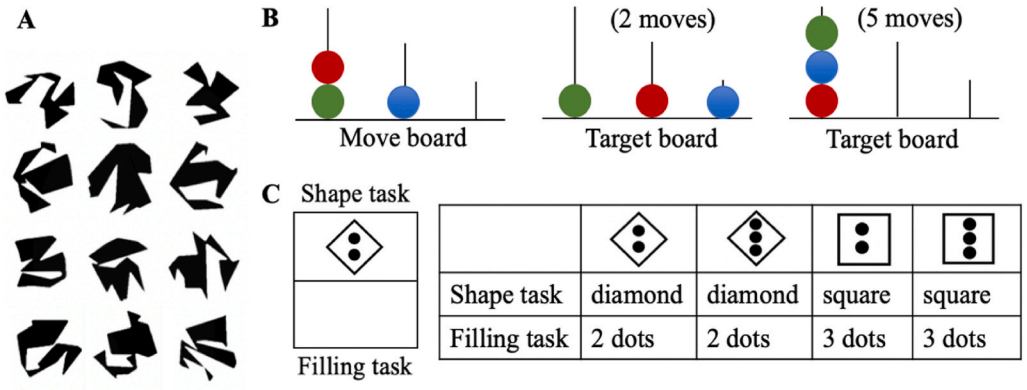
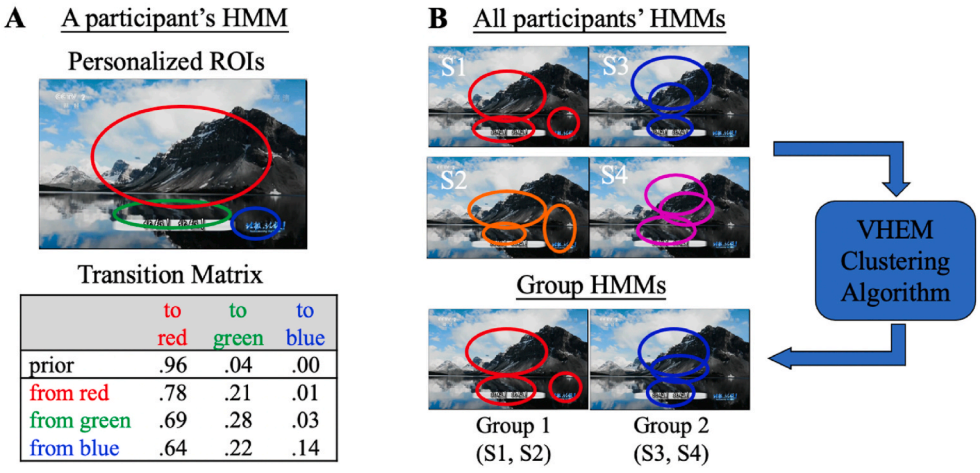



**Fig. 2.** (A) 12 pictorial stimuli (Attneave and Arnoult structures) used in the two-back test, (B) Example of the Tower of London test, and (C) Stimuli used in the multitasking test.




| | to red | to green | to blue |
|---|---|---|---|
| prior | .96 | .04 | .00 |
| from red | .78 | .21 | .01 |
| from green | .69 | .28 | .03 |
| from blue | .64 | .22 | .14 |

**Fig. 3.** (A) Example of an HMM summarizing a participant's eye movement pattern during video viewing. Ellipses show ROIs as 2-D Gaussian distributions. The table shows transition probabilities among the ROIs. Priors show the probabilities that a fixation sequence starts from the ellipse. In this example, the participant has 96% probability to start viewing the video with a fixation in the red region, and 4% with a fixation in the green region. After a fixation in the red region, the next fixation has 78% probability to stay in the red region, 21% to switch to the green region, and 1% to switch to the blue region. (B) Illustration of the VHEM clustering algorithm: S1 and S2 have similar eye movement patterns, whereas S3 and S4 have similar patterns. The clustering algorithm groups S1 and S2 to form Group 1, and S3 and S4 to form Group 2. Group 1 and Group 2 HMMs then can be used to quantify a participant's eye movement pattern as the participant's data likelihood given Group 1 HMM or Group 2 HMM.

scale was then used as the measure of participants' eye movement pattern to examine the associations between eye movement pattern, comprehension performance, and cognitive abilities.

**Table 2**
Means and standard deviations of comprehension accuracy in the four conditions in Experiment 1.

| Metrics | Conditions | Mean | SD | 95% confidence interval for mean | |
|---|---|---|---|---|---|
| | | | | Lower bound | Upper bound |
| Comprehension Accuracy (%) | Audio-only | 75.21 | 14.32 | 71.51 | 78.91 |
| | Audio + text | 80.97 | 9.84 | 78.43 | 83.51 |
| | Audio + video | 76.34 | 15.81 | 72.26 | 80.43 |
| | Audio + text + video | 79.44 | 12.22 | 76.29 | 82.60 |

## 2.2. Results

### 2.2.1. Effect of video and subtitles on comprehension performance

The means and standard deviations of comprehension performance in accuracy in different conditions were summarized in Table 2. The results showed a significant main effect of subtitles, $F(1, 59) = 13.359$, $p = .001$, $\eta^2_p = .185$. There was no main effect of video or interaction between video and subtitles. This result suggested that subtitles, but not video, facilitated comprehension of audio lessons. In the planned pairwise comparisons, a significant difference between the audio-only and the audio + text conditions was observed, $t(59) = -3.423$, $p = .001$, $d = -0.442$, consistent with the verbal redundancy effect observed in the literature. There was no significant difference between the audio-only and the audio + video conditions, $t(59) = -0.634$, $p = .528$, $d = -0.082$, suggesting that adding video did not enhance comprehension. This result was in contrast to the multimedia principle, which suggests facilitation effects when combining word and picture information during learning. When we compared the audio + video and the audio + text + video conditions, a marginal effect was observed, $t(59) = -1.809$, $p = .075$, $d = -0.234$, suggesting that adding subtitles to video clips with audio narratives might be beneficial. This result was in contrast to the redundancy principle, which posits that learners perform better when

[2] According to a power analysis, the sample size required for paired *t*-test, assuming small-to-medium effect size $d = 0.25$, $\alpha = 0.05$, and power = 0.8, is 101. Among the 103 participants recruited, one male participant did not complete the task due to technical problem. One male and one female participants' eye movement data were excluded from analysis due to poor calibration.

presented with a narrated animation without redundant visual text information (such as subtitles) than with the information.

### 2.2.2. Eye movement patterns for viewing videos

For the audio + video and audio + text + video conditions separately, we used the EMHMM method to summarize each participant's eye movement pattern using an HMM, and then clustered all individual HMMs into two groups to discover two representative patterns among the participants. The representative HMMs of the discovered clusters were generated with the number of ROIs set to 4, the median number of ROIs of the individual HMMs (See the Eye Movement Data Analysis section in the Method for details).

Fig. 4 shows the results of the audio + video condition. In Group 1 pattern, after an initial fixation at the center of the video, participants had 8% probability to look at either the blue ROI containing a logo on the bottom right or the green and the pink ROIs containing the fixed title at the bottom center of the screen. Afterwards, they tended to stay in the same ROI or switch back to the red ROI, or occasionally switched among the green, blue, and pink ROIs. We referred to this pattern as the distributed pattern. 46 participants were classified in this group. In contrast, in Group 2 pattern, there were overlapping ROIs around the screen center, suggesting that participants mainly focused at the screen center. We referred to this pattern as the centralized pattern. There were 14 participants in this group. The representative HMMs of the two patterns significantly differed from each other according to Kullback-Leibler (KL) divergence approximation using data log-likelihoods (Chuk et al., 2014): data from participants adopting the distributed pattern were significantly more likely to be generated by the distributed representative HMM than the centralized one, $t(45) = 7.144$, $p < .01$, $d = 1.053$, and data from those adopting the centralized pattern were more likely to be generated by the centralized representative HMM than the distributed one, $t(13) = 1.892$, $p = .04$, $d = 0.506$.

We then quantified participants' eye movement pattern in the audio + video condition using Group 1–2 scale as described in the Eye Movement Data Analysis section. Here we referred to the Group 1–2 scale as the Distributed-Centralized scale (D-C scale) to increase clarity. Specifically, the D-C scale was defined as:

$$Distributed - Centralized\ scale = \frac{D - C}{|D| + |C|}$$

Where D is the log-likelihood of the participant's eye movement data being generated by the representative HMM of the distributed pattern, and C is the log-likelihood of the participant's data being generated by the representative HMM of the centralized pattern. This log-likelihood measure reflects the similarity of the participant's eye movement pattern to the representative patterns resulting from clustering. A more positive value in D-C scale indicated higher similarity to the distributed pattern, whereas a more negative value indicated higher similarity to the centralized pattern (Table 3). We found that participants' eye movement pattern as measured in D-C scale was negatively correlated with comprehension accuracy in the audio + video condition, $r(58) = -.291$, $p = .024$: the more distributed the pattern, the lower the accuracy in comprehension (Fig. 5A).

A similar analysis was conducted with eye movement data in the audio + text + video condition. Fig. 6 shows the results of clustering participants' eye movement patterns into 2 groups. The 2 groups showed similar concentrations on the ROIs at the bottom left and bottom center of the screen, where the subtitles were located, in addition to the screen center. Group 1 pattern showed a higher probability to look at the subtitles regions after looking at the screen center. One-third of the participants (20 out of 60) adopted Group 1 pattern (one participant's eye movement data was invalid due to technical problems). Group 1 HMM and Group 2 HMM were significantly different from each other as indicated by KL divergence approximation: data from participants in Group 1 were significantly more likely to be generated by Group 1 HMM than Group 2 HMM, $t(19) = 4.473$, $p < .01$, $d = 1.001$, and data from those adopting Group 2 pattern were more likely to be generated by Group 2 HMM than Group 1 HMM, $t(39) = 10.336$, $p < .01$, $d = 1.634$. We also quantified participants' eye movement pattern using Group 1–2 scale (Table 3) and found that it did not correlate significantly with comprehension performance (Fig. 5B).

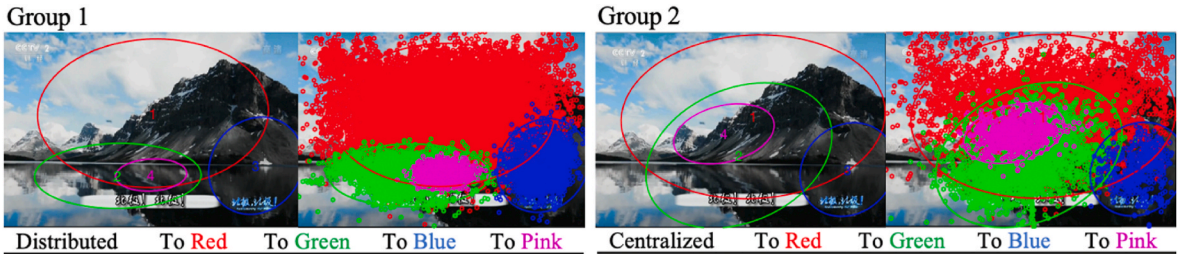
### 2.2.3. Do cognitive abilities and eye movement pattern predict comprehension performance?

In either the audio-only or the audio + text condition, no significant correlation between cognitive abilities and comprehension performance was found.

In the audio + video condition, in addition to the significant correlation between comprehension performance and eye movement pattern

**Table 3**
Means and standard deviations of participants' eye movement pattern as measured in D-C scale/Group 1–2 scale in the audio + video condition (D-C scale) and the audio + text + video condition (Group 1–2 scale) in Experiment 1.

| Metrics | Conditions | Mean | SD | 95% confidence interval for mean | |
|---|---|---|---|---|---|
| | | | | Lower bound | Upper bound |
| D-C scale/Group 1–2 scale | Audio + video | .004 | .0058 | .0025 | .0055 |
| | Audio + text + video | .0002 | .0137 | -.0033 | .0037 |




| Distributed | To Red | To Green | To Blue | To Pink |
|---|---|---|---|---|
| Priors | 1.0 | .00 | .00 | .00 |
| From Red | .92 | .05 | .02 | .01 |
| From Green | .37 | .56 | .07 | .00 |
| From Blue | .38 | .12 | .47 | .04 |
| From Pink | .27 | .00 | .08 | .64 |

| Centralized | To Red | To Green | To Blue | To Pink |
|---|---|---|---|---|
| Priors | .46 | .44 | .01 | .00 |
| From Red | .91 | .08 | .01 | .00 |
| From Green | .18 | .78 | .04 | .01 |
| From Blue | .17 | .39 | .44 | .00 |
| From Pink | .03 | .04 | .00 | .93 |

**Fig. 4.** Group 1 Distributed (left) and Group 2 centralized (right) patterns in the audio + video condition in Experiment 1. Ellipses show ROIs as 2-D Gaussian emissions; the size of the ellipses shows 2 times standard deviation. The table shows transition probabilities among the ROIs. Priors show the probabilities that a fixation sequence starts from the ellipse. The image on the right shows raw eye fixation data and their ROI assignments. Each fixation was assigned to the ROI with the highest data log-likelihood.

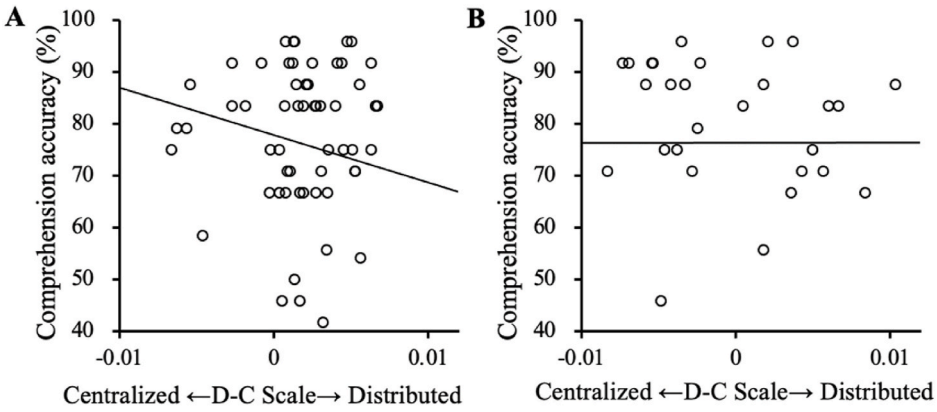
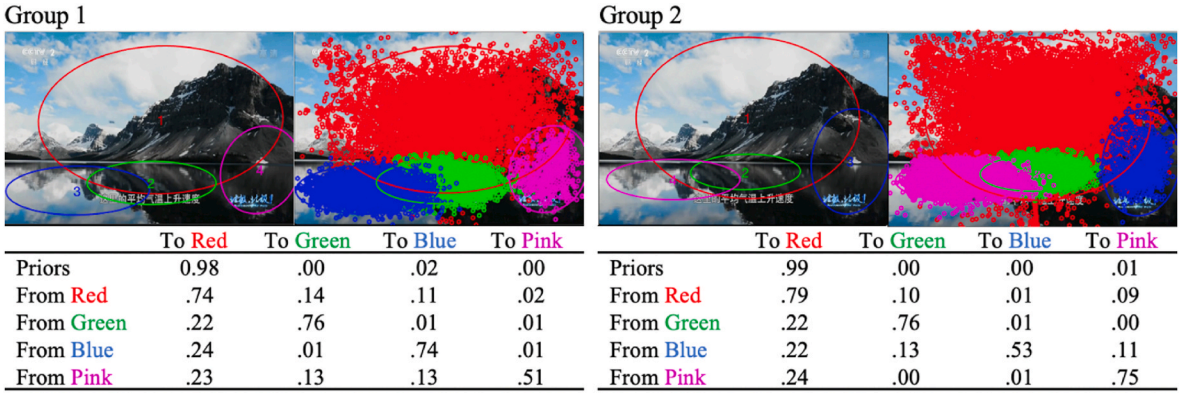



**Fig. 5.** Correlation between eye-movement pattern and comprehension performance in the (A) audio + video condition, and (B) audio + text + video condition in Experiment 1.




| | To Red | To Green | To Blue | To Pink |
|---|---|---|---|---|
| Priors | 0.98 | .00 | .02 | .00 |
| From Red | .74 | .14 | .11 | .02 |
| From Green | .22 | .76 | .01 | .01 |
| From Blue | .24 | .01 | .74 | .01 |
| From Pink | .23 | .13 | .13 | .51 |

| | To Red | To Green | To Blue | To Pink |
|---|---|---|---|---|
| Priors | .99 | .00 | .00 | .01 |
| From Red | .79 | .10 | .01 | .09 |
| From Green | .22 | .76 | .01 | .00 |
| From Blue | .22 | .13 | .53 | .11 |
| From Pink | .24 | .00 | .01 | .75 |

**Fig. 6.** The two representative eye movement patterns observed in the audio + text + video condition in Experiment 1.

(D-C scale), comprehension performance was also significantly correlated with auditory working memory ability as measured in the two-back task, $r(58) = 0.337$, $p = .008$, and task switching ability as measured in the multitasking test, $r(58) = 0.262$, $p = .043$. To examine whether eye movement pattern significantly contributed to comprehension after variation due to cognitive abilities was controlled, a three-stage hierarchical multiple regression was conducted to predict comprehension accuracy. At stage one, auditory working memory capacity (two-back test) contributed significantly to the regression model, $B = 0.283$, $F(1,58) = 7.432$, $p = .008$, and accounted for 11.4% of the variation. Adding task switching ability (multitasking test) to the regression model explained an additional 5.6% of the variation in comprehension accuracy and the change in $R^2$ was marginal, $B = 0.268$, $F(1,57) = 3.832$, $p = .055$. Finally, introducing eye movement pattern (D-C scale) explained an additional 7.9% of the variation in comprehension accuracy and this change in $R^2$ was significant, $B = -.285$, $F(1,56) = 5.891$, $p = .018$. The tests for multicollinearity indicated a low level of multicollinearity (tolerance = 0.978, 0.982, 0.975 for auditory working memory, multitasking, and eye movement patterns, respectively). Thus, this result suggested that when learning from audio lessons with video, participants' attention allocation reflected in eye movement behavior played an important role in comprehension performance in addition to cognitive abilities.

A similar analysis was conducted for predicting comprehension performance in the audio + text + video condition. In contrast to the audio + video condition, neither cognitive abilities nor eye movement pattern could significantly predict the comprehension accuracy. This result suggested that subtitles may facilitate learning by guiding attention allocation and thus reducing the reliance on cognitive abilities and eye movement pattern. To test this possibility, we examined the entropy of each participant's HMM as a measure of how consistent the participant's eye movement patterns were across video clips (Hsiao, Chan, et al., 2021). Entropy is a measure of inconsistency: the higher the entropy, the lower the consistency (Chan & Hsiao, 2016; Cover & Thomas, 2006; Hsiao & Cheung, 2016; Hsiao & Lam, 2013). The entropy in the audio + video condition, $M = 13.370$, $SD = 0.281$, was significantly higher than that in the audio + text + video condition, $M = 12.972$, $SD = 0.338$, $t(59) = 9.836$, $p < .001$, suggesting that eye movements when viewing videos with subtitles were more consistent across video clips than when viewing videos without subtitles.

## 3. Experiment 2

In Experiment 1, in contrast to the redundancy principle (Mayer et al., 2001), we observed a marginal facilitation effect of subtitles on the comprehension of audio lessons. In addition, comprehension performance of audio lessons with video but without subtitles depends on participants' cognitive abilities and eye movement pattern. In Experiment 2, we aimed to replicate these two findings by focusing on the audio + video and audio + text + video conditions.

### 3.1. Method

103 native Mandarin speakers (88 females, 18–30 years old, M = 21.07, SD = 3.32) from a local university were recruited.[2] The recruitment criteria were identical to Experiment 1. The materials were 4 lessons from Experiment 1 with subtitles presented at the bottom center to control for the subtitles position across video clips. The corresponding 24 MCQs (6 for each clip) from Experiment 1 were also used to test participants' comprehension accuracy. The video clips were presented

in either the audio + video condition (i.e., without subtitles) or the audio + text + video condition (i.e., with subtitles). The design consisted of a within-subject variable, subtitles (with vs. without). Each participant viewed each clip once, with 2 clips in each of the 2 subtitle conditions. Participants were randomly and equally assigned to one of the two presentation condition groups shown in Table 4. For each participant, the presentation order of the two lessons in each block in Table 4 and the presentation order of the two blocks were both randomized. The dependent variable was comprehension accuracy. Paired sample *t*-test was used for data analysis. Participants also performed the same cognitive ability tests as in Experiment 1. Similar to Experiment 1, in the two subtitle conditions separately, we used the EMHMM method to quantify participants' eye movement pattern, and then performed multiple regression to examine whether eye movement pattern and cognitive abilities predicted comprehension accuracy. The same apparatus and procedure as Experiment 1 were used.

### 3.2. Results

#### 3.2.1. Effect of subtitles on comprehension of audio lessons

The results showed a significant difference in comprehension accuracy between the audio + video and audio + text + video conditions, t (101) = 4.72, p < .001, d = 0.467 (Table 5). This result indicated that adding subtitles to audio lessons with video facilitated comprehension, in contrast to the redundancy principle (Mayer et al., 2001).

#### 3.2.2. Eye movement patterns for viewing videos

The EMHMM method was used to analyze the eye movement data in the audio + video and audio + text + video conditions separately. Fig. 7 shows the two representative eye movement patterns discovered in the audio + video condition. In Group 1 pattern, participants typically started from a fixation at the screen center (red ROI), and then either stayed around the center (red ROI, 91% probability) or switched to other regions with icons at the bottom center (green ROI), bottom right (blue ROI), and upper left (pink ROI) of the screen (total 9%). We referred to this pattern as the distributed pattern. 59 (out of 100) participants were classified in this group. In Group 2 pattern, participant typically started from and stayed at the screen center (red and green ROIs), with a slightly lower probability (7%) to look at other regions (also fewer regions: blue and pink ROIs) than Group 1 pattern. We referred to this pattern as the centralized pattern. 41 participants were clustered into this group. The representative HMMs of the two patterns differed from each other according to KL divergence approximation using data log-likelihoods (Chuk et al., 2014): data from participants adopting the distributed pattern were marginally more likely to be generated by the distributed HMM than the centralized HMM, $t(58) = 1.31$, $p = .09$, $d = 0.17$, and data from those adopting the centralized pattern were significantly more likely to be generated by the centralized HMM than the distributed HMM, $t(40) = 10.59$, $p < .001$, $d = 1.65$.

We used the Distributed-Centralized scale (D-C scale) to quantify each participant's eye movement pattern as in Experiment 1. A more positive D-C scale indicated higher similarity to the distributed pattern.

**Table 4**
Presentation condition groups to counterbalance the lessons used in the two different experimental conditions (audio + video, and audio + text + video).

| Block no. | Lessons | Presentation condition groups | |
|---|---|---|---|
| | | Group 1 | Group 2 |
| Block 1 | Geography 3 – Icebreaker | Audio + video | Audio + text + video |
| | Astronomy 3 – Mebula | Audio + text + video | Audio + video |
| Block 2 | Geography 2 – Glacier melting | Audio + text + video | Audio + video |
| | Astronomy 4 – Telescope | Audio + video | Audio + text + video |

**Table 5**
Means and standard deviations of comprehension accuracy and eye movement pattern as measured in D-C scale in the audio + video and audio + text + video conditions in Experiment 2.

| Metrics | Conditions | Mean | SD | 95% confidence interval for mean | |
|---|---|---|---|---|---|
| | | | | Lower bound | Upper bound |
| Comprehension Score (%) | Audio + video | 75.00 | 10.68 | 72.87 | 77.13 |
| | Audio + text + video | 80.56 | 9.98 | 78.57 | 82.55 |
| Eye movement pattern (D-C scale) | Audio + video | -.0022 | .0046 | -.0031 | -.0012 |
| | Audio + text + video | .000018 | .0125 | -.0025 | .0025 |

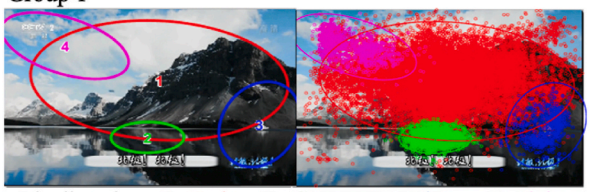
Consistent with Experiment 1, participants' eye movement pattern as measured in D-C scale was negatively correlated with comprehension accuracy in the audio + video condition, $r(99) = -.230$, $p = .021$: the more distributed the pattern, the lower the comprehension performance (Fig. 8A).

In the audio + text + video condition, the two representative eye movement patterns discovered through clustering were shown in Fig. 9. In Group 1 pattern, in addition to the screen center (red ROI) and the subtitle region (green and blue ROIs), participants also occasionally looked at the irrelevant icon at the bottom right corner of the screen (pink ROI). In contrast, in Group 2 pattern, participants mainly focused on the screen center (red and blue ROIs) and the subtitles (pink and green ROIs). Accordingly, we referred to Group 1 pattern as the distributed pattern, and Group 2 pattern as the centralized pattern. 51 (out of 100) participants were classified as adopting the distributed pattern, and 49 participants adopted the centralized pattern. The representative HMMs of the two patterns differed from each other according to KL divergence approximation using data log-likelihoods (Chuk et al., 2014): data from participants adopting the distributed pattern were more likely to be generated by the distributed HMM than the centralized HMM, $t(50) = 12.58$, $p < .001$, $d = 1.76$, and data from those adopting the centralized pattern were more likely to be generated by the centralized HMM than the distributed HMM, $t(48) = 6.99$, $p < .001$, $d = 1.00$. Interesting, similar to the audio + video condition, we found that participants' eye movement pattern as measured in D-C scale was negatively correlated with comprehension accuracy in the audio + text + video condition, $r(99) = -.289$, $p = .004$: the more distributed the pattern, the lower the comprehension performance (Fig. 8B).

#### 3.2.3. Do cognitive abilities/eye movement pattern predict comprehension performance?

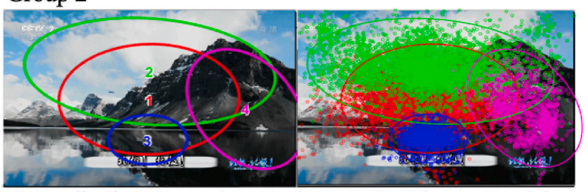
In the audio + video condition, in addition to eye movement pattern as measured in D-C scale, comprehension accuracy was significantly correlated with Tower of London executing time, $r(101) = -.276$, $p = .005$. To examine whether eye movement pattern significantly contributed to comprehension accuracy after variance due to cognitive abilities was controlled, a two-stage hierarchical multiple regression was conducted to predict comprehension accuracy. At stage one, Tower of London execution time contributed significantly to the regression model, $B = -.269$, $F(1,98) = 7.639$, $p = .007$, and accounted for 7.2% of the variance. Introducing the eye movement pattern measure (D-C scale) explained an additional 5.9% of the variation in comprehension accuracy and the change in $R^2$ was significant, $B = -.244$, $F(1,97) = 6.618$, $p = .015$. The tests for multicollinearity indicated a low level of multicollinearity (tolerance = 0.998 and 0.998 for executing functioning and eye movement pattern respectively). Thus, this result suggested that participants' attention allocation during video viewing as reflected in eye movement behavior contributed to comprehension accuracy after variance in cognitive abilities was taken into account.

Group 1



| Distributed | To Red | To Green | To Blue | To Pink |
|---|---|---|---|---|
| Priors | 1.0 | .00 | .00 | .00 |
| From Red | .91 | .07 | .01 | .01 |
| From Green | .53 | .42 | .05 | .01 |
| From Blue | .37 | .17 | .44 | .01 |
| From Pink | .35 | .07 | .01 | .56 |

Group 2



| Centralized | To Red | To Green | To Blue | To Pink |
|---|---|---|---|---|
| Priors | .93 | .07 | .00 | .00 |
| From Red | .86 | .06 | .04 | .03 |
| From Green | .09 | .84 | .04 | .03 |
| From Blue | .32 | .11 | .47 | .10 |
| From Pink | .19 | .13 | .09 | .59 |

**Fig. 7.** Distributed (left) and centralized (right) eye movement patterns discovered in the audio + video condition in Experiment 2.

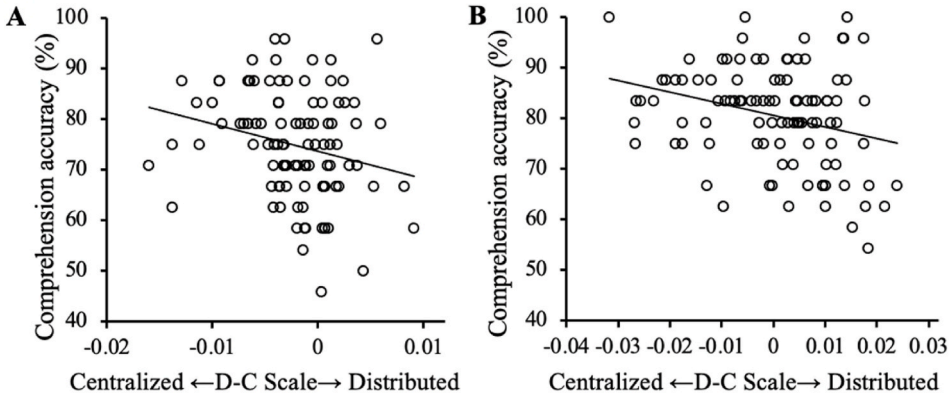
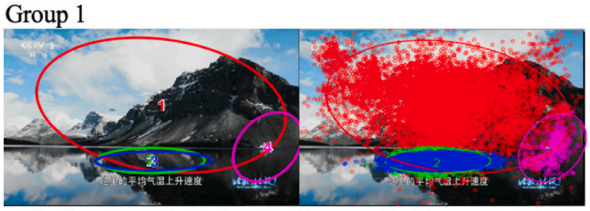


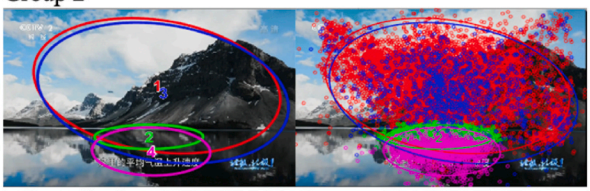

**Fig. 8.** Correlation between eye-movement pattern as measured in D-C scale and comprehension accuracy in the (A) audio + video condition, and (B) audio + text + video condition in Experiment 2.

Group 1



| Distributed | To Red | To Green | To Blue | To Pink |
|---|---|---|---|---|
| Priors | .91 | .05 | .04 | .00 |
| From Red | .73 | .25 | .01 | .01 |
| From Green | .23 | .27 | .49 | .00 |
| From Blue | .25 | .07 | .68 | .00 |
| From Pink | .27 | .21 | .02 | .50 |

Group 2



| Centralized | To Red | To Green | To Blue | To Pink |
|---|---|---|---|---|
| Priors | .84 | .02 | .14 | .00 |
| From Red | .75 | .25 | .00 | .00 |
| From Green | .29 | .71 | .00 | .00 |
| From Blue | .00 | .00 | .69 | .31 |
| From Pink | .00 | .00 | .27 | .73 |

**Fig. 9.** Distributed (left) and centralized (right) eye movement patterns discovered through clustering in the audio + text + video condition in Experiment 2.

In the audio + text + video condition, in addition to eye movement pattern as measured in D-C scale, comprehension accuracy was significantly correlated with Tower of London planning time, $r(101) = 0.311$, $p = .001$. A similar two-stage hierarchical multiple regression was conducted to predict comprehension accuracy. At stage one, Tower of London planning time contributed significantly to the regression model, $B = 0.301$, $F(1, 97) = 9.696$, $p = .002$, and accounted for 9.1% of the variance. Eye movement pattern as measured in D-C scale accounted for an additional 5.8% of the variance in comprehension accuracy and the $R^2$ change was significant, $B = -.244$, $F(1, 96) = 6.499$, $p = .012$. The tests for multicollinearity indicated a low level of multicollinearity (tolerance = 0.969 and 0.969 for Tower of London planning time and eye movement pattern respectively). Thus, when learning from audio lessons with video, participants' online eye movement behavior played a vital role in comprehension accuracy in addition to cognitive abilities, regardless of whether subtitles were presented.

We also examined whether participants demonstrated more consistent eye movement patterns (as measured in entropy) in the audio + text + video condition than the audio + video condition as observed in Experiment 1. Consistent with Experiment 1, the entropy in the audio + video condition, $M = 13.231$, $SD = 0.250$, was significantly higher than that in the audio + text + video condition, $M = 12.577$, $SD = 0.338$, $t(99) = 19.809$, $p < .001$, suggesting that adding subtitles when learning from audio lessons with video made participants' eye movement patterns across videos more consistent.

## 4. Discussion

Here we investigated the effect of video and subtitles on the comprehension of audio lessons, and whether the comprehension depended on individual differences in cognitive abilities and attention allocation behavior as reflected in eye movement pattern. In sum, we showed that subtitles facilitated comprehension whereas video did not. When learning from audio lessons with video, comprehension performance depended on both the learner's cognitive abilities and attention allocation reflected in eye movement behavior.

### 4.1. Empirical contributions

We showed that for audio lessons consisted of statements of facts, subtitles, but not video with relevant images, facilitated comprehension. More specifically, adding subtitles to the audio lessons facilitated comprehension, consistent with the verbal redundancy effect (Adesope & Nesbit, 2012; Moreno & Mayer, 2002). In contrast to the multimedia principle (Mayer & Anderson, 1992), adding video to the audio lessons failed to enhance comprehension, consistent with our hypothesis that the multimedia principle may depend on individual differences in learning ability. Indeed, we found that the comprehension of audio lessons with video depended on both the learner's online eye movement behavior and cognitive abilities: a more centralized eye movement pattern focusing at the screen center and better cognitive abilities predicted better comprehension. This result was consistently observed in both Experiment 1 and 2. Since in our lessons, video contained images relevant to the lesson content, but was not required for comprehending and memorising the lesson content, and thus it may have distracted some learners from the lesson content as reflected in their distributed eye movement pattern, especially those who had lower cognitive abilities.

We also found that adding subtitles to audio lessons with video enhanced comprehension, in contrast to the redundancy principle (Mayer et al., 2001). This finding was consistent with our hypothesis: since important lesson information for comprehension was in the audio but not the video, adding subtitles may help maintain the viewers' attention on the subtitles and enhance comprehension through verbal redundancy (e.g., Kruger & Steyn, 2014). Indeed, we found that adding subtitles to the video clips made participants' eye movements more consistent across videos (as measured in entropy), suggesting that adding subtitles may help guide and maintain participants' attention to the location of the subtitles, which may consequently enhance attention to the content of the audio lessons through verbal redundancy. These results were consistently observed in Experiment 1 and 2. Here we used the machine-learning based, EMHMM method (Chuk et al., 2014) to quantify participants' eye movement pattern, and also used the entropy of HMM to measure eye movement consistency. The EMHMM method takes both regions gazed and the order of the regions gazed into account. Thus, it provides a most comprehensive eye movement pattern and consistency measures than previous studies (e.g., Shic et al., 2008; Stark et al., 2018).

In Experiment 1, we found that the comprehension of audio lessons with video and subtitles could not be predicted by either cognitive abilities or eye movement pattern. This finding was in contrast to the comprehension of audio lessons with video but not subtitles, where both cognitive abilities and eye movement pattern were significant predictors. Nevertheless, with better control of the subtitle location and a larger sample size in Experiment 2, we found consistent results where the comprehension of audio lessons with video depended on both individual differences in cognitive abilities and online eye movement behavior, regardless of whether subtitles were presented or not.

### 4.2. Theoretical implications

When learning from multimedia materials, previous studies on the effect of subtitles have reported inconsistent findings, with some showing that on-screen text redundant to the spoken content (such as subtitles) is distracting (Mayer et al., 2001) whereas others suggesting facilitating effects (Kruger & Steyn, 2014). This inconsistency may be due to differences in the amount of information carried in each medium during multimedia learning. When pictorial stimuli contain important content for knowledge acquisition, simultaneous on-screen texts may be distracting and compete for cognitive resources, as the two sources of information are both from the visual channel (Mayer, 2014b), leading to poorer comprehension. Thus, in this case, eliminating redundant on-screen text may facilitate learning, an effect referred as the redundancy principle (Hoffman, 2006; Mayer et al., 2001). In contrast, when the audio narration provide the most important information, such as the lessons used in the current study, subtitles that match well with the audio narration facilitated comprehension. This phenomenon may be because subtitles help maintain participants' attention to the content of the knowledge, decrease information loss due to concurrent processing of video information, and consequently facilitate comprehension (Kruger et al., 2013). Indeed, we found that participants' eye movements became more consistent across video clips when subtitles were presented.

Our results also did not support the multimedia principle. Indeed, previous research has suggested that the multimedia principle can be affected by the learning materials used and the learner's cognitive ability (Fletcher & Tobias, 2005). Here we further showed that when learning from audio lessons with video, where important information for comprehension was in the audio but not the video, individual differences in both online eye movement behavior and cognitive abilities could affect comprehension: a more centralized eye movement pattern focusing on the screen center (and subtitles at the bottom center when they were presented) and better cognitive abilities predicted better comprehension. Interestingly, a similar finding was observed when learning from audio lessons with video and subtitles. These individual differences in cognitive abilities and online eye movement behavior may also be related to the inconsistent findings in the literature. In particular, the importance of online eye movement behavior during multimedia learning has long been overlooked. The finding that the centralized eye movement pattern, as opposed to the distributed pattern, was correlated with better comprehension may be related to cognitive load. According to the cognitive load theory (e.g., Low & Sweller, 2014), engaging in active eye movement planning as demonstrated in the distributed pattern where ROIs were located at different regions of the video may increase cognitive load, resulting in decreased attentional resources for listening comprehension. The EMHMM method allows discovery of representative eye movement patterns from individual patterns in a data-driven fashion and provides quantitative measures of eye movement pattern similarity, leading to this novel finding. This finding has important implications for the literature on the multimedia principle and the redundancy principle in multimedia learning, since it suggests that the learner's online attention allocation behavior as reflected in eye movement pattern may be an important factor to consider when examining the situations under which these principles do or do not hold.

### 4.3. Practical implications

Our results suggest that when using multimedia to facilitate learning, the importance of the information carried in each medium for comprehension may be an important factor to consider. For learning materials where audios contain more critical information than videos, such as lecture recordings or audio lessons as used in our study, adding subtitles can lead to better learning outcomes. Our results also suggested that when learning from audio lessons with images/videos that are relevant to the lesson content, while learners' cognitive abilities, including working memory capacity, task switching, and executive planning abilities, may affect comprehension performance, providing learners with explicit instructions to adopt a more centralized eye movement

pattern as discovered in the current study may enhance their comprehension, especially for those who use distributed eye movement patterns during video viewing. Future work will examine this possibility.

### 4.4. Limitations and future directions

In the current study, we used lessons where the audio contained more important information for comprehension than the video. Thus, the current findings might be limited to the specific type of lessons used. For example, the centralized eye movement pattern may not be advantageous in multimedia materials where videos contained critical information at different screen locations. Adding subtitles may distract learners from the critical visual information, resulting in poorer comprehension performance. In addition, here we mainly tested the rote memory of the lessons and did not use deeper comprehension questions such as applications of the learned knowledge or transfer tests used in previous multimedia learning studies (e.g., Austin, 2009; Mautone & Mayer, 2001; Moreno & Mayer, 1999). Thus, how well our current results can be generalized to these scenarios requires further examinations. For example, redundant on-screen text has been shown to facilitate learning from narrated graphic presentations on retention but not on transfer (Mayer & Johnson, 2008; McCrudden et al., 2014). Thus, participants' better performance in the audio + text + video condition over the audio + video condition observed here may be limited to rote memory but not on transfer. Their performance on transfer may also be affected by how they allocate attention when viewing the videos as reflected in their eye movement patterns. Future work will examine these possibilities.

### 4.5. Conclusions

In conclusion, here we showed that for knowledge acquisition from lessons where the important information is mainly in the audio, subtitles facilitated comprehension, whereas video did not. More specifically, the comprehension of audio lessons depended on participants' cognitive abilities and online eye movement behavior. These findings have important implications for the multimedia principle and the redundancy principle in multimedia learning, since it suggests that the learner's online attention allocation strategy as reflected in eye movement behavior may be an important factor to consider in addition to individual differences in cognitive abilities. It also has important practical implications for instructional design to consider learners' cognitive abilities and explicit instructions on attention allocation. In sum, our findings demonstrated the importance of taking individual differences into account in the research on the science of learning, and eye tracking with EMHMM provides a useful tool for revealing and quantitatively assessing these individual differences.

## Open practices statement

The experiment reported in this article was not formally preregistered. The data and the materials have been made available on a permanent third-party archive at http://doi.org/10.17605/OSF.IO/CEKVR.

## Author contributions

X. Ye and J. H. Hsiao developed the initial idea of the study. All authors contributed to the study design. X. Ye and Y. Zheng performed the programming and data collection. X. Ye performed the data analysis for the earlier version of the paper. Y. Zheng performed the data analysis and interpretation under the supervision of J. H. Hsiao. Y. Zheng drafted the manuscript, and J. H. Hsiao provided critical suggestions and revisions. All authors approved the final version of the manuscript for submission.

## Acknowledgements

We are grateful to Research Grants Council of Hong Kong (Project # 17609117 to Janet Hsiao). We thank the Editor and two anonymous reviewers for the helpful comments.

## Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.learninstruc.2021.101542.

## References

Adesope, O. O., & Nesbit, J. C. (2012). Verbal redundancy in multimedia learning environments: A meta-analysis. *Journal of Educational Psychology, 104*(1), 250.
Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education, 125*, 413–428.
An, J., & Hsiao, J. H. (2021). Modulation of mood on eye movement pattern and performance in face recognition. *Emotion, 21*(3), 617–630. https://doi.org/10.1037/emo0000724
Attneave, F., & Arnoult, M. D. (1956). The quantitative study of shape and pattern perception. *Psychological Bulletin, 53*, 452–471.
Austin, K. A. (2009). Multimedia learning: Cognitive individual differences and display design techniques predict transfer learning with multimedia learning modules. *Computers & Education, 53*(4), 1339–1354.
Baadte, C., Rasch, T., & Honstein, H. (2015). Attention switching and multimedia learning: The impact of executive resources on the integrative comprehension of texts and pictures. *Scandinavian Journal of Educational Research, 59*(4), 478–498.
Brasel, S. A., & Gips, J. (2014). Enhancing television advertising: Same-language subtitles can improve brand recall, verbal memory, and behavioral intent. *Journal of the Academy of Marketing Science, 42*, 322–336.
Chan, A. B., & Hsiao, J. H. (2016). Information distribution within musical segments. *Music Perception, 34*(2), 218–242. https://doi.org/10.1525/mp.2016.34.2.218
Chan, C. Y. H., Chan, A. B., Lee, T. M. C., & Hsiao, J. H. (2018). Eye-movement patterns in face recognition are associated with cognitive decline in older adults. *Psychonomic Bulletin & Review, 25*(6), 2200–2207.
Chan, F. H. F., Barry, T. J., Chan, A. B., & Hsiao, J. H. (2020). Understanding visual attention to face emotions in social anxiety using hidden Markov models. *Cognition and Emotion, 34*(8), 1704–1710. https://doi.org/10.1080/02699931.2020.1781599
Chan, F. H. F., Jackson, T., Hsiao, J. H., Chan, A. B., & Barry, T. J. (2020). The interrelation between interpretation biases, threat expectancies and pain-related attentional processing. *European Journal of Pain, 24*(10), 1956–1967. https://doi.org/10.1002/ejp.1646
Chan, F. H. F., Suen, H., Chan, A. B., Hsiao, J. H., & Barry, T. J. (2021). The effects of attentional and interpretation biases on later pain outcomes among younger and older adults: A prospective study. *European Journal of Pain*. https://doi.org/10.1002/ejp.1853
Chan, F. H. F, Suen, H., Hsiao, J. H., Chan, A. B., & Barry, T. J. (2020). Interpretation biases and visual attention in the processing of ambiguous information in chronic pain. *European Journal of Pain, 24*(7), 1242–1256. https://doi.org/10.1002/ejp.1565
Chuk, T., Chan, A. B., & Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden Markov models. *Journal of Vision, 14*(11), 8-8.
Chuk, T., Chan, A. B., & Hsiao, J. H. (2017a). Is having similar eye movement patterns during face learning and recognition beneficial for recognition performance? Evidence from hidden Markov modeling. *Vision Research, 141*, 204–216.
Chuk, T, Chan, A. B., Shimojo, S., & Hsiao, J. H. (2020). Eye movement analysis with switching hidden Markov models. *Behavior Research Methods, 52*(3), 1026–1043. https://doi.org/10.3758/s13428-019-01298-y
Chuk, T., Crookes, K., Hayward, W. G., Chan, A. B., & Hsiao, J. H. (2017b). Hidden Markov model analysis reveals the advantage of analytic eye movement patterns in face recognition across cultures. *Cognition, 169*, 102–117.
Cover, T. M., & Thomas, J. A. (2006). *Chapter 1 Entropy, relative entropy and mutual information. Elements of Information Theory* (Vols. 12–49). Wiley.
Coviello, E., Chan, A. B., & Lanckriet, G. R. (2014). Clustering hidden Markov models with variational HEM. *Journal of Machine Learning Research, 15*(1), 697–747.
Crottaz-Herbette, S., Anagnoson, R. T., & Menon, V. (2004). Modality effects in verbal working memory: Differential prefrontal and parietal responses to auditory and visual stimuli. *NeuroImage, 21*, 340–351.
D'Ydewalle, G., & De Bruycker, W. (2007). Eye movements of children and adults while reading television subtitles. *European Psychologist, 12*(3), 196–205.
Fenesi, B., Kramer, E., & Kim, J. (2016). Split-attention and coherence principles in multimedia instruction can rescue performance for learners with lower working memory capacity. *Applied Cognitive Psychology, 30*(5), 691–699.
Fletcher, J. D., & Tobias, S. (2005). The multimedia principle. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 117–133). Cambridge University Press.
Florax, M., & Ploetzner, R. (2010). What contributes to the split-attention effect? The role of text segmentation, picture labeling, and spatial proximity. *Learning and Instruction, 20*, 216–224.
Hoffman, B. (2006). *The encyclopedia of educational technology*. San Diego, CA: Montezuma Press.

Hsiao, J. H., An, J., Zheng, Y., & Chan, A. B. (2021). Do portrait artists have enhanced face processing abilities? Evidence from hidden Markov modeling of eye movements. *Cognition, 211*. https://doi.org/10.1016/j.cognition.2021.104616, 104616.

Hsiao, J. H., Chan, A. B., An, J., Yeh, S.-L., & Jingling, L. (2021). Understanding the collinear masking effect in visual search through eye tracking. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-021-01944-7

Hsiao, J. H., & Cheung, K. (2016). Visual similarity of words alone can modulate hemispheric lateralization in visual word recognition: Evidence from modeling Chinese character recognition. *Cognitive Science, 40*(2), 351–372. https://doi.org/10.1111/cogs.12233

Hsiao, J. H., & Lam, S. M. (2013). The modulation of visual and task characteristics of a writing system on hemispheric lateralization in visual word recognition - A computational exploration. *Cognitive Science, 37*(5), 861–890. https://doi.org/10.1111/cogs.12033

Hsiao, J. H., Lan, H., Zheng, Y., & Chan, A. B. (2021). Eye Movement analysis with Hidden Markov Models (EMHMM) with co-clustering. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01541-5

Hyona, J. (2010). The use of eye movements in the study of multimedia learning. *Learning and Instruction, 20*(2), 172–176.

Kruger, J. L., Hefer, E., & Matthew, G. (2013). Measuring the impact of subtitles on cognitive load: Eye tracking and dynamic audiovisual texts. In *Proceedings of the 2013 conference on eye tracking South Africa* (pp. 62–66). ACM.

Kruger, J. L., & Steyn, F. (2014). Subtitles and eye tracking: Reading and performance. *Reading Research Quarterly, 49*, 105–120.

Lau, E. Y. Y., Eskes, G. A., Morrison, D. L., Rajda, M., & Spurr, K. F. (2010). Executive function in patients with obstructive sleep apnea treated with continuous positive airway pressure. *Journal of the International Neuropsychological Society, 16*, 1077–1088.

Low, R., & Sweller, J. (2014). *The modality principle in multimedia learning. The Cambridge Handbook of Multimedia Learning* (2nd ed., pp. 227–246). Cambridge: Cambridge University Press.

Mautone, P. D., & Mayer, R. E. (2001). Signaling as a cognitive guide in multimedia learning. *Journal of Educational Psychology, 93*(2), 377.

Mayer, R. E. (2014a). *Introduction to multimedia learning. The cambridge handbook of multimedia learning* (2nd ed., pp. 1–24). Cambridge: Cambridge University Press.

Mayer, R. E. (2014b). *Cognitive theory of multimedia learning. The cambridge handbook of multimedia learning* (2nd ed., pp. 43–71). Cambridge University Press.

Mayer, R. E., & Anderson, R. B. (1992). The instructive animation: Helping students build connections between words and pictures in multimedia learning. *Journal of Educational Psychology, 84*(4), 444–452.

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.

Mayer, R. E., & Johnson, C. I. (2008). Revising the redundancy principle in multimedia learning. *Journal of Educational Psychology, 100*(2), 380–386.

McCrudden, M. T., Hushman, C. J., & Marley, S. C. (2014). Exploring the boundary conditions of the redundancy principle. *The Journal of Experimental Education, 82*(4), 537–554.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49–100.

Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia design: The role of modality and contiguity. *Journal of Educational Psychology, 91*, 358–368.

Moreno, R., & Mayer, R. E. (2002). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology, 94*, 156–163.

Nystrom, L. E., Braver, T. S., Sabb, F. W., Delgado, M. R., Noll, D. C., & Cohen, J. D. (2000). Working memory for letters, shapes, and locations: fMRI evidence against stimulus-based regional organization in human prefrontal cortex. *NeuroImage, 11*, 424–446.

Phillips, L. H., Wynn, V. E., McPherson, S., & Gilhooly, K. J. (2001). Mental planning and the tower of London task. *Quarterly Journal of Experimental Psychology, 54*, 579–597.

Schnotz, W. (2005). *Integrated model of text and picture comprehension. The cambridge handbook of multimedia learning* (2nd ed., pp. 72–103). Cambridge: Cambridge University Press.

Shic, F., Chawarska, K., Bradshaw, J., & Scassellati, B. (2008). Autism, eye-tracking, entropy. In *Proceedings of the 2008 7th IEEE international conference on development and learning* (Vols. 73–78)Monterey, CA: IEEE.

Starbek, P., Erjavec, M. S., & Peklaj, C. (2010). Teaching genetics with multimedia results in better acquisition of knowledge and improvement in comprehension. *Journal of Computer Assisted Learning, 26*(3), 214–224.

Stark, L., Brünken, R., & Park, B. (2018). Emotional text design in multimedia learning: A mixed-methods study using eye tracking. *Computers & Education, 120*, 185–196.

Stoet, G., O'connor, D., Conner, M., & Laws, K. (2013). Are women better than men at multi-tasking? *BMC Psychology, 1*(1), 18.

Sweller, J. (2005). The redundancy principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning, 159–167*. New York, NY: Cambridge University Press.

Van Gog, T., & Scheiter, K. (2010). Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction, 20*(2), 95–99.

Zhang, J., Chan, A. B., Lau, E. Y. Y., & Hsiao, J. H. (2019). Individuals with insomnia misrecognize angry faces as fearful faces while missing the eyes: An eye-tracking study. *Sleep, 42*(2), zsy220. https://doi.org/10.1093/sleep/zsy220