






Cognitive Science 47 (2023) e13247

© 2022 The Authors. *Cognitive Science* published by Wiley Periodicals LLC on behalf of Cognitive Science Society (CSS).

ISSN: 1551-6709 online

DOI: 10.1111/cogs.13247

Did You Get That? Predicting Learners' Comprehension of a Video Lecture from Visualizations of Their Gaze Data

Ellen M. Kok,^{a,b}  Halszka Jarodzka,^b  Matt Sibbald,^c  Tamara van Gog^a

^a*Department of Education, Utrecht University*

^b*Department of Online Learning and Instruction, Open University of the Netherlands*

^c*McMaster Education Research, Innovation and Theory (MERIT) Program, Faculty of Health Sciences, McMaster University*

Received 15 August 2021; received in revised form 20 December 2022; accepted 4 January 2023

Abstract

In online lectures, unlike in face-to-face lectures, teachers lack access to (nonverbal) cues to check if their students are still “with them” and comprehend the lecture. The increasing availability of low-cost eye-trackers provides a promising solution. These devices measure unobtrusively where students look and can visualize these data to teachers. These visualizations might inform teachers about students' level of “with-me-ness” (i.e., do students look at the information that the teacher is currently talking about) and comprehension of the lecture, provided that (1) gaze measures of “with-me-ness” are related to comprehension, (2) people not trained in eye-tracking can predict students' comprehension from gaze visualizations, (3) we understand how different visualization techniques impact this prediction. We addressed these issues in two studies. In Study 1, 36 students watched a video lecture while being eye-tracked. The extent to which students looked at relevant information and the extent to which they looked at the same location as the teacher both correlated with students' comprehension (score on an open question) of the lecture. In Study 2, 50 participants watched visualizations of students' gaze (from Study 1), using six visualization techniques (dynamic and static versions of scanpaths, heatmaps, and focus maps) and were asked to predict students' posttest performance and to rate their ease of prediction. We found that people can use gaze visualizations to predict learners' comprehension above chance level, with minor differences between visualization techniques. Further research should

Correspondence should be sent to Ellen M. Kok, Department of Education, Utrecht University, P.O. Box 80140, 3508 CS, Utrecht, The Netherlands. E-mail: e.m.kok@uu.nl

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

investigate if teachers can act on the information provided by gaze visualizations and thereby improve students' learning.

Keywords: Eye-tracking; Gaze; Gaze visualization; Video lectures; Teacher assessment; With-me-ness

1. Introduction

The use of online (video) lectures was already on the rise but has expanded due to the COVID-19 pandemic, revealing the many challenges involved in online lecturing. One of them is that in online lectures, unlike in face-to-face lectures, teachers have reduced access to (nonverbal) cues to assess whether their students can comprehend the lecture and thus learn from it (Hew & Cheung, 2014). This is a problem as teachers base their instructional decisions on their assessments of students' comprehension (Südkamp, Kaiser, & Möller, 2012). Teachers' assessment of whether the students can "follow" the lecture and are "with the teacher," is called "with-me-ness", which correlates with learning (Sharma, Jermann, & Dillenbourg, 2014).

The "with-me-ness" of a student in an online setting can be quantified using eye-tracking technology (Sharma et al., 2014). An eye tracker is a device that can be used to unobtrusively capture at what part of the instructional material a learner looks, for how long, and in which order (Holmqvist et al., 2011). Since there is a tight link between eye movements and attention (Liversedge & Findlay, 2000), and attention is often a prerequisite for learning (Kok & Jarodzka, 2017b), eye tracking can be used to assess cognitive processes related to learning. As eye trackers are becoming cheaper and smaller (e.g., Tobii, <https://gaming.tobii.com/product/eye-tracker-5/>), and the quality of webcam-based eye-tracking is increasing (Papoutsaki et al., 2016), they become increasingly interesting for educational practice. Thus, recent research has started to explore how eye tracking can not only be used to investigate but also improve education (Jarodzka, Holmqvist, & Gruber, 2017; Scheiter & van Gog, 2009; van Gog & Jarodzka, 2013; Van Gog & Scheiter, 2010; van Gog, Kester, Nievelstein, Giesbers, & Paas, 2009).

Eye-tracking devices allow not only for recording gaze but also for visualizing gaze on top of instructional material. Three applications of visualizing gaze can be imagined in education: First, teachers' gaze can be visualized to students (eye movement modeling examples) to improve their learning from video examples (Chisari et al., 2020; Jarodzka, van Gog, Dorr, Scheiter, & Gerjets, 2013; Mason, Pluchino, & Tornatora, 2015; Scheiter, Schubert, & Schüler, 2018; Van Gog, Jarodzka, Scheiter, Gerjets, & Paas, 2009). Second, we can show learners where they themselves were looking to improve their self-assessment and self-regulated learning (Donovan, Manning, & Crawford, 2008; Eder et al., 2020; Henneman et al., 2014; Kok et al., 2022; Kok, Aizenman, Vö, & Wolfe, 2017; Kostons, van Gog, & Paas, 2009). A third possible application is to show teachers where students were looking to improve teachers' assessment of students' performance on a task (see Knoop-van Campen et al., 2021; Špakov, Siirtola, Istance, & Riih , 2017, e.g., in reading). Several researchers have investigated to what extent people can interpret visualizations of other people's gaze

(Bahle, Mills, & Dodd, 2017; Emhardt, van Wermeskerken, Scheiter, & van Gog, 2020; Foulsham & Lock, 2015; Greene, Liu, & Wolfe, 2012; Van Wermeskerken, Litchfield, & van Gog, 2018; Zelinsky, Peng, & Samaras, 2013). This is a promising application that might be able to provide teachers with cues about their students' with-me-ness, and thereby, how much they learn from the lectures.

There are three prerequisites for presenting students' gaze to teachers to provide them with cues about students' with-me-ness and comprehension, which we will investigate in two studies. First, gaze measures should be related to measures of learning from video lectures (posttest performance, which reflect learners' comprehension) to be useful. Second, teachers without training or prior experience with gaze data (effectively laypeople with respect to eye tracking) would have to be able to predict students' posttest performance based on their gaze visualizations. Finally, we need to understand whether and how different gaze visualization techniques impact this prediction. As there are many different possibilities for visualizing gaze data, the question is which gaze visualization technique is best in terms of speed and perceived ease of interpretation. The first prerequisite is addressed in Study 1, and the second and third prerequisites are addressed in Study 2, using visualizations of the gaze data collected in Study 1. Below, we discuss each of the three prerequisites in more detail before introducing the studies.

1.1. Are gaze measures related to comprehension of instructional videos?

The underlying rationale of expecting gaze measures to relate to comprehension of instruction videos comes from the cognitive theory of multimedia learning (Mayer, 2014). This theory outlines that when learning from video lectures that use multimedia materials (text and pictures), learners have to select (i.e., attend to) the relevant verbal and pictorial information, mentally organize the selected information, and integrate it with their existing knowledge (Mayer, 2014). Since the information in videos is mostly transient, selecting the right information at the right time (i.e., selecting the related visual and verbal information) is necessary for building a coherent mental representation (Ayres & Paas, 2007; de Koning & Jarodzka, 2017). Usually, selecting visual information is done by looking at the information, so we can measure the selection process with an eye-tracking device (Jarodzka et al., 2017; Kok & Jarodzka, 2017a). Note here that looking at the relevant information is a necessary (but not sufficient) first step for organizing and integrating it with existing knowledge (Kok & Jarodzka, 2017b): If a learner does not "follow" what the teacher is currently talking about, for example, if the teacher uses jargon that the learner does not know yet or if teachers verbal utterances are ambiguous, learners could have trouble locating the visual information related to the verbal information (Van Marlen, van Wermeskerken, & van Gog, 2019). This could lead to the learner missing (i.e., not selecting) important visual information and thus not organizing verbal and visual information together, and integrating it with existing knowledge. This means that learning might be hampered (Richter, Scheiter, & Eitel, 2016).

"With-me-ness" is the extent to which learners look at visual information that the teacher refers to verbally (Sharma and colleagues call this *conceptual* with-me-ness). Sharma et al. (2014) found that students who looked (fixated) longer at the information that the teacher

was talking about (relevant information) had higher posttest scores than students who looked at this information for a shorter amount of time. In the context of multimedia learning, a review found that often, but not always, the amount of attention allocated to relevant pictures was associated with learning performance (Alemdag & Cagiltay, 2018). Likewise, experts in visual tasks are found to spend a larger percentage of their time looking at relevant information compared to novices (Reingold & Sheridan, 2011), and with increased experience in a task, people start ignoring irrelevant information and focus more on relevant information (Haider & Frensch, 1996).

Furthermore, Sharma et al. (2014) posed, but did not test, that with-me-ness is comparable to “gaze coupling.” Gaze coupling, or cross-recurrence, is a measure of how much the gazes of two people follow each other during interaction (Richardson & Dale, 2005). Richardson and Dale (2005) found that the more alike a speaker’s gaze and a listener’s gaze on a visual stimulus were (gaze positions in terms of time and space), the better the listener understood what the speaker said (i.e., performance on a later comprehension test). Note that they found a lag of approximately 2 s between the moment a speaker would look at the information they talked about, and the moment the listener would look at that information, partly caused by the speaker looking at information before talking about it (cf. Griffin & Bock, 2000) and partly caused by the speaker searching for the information after hearing this information, meaning that it is important to consider this lag when analyzing gaze coupling.

In line with the findings of Richardson and Dale (2005), it was found that a higher gaze coupling correlates with increased interaction quality (Jermann & Nüssli, 2012; Sharma, Leftheriotis, Noor, & Giannakos, 2017), and this results in better learning (Sharma et al., 2017). In the context of collaborative learning, Cakir and Uzunosmanoğlu (2014) found that higher-achieving dyads on average exhibited higher gaze coupling, although Villamor and Rodrigo (2018) found no significant correlations between gaze coupling rate and task performance in a pair-programming task. Overall, gaze coupling seems to be related to interaction quality and performance, but it is yet unknown if the gaze coupling between a teacher and a student in video lectures is correlated with students’ learning from video lectures.

1.2. Can laypeople predict students’ comprehension from gaze visualizations?

If the prerequisite that gaze measures are related to students’ posttest scores is met (prerequisite 1), the next prerequisite for using gaze visualizations in education is that teachers, who are laypeople in the interpretation of gaze data, should also be able to predict posttest performance based on these gaze visualizations (prerequisite 2). Several studies have shown that people can, to some extent, interpret gaze visualizations in terms of the perceptual and/or cognitive processes of the observer (Bahle et al., 2017; Emhardt et al., 2020; Foulsham & Lock, 2015; Greene et al., 2012; Van Wermeskerken et al., 2018; Zelinsky et al., 2013). In those studies, the gaze data of a group of participants (observers) are first recorded and subsequently presented to a new group of participants with the task to infer those processes from the visualization.

For example, Zelinsky et al. (2013) presented participants with visualizations of the gaze of observers who searched for bears or butterflies among distractors. The visualizations were

static with the first fixated object marked with a green circle and the other objects marked with red circles. Participants could generally infer the search target from a visualization of observers' gaze at above-chance levels. Accuracy was especially high when inferring the target of the observer after several subsequent trials (76–100%). Foulsham and colleagues (2015) presented observers with four abstract pictures (fractals) and asked them for their preferences. Participants could guess which picture was preferred by initial observers at above-chance levels (over 60% correct) based on dynamic gaze visualizations (i.e., a depiction of the visual focus over time in the form of a moving dot). In both situations, gaze visualizations showed which part of a stimulus was looked at most/longest, and this information was correctly used by participants to infer picture preference. Similar results were found in an educational context. Emhardt et al. (2020) presented participants with gaze visualizations of observers doing graph comprehension tasks with the related multiple-choice question presented underneath. Based on dynamic gaze visualizations, participants could infer observers' choice for one of the four answer options at above-chance levels (overall accuracy 54%). Presumably, participants' inferences were typically based on which of the answers was looked at most.

In all three studies, participants seemed to have relied mostly on the gaze bias effect (Lindner et al., 2014; Shimojo, Simion, Shimojo, & Scheier, 2003): Observers show an attentional bias toward the preferred answer option, and this can be used by participants as an indication of their choice (Emhardt et al., 2020). Even though not all of the above-mentioned studies took place in the context of education, it seems that participants are generally able to use information about how long people looked at (relevant) information as a cue to infer cognitive processes, which makes it likely that participants can also use this cue in the context of video lectures to predict posttest performance. No study has yet investigated whether a participant can also use information about gaze coupling between teacher and student to predict posttest performance.

1.3. How do different visualization techniques impact performance predictions?

So far, studies that investigated gaze visualization interpretation have mostly displayed gaze as (simplified) scanpaths, in which fixations (moments during which the eyes are relatively still and take in information) are shown as circles or crosshairs, and saccades (jumps between fixations that represent attention reallocation) as lines between those fixations (Bahle et al., 2017; Emhardt et al., 2020; Foulsham & Lock, 2015; Van Wermeskerken et al., 2018; Zelinsky et al., 2013). However, different types of visualizations might have different affordances for interpretation (Blascheck et al., 2014; Kurzhals et al., 2015). Thus, investigating how the type of visualization impacts interpretation performance, speed, and preference is important to optimize this process (prerequisite 3). Indeed, Bahle et al. (2017) required participants to classify gaze visualizations as coming from a search, memory, or rating task. They found that the type of visualization impacted the classification result. For instance, participants were most likely to correctly classify a gaze visualization as reflecting that the observer was executing a search task if it only showed fixations and not saccades. This visualization conveyed the relevant gaze information (when searching, observers make a large number of short fixations).

The two most commonly used visualizations of gaze data are the scanpath (see Fig. 1a) and the attention map (see Fig. 1b and c) (Blascheck et al., 2014; Kurzhals et al., 2015).

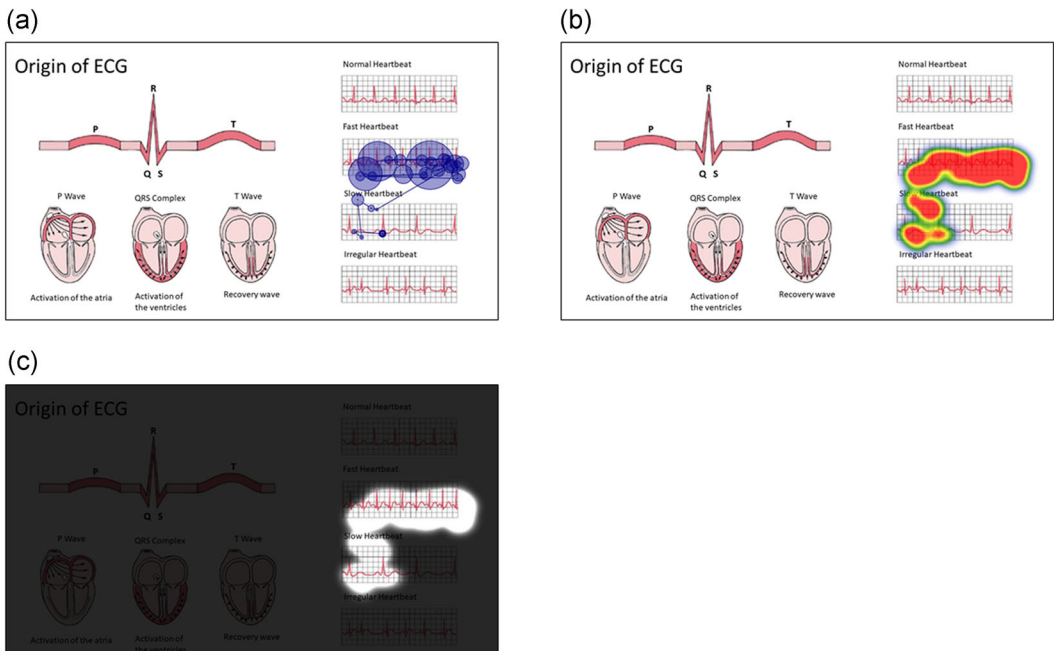


Fig. 1. Example gaze visualizations (a) scanpath, (b) heatmap, and (c) focus map.

Note. The authors' research findings have been superimposed on an electrocardiogram drawing (stimulus in the present study), supplied by the MSD Manuals, edited by Robert Porter. Copyright (2021) by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ. Available at <https://www.msdmanuals.com/home/heart-and-blood-vessel-disorders/diagnosis-of-heart-and-blood-vessel-disorders/electrocardiography>, accessed (3-8-2021).

A *scanpath* is a sequence of fixations and saccades on a stimulus (Blascheck et al., 2014; Holmqvist & Anderson, 2017; Holmqvist et al., 2011), in which, typically, fixations are indicated by circles and saccades are indicated by lines. Some studies use only visualizations of saccades as scanpaths (Bahle et al., 2017). Whereas a scanpath allows for seeing the order of fixations, it is argued that this kind of visualization, especially in a static format, results in visual clutter which makes it difficult to appreciate patterns (Blascheck et al., 2014).

Attention maps aggregate fixations (or raw data) over time (Blascheck et al., 2014; Holmqvist & Anderson, 2017; Holmqvist et al., 2011). Two different types of attention maps are heatmaps and focus maps. The *heatmap* is among the most widely used attention maps (Bojko, 2009). Heatmaps represent values as colors to visualize spatial patterns of attention, that is, how much attention was allocated to certain regions (Bojko, 2009). Warmer colors usually reflect longer and/or more fixations. Heatmaps can be agnostic of fixation duration or be weighted by fixation duration. In the former case, "hot" regions are looked at more often than "cold" regions, whereas in the latter case, "hot" regions are looked at more often and longer than "cold" regions. Attention maps make it easy to identify which regions on a

stimulus attract much attention (Kurzahls et al., 2015), and they can be aggregated across people to understand group trends in where people look. They are considered compelling and intuitive (Bojko, 2009). Unlike scanpaths, however, the duration of individual fixations is not represented in an attention map. Since the duration of a fixation tells us about cognitive processing (Rayner, 1998) attention maps be misleading (Bojko, 2009).

Focus maps or luminance maps have the same affordance as heatmaps, but use filter approaches to reduce color saturation or sharpness in unattended regions instead of adding information (Blascheck et al., 2014). The benefit of this approach is that irrelevant information is removed, whereas in heatmaps, the relevant information is added. From a perspective of cognitive (load) theories on multimedia learning (Mayer, 2014; Sweller, Van Merriënboer, & Paas, 1998), it is important not to overload a beginning learner with too much information, otherwise learning cannot take place. Because in focus maps, unimportant information is filtered out, it could be argued that cognitive load is reduced and interpretation performance and perceived ease of interpretation might be improved. Indeed, there is some evidence that students may learn more from seeing a teacher's gaze as a focus map (irrelevant information removed), than from seeing a teacher's gaze as a scanpath (Jarodzka et al., 2012; Jarodzka et al., 2013). In addition, the part of the instructional material that is focused on is clearly visible as opposed to obscured by the graphical representation, like in the scanpath and heatmap visualizations.

Furthermore, gaze visualizations can either be presented in a *static* (i.e., a picture with all gaze data overlaid) or *dynamic* (i.e., a video with gaze overlaid in a moment-to-moment fashion) version. On the one hand, dynamic gaze visualizations provide information about the order of looking, which is completely missing in static focus maps or heatmaps, and more difficult to extract in scanpaths. Van Wermeskerken et al. (2018) found that this order information helped participants to interpret gaze visualizations for some of the tasks. On the other hand, since dynamic gaze visualizations provide information in a transient manner, this might result in a higher cognitive load for viewers, as they have to keep track of and integrate the gaze information over time.

Overall, several claims have been made regarding the affordances of different types of visualizations for the correct interpretation of gaze visualizations. However, it has not yet been investigated if there are indeed differences between static and dynamic versions of scanpaths, heatmaps, and focus maps with regard to the interpretation performance, perceived ease, and speed of interpretation of gaze visualizations of learners looking at instruction videos.

1.4. *The present studies*

In sum, this research investigates three prerequisites for the utility of showing visualizations of learners' gaze to teachers for online monitoring of learners' comprehension. Study 1 addresses the question (1) Are gaze measures related to posttest performance in instruction videos? (prerequisite 1) Study 2 addresses the questions (2) Can laypeople predict students' posttest performance from gaze visualizations? (3) How do different visualization techniques impact performance predictions? (prerequisites 2 and 3). In Study 2, we use gaze visualizations developed based on data from Study 1.

2. Study 1

The aim of Study 1 was to investigate whether gaze measures of with-me-ness correlate with learning from instruction videos. Two gaze measures that reflect “with-me-ness” were included: the proportion fixation time on relevant information, and gaze coupling. The proportion fixation time on relevant information is the time spent fixating information that the teacher talks about divided by the sum of all fixation durations. In line with the cognitive theory of multimedia learning (Mayer, 2014), we expect to replicate the finding by Alemdag and Cagiltay (2018) that the amount of attention allocated to relevant pictures is associated with learning outcomes. How closely teachers’ and learners’ gaze are coupled (gaze coupling) is quantified using cross-recurrence analysis (see Coco & Dale, 2014; Richardson & Dale, 2005, for more information about this analysis). We expect to replicate the findings by Richardson and Dale (2005) in the context of communication in an educational setting, and expect that the closer a speaker’s (in our case: teacher’s) gaze and a listener’s (in our case: learner’s) gaze on a visual stimulus are coupled (in terms of time and space, i.e., higher gaze coupling), the better the learner performs on a posttest. Listeners, however, need some time to look at the information that the speaker refers to, and speakers tend to look at information shortly before the information is mentioned, so the gaze coupling between speaker and listener is higher if we take into account that there is a lag between them (Richardson & Dale, 2005), so we look at the gaze coupling for the optimal lag between speaker and each participant. Furthermore, we extend their measure of posttest performance (participants’ judgments of whether the speaker said certain information) with a set of true/false questions (similar in form, but more common in education), and an open question.

2.1. Methods

2.1.1. Participants and design

Participants were 36 (30 female, 6 male) higher education students or recent graduates (< 1 year ago) with a nonmedical background (i.e., laypeople), mean age 22.9 ($SD = 2.1$). English was the mother tongue ($n = 1$) or second language ($n = 35$) of participants. Three participants reported some prior knowledge of the task (e.g., knew about different heartbeats). All participants reported normal or corrected-to-normal vision and none were found to be colorblind. All participants provided written informed consent. This study was approved by the Research Ethics Committee of the first author’s institute. Participants received a compensation of 2 euros.

An a-priori power-analysis in G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) showed that with a sample of 52 participants ($\alpha = 0.05$, one-tailed test), assuming null-hypothesis significance testing as originally planned, we would have a power of 0.8 to find a correlation of 0.33 between gaze coupling and comprehension performance. However, due to lab closure as a result of the first corona-wave, we had to finish data collection after 36 participants already. Since this was the same number of participants of that of Richardson and Dale (2005), we decided to analyze this sample.

2.1.2. Apparatus

The teacher's gaze was recorded at 120 Hz using an SMI RED Mobile eye tracker using SMI Experiment Center software (Version 3.7), and participants' gaze was recorded at 250 Hz using an SMI RED eye tracker using SMI Experiment Center software (Version 3.7). For participants, a headrest was used to stabilize the head position at approximately 60 cm from the center of the screen. The stimuli were presented on a 22-in monitor (1680×1050 pixels) which subtended 44° of visual angle horizontally and 28° of visual angle vertically. The SMI high-velocity event-detection algorithm was used with a velocity threshold of 40 degrees/s and a minimum fixation duration of 50 ms. The system was calibrated using a nine-point calibration procedure with four-point validation. Calibration was repeated a maximum of three times until the average accuracy for the x-axis and y-axis was lower than 1.0 degrees of visual angle and preferably lower than 0.5. The average deviation was $M_x = 0.48$ ($SD_x = 0.21$), $M_y = 0.50$ ($SD_y = 0.16$). No drift correction was applied.

2.1.3. Materials

Instruction videos A medical teacher and expert in electrocardiogram (ECG) interpretation (MS) developed two videos: A 30-s introductory video that familiarized participants with the material and explained how an ECG is made and an instruction video of 3 min and 47 s explaining how an ECG shows the functioning of the heart. The instruction video featured a single image (see background picture in Fig. 1) that had different elements (the ECG, the pictures of the heart, and the pictures of example ECGs), and included a spoken narration by the teacher. The narrative was recorded at the same time as the teacher's eye movements. In Study 1, the video did not show the gaze of the expert nor a video of the teacher overlaid on the screen, just a static picture with spoken text. The spoken text was not captioned.

Posttests Three posttests were used: First, an open question was posed. Next, to replicate Richardson and Dale's (2005) posttest measure, we included questions of the type "did the speaker say that..." As a similar, but more educationally relevant measure, we added true/false questions.

Open question The open question was: "In this video, you were taught how an ECG reflects the electrical activation of the heart. Explain in your own words the heart cycle and how this is represented in the ECG (p-wave, QRS complex, and t-wave). Next, explain how you can recognize a slow, fast, and irregular heartbeat."

"Did the speaker say?" questions The "did the speaker say..." questions were 12 sentences (presented one by one) of which the participant had to decide whether or not the speaker said that (by clicking yes or no). Participants were instructed to click no if a certain sentence was not said, even if the sentence were true. Five of the sentences were said, for five of the sentences, one or a few words were changed (e.g., "Did the speaker say that the heart has two lower pumping chambers called the atria?" while the speaker said ventricles instead of atria), and two sentences featured information that was not conveyed in the video (e.g., Did the speaker say that the right atrium pumps blood to the lungs?). Additionally, an

example question was developed that was presented after the introduction video to familiarize participants with the question format.

True/false questions The true/false questions consisted of single statements and participants were required to state whether the statement was true or false based on the video (e.g., The relaxation of the heart results in the T wave). Those statements were not literally present in the video, but were worded slightly different. Eight of the statements were false, and eight were true. For half of the questions, the visual information in the video was useful, for the other half, only the verbal information sufficed. Additionally, an example question was developed that was presented after the introduction video to familiarize participants with the question format.

Color blindness test To screen for color blindness, a paper-based Ishihara color blindness test was used (Ishihara, 2017). This test consists of 11 plates (and three additional plates for detailed screening) with embedded numbers that participants are required to read aloud. The numbers are clearly visible for participants with normal color vision but not for participants with different color deficiencies. Participants with more than one error might have color deficiencies. None of the participants had to be excluded for making more than one error.

2.1.4. Procedure

Participants were tested in individual sessions of approximately 15 min in a sound-proof room. First, participants filled out a questionnaire to report demographic data and were screened for color-blindness. Next, the eye tracker was calibrated. Next, participants saw the introduction video and answered the two example questions. Once the procedure was clear, participants were asked to watch the instructional video while their gaze was tracked. They were not able to pause or replay the video. After the video, participants answered the open question, the “did the speaker say that...” questions and true/false questions. All questions were presented visually and participants clicked on or typed in their answers. The order of the “did the speaker say that...” and “true/false” questions was randomized within each type. Finally, to check that they had indeed learned the information from the video (instead of already knowing the information), participants were asked to report whether there were any aspects of the information taught in the video that were already known to them.

2.1.5. Measures and analyses

Posttest scores The answers to the open question were scored using a coding scheme of 17 items, in which each item was a specific idea unit, and that together specified the eight topics: heart cycle, P-wave, break, QRS complex, T-wave, slow heartbeat, fast heartbeat, and irregular heartbeat. A score of 0.5 was awarded for each idea unit with possible scores between 0 and 8.5. A random selection of 10% of the data (four participants) was scored by two coders, the Krippendorff’s alpha as calculated with the KALPHA macro in SPSS (Hayes & Krippendorff, 2007) was acceptable at .77. The other answers were scored separately by one of the coders. For the “did the speaker say...” questions and the true/false questions, the

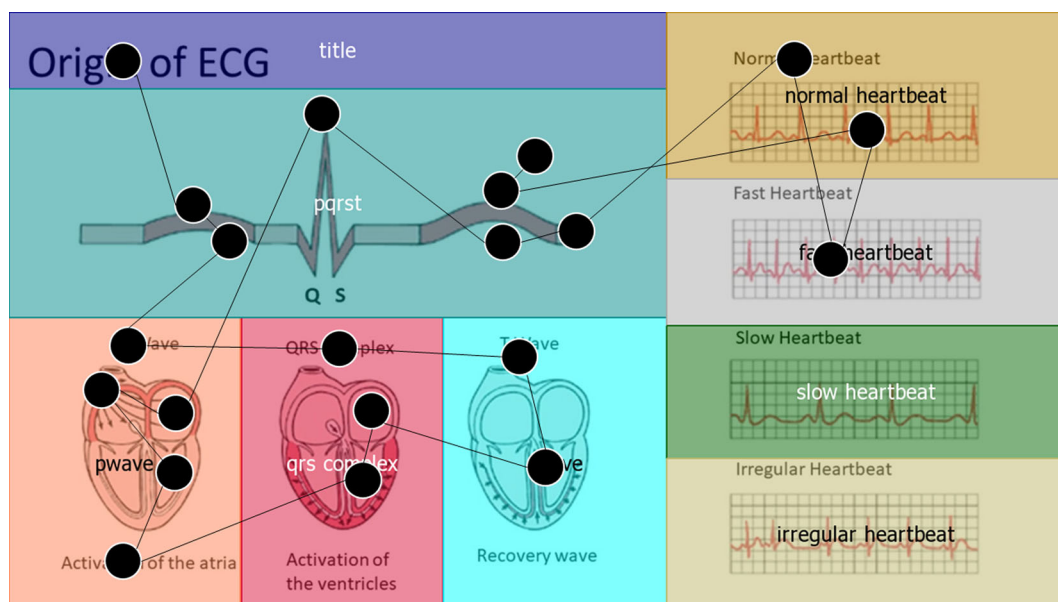


Fig. 2. Areas of interest and an example (simplified) scanpath of a listener.

Note. Areas of interest (as designated by colors) and an example scanpath have been superimposed on an electrocardiogram drawing, supplied by the MSD Manuals, edited by Robert Porter. Copyright (2021) by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ. Available at <https://www.msmanuals.com/home/heart-and-blood-vessel-disorders/diagnosis-of-heart-and-blood-vessel-disorders/electrocardiography>, accessed (3-8-2021). Circles denote fixations, lines denote saccades. Note that unlike in actual data, all fixations are of equal duration (same size) in this picture.

total score was the number of correct answers (scores between 0 and 12 for “did the speaker say...” and between 0 and 14 for true/false questions).

Average proportion of fixation time on relevant information For each fragment of the video, an area of interest (AOI) was drawn around the relevant information (see Fig. 2). The relevant information was the element of the image that the teacher was talking about. The AOIs covered between 8.4% and 23.8% of the screen ($M = 11.48\%$). The average proportion of fixation time on relevant AOIs was the sum of all fixation durations inside these AOIs divided by the total fixation time for the entire video (sum of all fixation durations).

Gaze coupling The basis of this analysis is that the gaze of the teacher and the learner are recurrent (coupled) if they are on the same object (in the same AOI) at the same time. However, it was found that speakers look at an object shortly before mentioning it, and listeners look at it slightly after it was mentioned (Richardson & Dale, 2005). Therefore, we established the optimal lag for each participant, and quantified gaze coupling as the proportion gaze coupling for the optimal lag.

First, participants' gaze data were resampled using a custom script written in the programming language TCL Version 8.5 (TCL, 2021) to be exactly 120 Hz so that each sample of each participant and the teacher had the exact same timing. Resampling was necessary because the SMI down samples when the eye is not detected, so minor differences in file length exist between participants and experts. Resampling improved the precision of the gaze-coupling measure. Recurrence between the teacher's and a participant's gaze (gaze coupling) was calculated as the maximum proportion of recurrence between teacher and student using the same procedure as Richardson and Dale (2005). We did not conduct event classification but worked with gaze samples. For each sample of gaze data, the gaze location was assigned to one of the AOIs (see Fig. 2). This can be graphically depicted using a scarf plot, which represents each gaze sample as a rectangle with the color reflecting the AOI looked at (see Fig. 3). Scarf plots of the speaker and the listener were compared for each time point to see if they were in the same AOI. If data for either the speaker or the listener were missing (e.g., during blinks), a sample was considered to be not overlapping. The total gaze coupling was the proportion of samples in the same AOI. Listeners, however, need some time to look at the information that the speaker refers to, and speakers tend to look at information shortly before the information is mentioned, so the gaze coupling between speaker and listener is higher if we take into account that there is a lag between them (Richardson & Dale, 2005). Thus, in line with Richardson and Dale (2005), we calculated the individual proportion gaze coupling for a range of 200 ms lags between $-4,000$ and $+10,000$ ms. The proportion gaze coupling for the lag that results in the highest proportion gaze coupling (i.e., optimal lag) was used as a measure of gaze coupling between the teacher and the student. Additionally, to check that the gaze coupling measure reflects actual gaze coupling, and not just alignment of the listeners' gaze with the teachers' gaze based on chance, a baseline measure of gaze coupling was created by randomly reordering the gaze samples of the listener before calculating the gaze coupling as described above.

Analyses Analyses were conducted in JASP (JASP Team, 2021). For all analyses, we report the Bayes factor (BF_{10}) to quantify the evidence in favor of the hypothesis that there is a correlation between gaze measures and posttest scores (i.e., a $BF_{10} = 3$ means the data are three times more likely under $H_a: r \neq 0$ than under H_0) (Marsman & Wagenmakers, 2017). The evidence in favor of the null-hypothesis (BF_{01}) can be calculated as $1/BF_{10}$, so if the BF_{10} is lower than 1 and approaches 0, the data are increasingly more likely under the null-hypothesis and a $BF_{10} = 1/3$ means data are three times more likely under H_0 than under $H_a: r \neq 0$. Guidelines for the interpretation of Bayes factors differ widely and many statisticians argue against any cut-off values as they are arbitrary (cf. $p < .05$). Even so, most guidelines are similar in that Bayes factors between $1/3$ and 3 are considered ignorable evidence (e.g., Kass & Raftery, 1995), that is, there is uncertainty whether there is a correlation between the gaze measure and posttest score. Bayes factors larger than 3 and smaller than $1/3$ are interpreted as increasingly stronger evidence in favor of the alternative hypothesis and null-hypothesis, respectively.

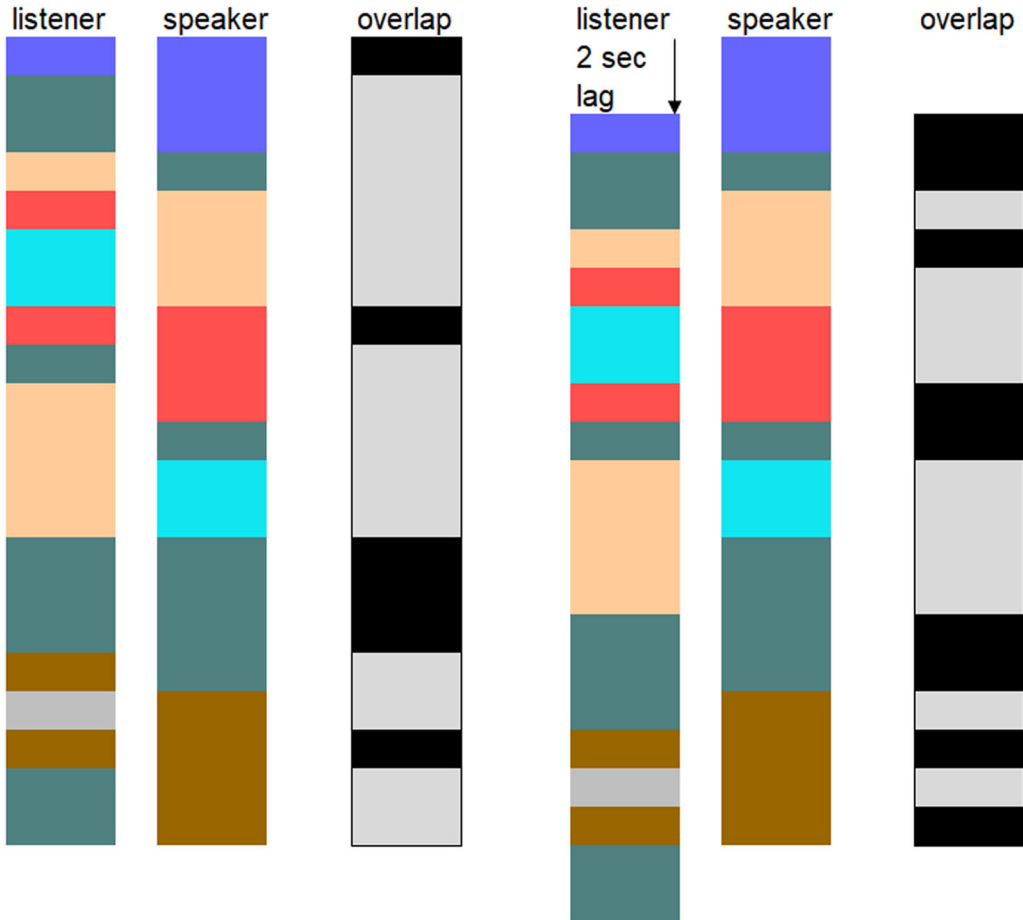


Fig. 3. Example scarf plots and gaze coupling calculation.

Note. Scarf plots represent each fixation in Fig. 2 (listener) as a rectangle with the color reflecting the AOI looked at. The order of fixations is plotted from top to bottom. The overlap bar is black for overlap in AOIs and gray if there was no overlap. In this example, gaze coupling without lag is 28.6%, with lag is 47.4%. Note that in the analyses, we did not compare fixations (as they differ in length) but samples, which include blinks (missing data), fixations, and saccades.

2.2. Results

Descriptive statistics for the posttest scores and the gaze measures can be found in Table 1. All correlations between posttest scores and gaze measures can be found in Table 2.

To check that the gaze coupling measure reflects actual gaze coupling, and not just alignment of the listeners' gaze with the teachers' gaze based on chance, we calculated the baseline gaze coupling, which is the maximum overlap between teachers' gaze and a randomly reordered version of the learner's gaze. The average maximum gaze coupling is higher than

Table 1
Descriptive statistics for the posttest scores and gaze measures

Statistic	<i>N</i>	Minimum	Maximum	Mean	<i>SD</i>
<i>Posttest scores</i>					
Total score open question ^a	36	0.5	6.5	2.93	1.55
Total score did the speaker say... ^b	36	5.00	11.00	8.03	1.54
Total score true/false ^c	36	5.00	13.00	9.14	2.14
<i>Gaze measures</i>					
Average proportion of fixation time on relevant information	36	.43	.78	.65	.08
Optimal lag in seconds	36	-0.8	8.2	2.79	1.67
Maximum gaze coupling	36	.34	.72	.50	.09

^aMaximum possible score = 8.5.

^bMaximum possible score = 12.

^cMaximum possible score = 14.

Table 2
Correlations between the two gaze measures and the three posttest scores

Variable	Did the speaker say ... questions			True/false questions			Open question		
	<i>r</i>	<i>BF</i> ₁₀		<i>r</i>	<i>BF</i> ₁₀		<i>r</i>	<i>BF</i> ₁₀	
Maximum gaze coupling	.31	1.05	U	.08	0.23	E	.54	54.26	I
Average proportion fixation time on relevant information	.21	0.42	U	.06	0.22	E	.48	13.18	I

Note. E denotes factors for which the $BF_{inclusion} < 1/3$ (substantial evidence for exclusion of this predictor), I denotes factors for which the $BF_{inclusion} > 3$ (substantial evidence for inclusion of this predictor), and U denotes factors with $1.3 > BF_{inclusion} > 3$ (uncertainty regarding inclusion or exclusion of this predictor).

the baseline gaze coupling ($M = .14$, $SD = .01$), $BF_{10} > 1000$. The baseline gaze coupling (reflecting a random gaze order of the participant) did not correlate with the score on the open question ($BF_{10} = 0.36$), the score on the true/false questions ($BF_{10} = 0.21$), and the “did the speaker say...” questions ($BF_{10} = 0.26$).

The Bayesian correlation analyses provide strong support for a correlation between the score on the open question and the two gaze measures. For the true/false questions, there is substantial support for a lack of correlation between the score and the two gaze measures. For the “did they speaker say...” questions, there is uncertainty regarding whether or not there is a correlation between the score and the two gaze measures. Scatterplots that show the correlations between the posttest scores and the two gaze measures can be found in Fig. 4. As can be seen, the patterns of the correlations are very similar between the gaze two measures. Indeed, the maximum gaze coupling correlated strongly with the average proportion fixation time on relevant information, $r = .88$, $BF_{10} > 1000$. Additionally, we checked whether the

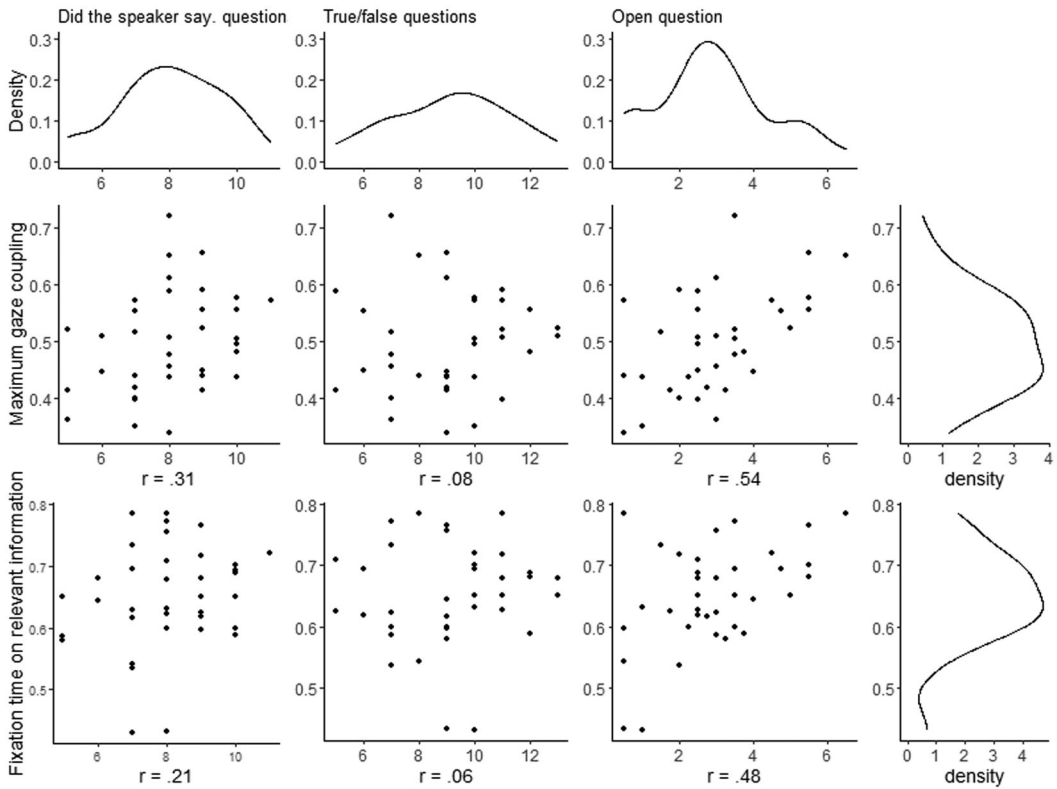


Fig. 4. Scatter plots and density plots of the two gaze measures with the three posttest scores.

Note. The correlation between maximum gaze coupling and fixation time on relevant information was high, $r = .88$.

optimal lag predicts the posttest scores. The Bayes factor expresses uncertainty regarding the correlation between the score on the open question and the optimal lag, $r = .22$, $BF_{10} = 0.46$. For the two other tests, there was substantial evidence that there is no correlation: For the “did the speaker say...” questions, $r = -.002$, $BF_{10} = 0.21$. For the true/false questions, $r = .11$, $BF_{10} = 0.25$.

2.3. Discussion

Study 1 aimed to establish a correlation between gaze measures of with-me-ness and posttest scores. As expected, we found strong correlations between the average proportion of fixation time on relevant information, gaze coupling, and the score on the open question. We thus replicated earlier findings that looking more at relevant information, and having higher gaze coupling with the teacher (both indications of with-me-ness) correlate with higher posttest performance. Richardson and Dale (2005) found a correlation between the score on the Did the speaker say... questions and maximum gaze coupling. In our sample, there was

uncertainty regarding the strength of the correlation. Furthermore, we did not find a correlation with gaze measures for true/false questions.

Study 1 shows that gaze measures indeed hold information that relates to posttest scores (i.e., the open question, but not the multiple-choice measures). It could be argued that the two multiple-choice measures are measures of recall, and the open question is a measure of comprehension. Recall is arguably a necessary but not a sufficient condition for comprehension, and predicting comprehension of a video lecture seems more important than predicting recall. Thus, the findings from Study 1 suggest that the first prerequisite for being able to use gaze measures to predict comprehension was met. With this prerequisite covered, we can now investigate whether observers can use visualizations of these data to predict students' posttest performance.

3. Study 2

Study 2 aimed to investigate whether laypeople can predict students' posttest performance (score on the open question) from gaze visualizations and whether different visualization techniques impact performance predictions (i.e., perceived ease and speed).

To investigate this, participants' scores on the open questions and their gaze measures were binned into three equally sized bins (below average, average, and above average). Next, we selected stimuli from participants for whom gaze measures and posttest performance match (i.e., are in the same bin), and learners for whom there is a mismatch (see Fig. 5). Two groups of mismatches were selected: learners for whom the bin of the gaze measures was higher than that of the score (underestimation), and those for whom the bin of the gaze measures was lower than that of the score (overestimation).

We then used the Bayesian informative hypotheses approach (Hojtink, Mulder, van Lissa, & Gu, 2019). In this approach, meaningful hypotheses were specified that each have a theoretical interpretation. For each hypothesis, the evidence for this hypothesis given the data is calculated (i.e., how likely is the data to be observed under the different hypotheses). Finally, Bayes factors can be used to quantify how much more likely the preferred hypothesis is than the other hypotheses given the data. We consider four hypotheses about participants' predictions of posttest performance based on gaze visualizations and quantify the evidence in favor of each hypothesis to answer the research question.

H₁: If performance is on chance level for mismatch stimuli and above chance level for match stimuli, participants base their estimate of posttest performance on the two gaze measures that we found to relate to learning (average proportion of fixation time on relevant information and gaze coupling).

H₂: If performance is above chance level for mismatch stimuli but still lower than performance for match stimuli, participants base their estimate of posttest performance on the two gaze measures and other information.

H₃: If performance is above chance level for mismatch stimuli and performance is similar to that of match stimuli, participants base their estimate of posttest performance on other information.

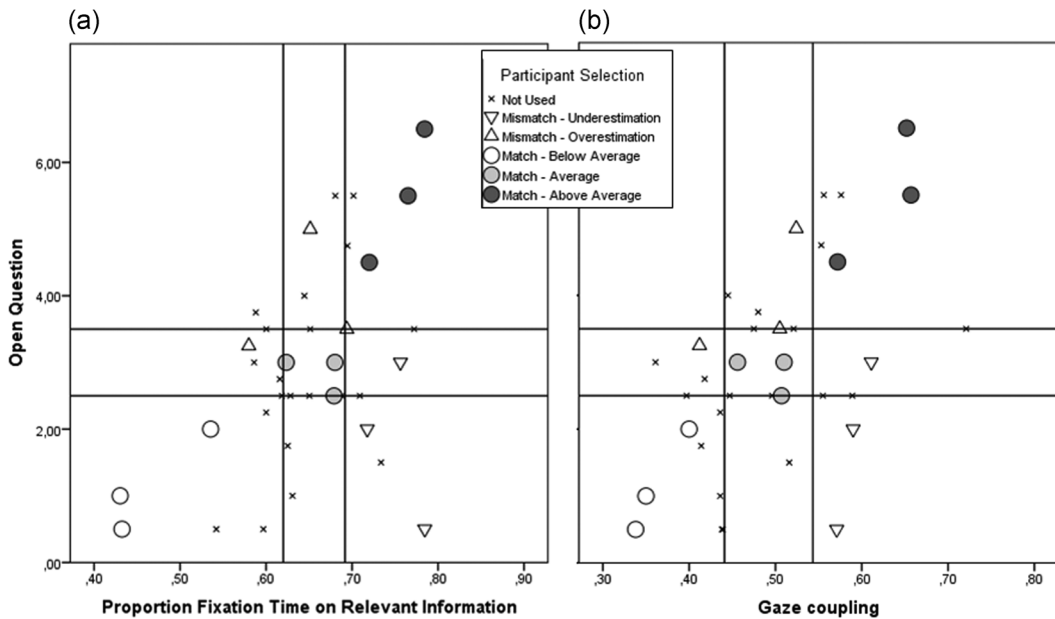


Fig. 5. Scatterplots of proportion fixation time on relevant information (a) and gaze coupling (b) with the score on the open question and participants selected.

Note. Lines denote the cut-off values for below average, average, and above average. Circles and triangles denote participants included in Study 2. Circles denote participants for whom the gaze-measure bin and open-question bin match, triangles denote participants for whom there is a mismatch.

H₄: If performance is on chance level for both match and mismatch stimuli, participants cannot interpret gaze visualizations in terms of posttest performance.

To investigate the effect of visualization techniques on interpretation accuracy, time, and preference, we compared differences between heatmaps, focus maps, and scanpaths, and between static and dynamic versions of those using Bayesian ANOVAs.

3.1. Methods

3.1.1. Participants and design

Fifty participants (27 male) were recruited via Prolific (<https://www.prolific.co/>). Their mean age was 29.5 years (range 19–60). Participants were selected to have at least a high school diploma/A-levels and reported having a high-school diploma or community college diploma ($n = 6$), undergraduate degree ($n = 24$), or graduate/doctorate degree ($n = 20$). Participants reported being fluent in English, 32 reported English as their mother tongue, and 18 reported speaking English as a second language. The country of residence was the United Kingdom ($n = 26$), non-English speaking European countries ($n = 18$), the United States of America ($n = 5$), and South Africa ($n = 1$). Most participants ($n = 40$) reported no teaching experience, with the other 10 working as teachers or giving lectures as part of another job.

Most participants ($n = 43$) did not have a background in medicine, and participants worked or studied in the health professions. No participants reported having experience with the use of eye tracking. One additional participant was originally enrolled but did not complete the study and was thus excluded. Three validity checks were done before participants were included: First of all, it was checked whether there was no obvious response pattern in the multiple-choice questions (e.g., always the same answer). Second, six open questions were included, and participants were only included if those answers were sensible. Third, response times per stimulus and per block were checked to see if those were sensible. None of the participants had to be excluded based on these quality checks. All participants provided written informed consent and were paid £5.00 for their participation. This study was approved by the Research Ethics Committee of the first author's institute.

An a-priori power analysis in G*power (Faul et al., 2007) showed that in order to find a medium-sized effect ($d_z = 0.5$) for paired t -tests (assuming null-hypothesis significance testing, as originally planned) with a power of 0.8 and alpha of .05, at least 34 participants should be included. Power analyses for the Bayesian informative hypothesis approach are not straightforward, and an estimate for the number of necessary participants was made in consultation with a statistician.

3.1.2. Materials

Data were collected using Gorilla (<http://www.gorilla.sc>), (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020). Only participants using a PC or laptop could access the study, and the study looked identical in all allowed browsers (Chrome, Safari, Edge, and Microsoft Internet Explorer).

Gaze visualizations To investigate whether laypeople can interpret gaze visualizations in terms of learners' posttest performance, we selected data from Study 1 to develop a varied stimulus set in which the relation between gaze data and posttest performance is representative of that in Study 1. Selection of the stimuli for Study 2 from data of Study 1 took place in two steps: 1. Selection of learners (participants from Study 1). 2. Selection of representative fragments.

Selection of learners Learners' total score on the open question, proportion fixation time on relevant information, and maximum gaze coupling were recoded into three equally sized bins: below average, average, and above average. The average proportion fixation time on relevant information and maximum gaze coupling were highly correlated, so only learners for whom the two gaze measures (proportion fixation time on relevant information and maximum gaze coupling) were in the same bin ($n = 26$) were selected and further subdivided into two groups: matches and mismatches (see Fig. 5). Three matching groups were selected: those with below-average gaze and open question scores, those with average gaze and open question scores, and those with above-average gaze and open question scores. Finally, two groups of mismatches were selected: learners for whom the bin of the gaze measures was higher than that of the score (underestimation), and those for whom the bin of the gaze measures was

lower than that of the score (overestimation). For each of the five groups, three learners were selected for whom a representative fragment could be selected (see next section).

Fragment selection For each learner, an 8–10 s fragment (cf. Van Wermeskerken et al., 2018) was selected that had the same average proportion of fixation duration on relevant information ($M = .66$, $SD = .12$) as the proportion for the whole video ($M = .66$, $SD = .12$), which was similar to that of the whole sample of Study 1 ($M = .65$, $SD = .08$). This ensured that the fragment was a good representation of the whole video, and as such, allowed us to ask observers to estimate performance on the whole video based on a single fragment. Those fragments were selected from roughly the same timespan so the relevant AOI was the same throughout the fragment and the same for all learners.

Gaze visualization development For each learner and the expert, six gaze visualizations were developed using SMI BeGaze 3.7: a static and a dynamic version of a heatmap, scanpath, and focus map (see Fig. 1). For the scanpaths, the Scanpath utility of BeGaze was used. Participants' scanpaths were shown in dark green, and those of the expert in blue. Circle size was reflective of the duration of the fixation, with 0.5 degrees representing 500 ms. All other settings were kept to default. For heatmaps, the accumulated time that a participant looked at the different areas of the stimulus was visualized. A color-coding from blue (lowest fixation duration) to red (highest fixation duration) was used, the data range was 0–103 ms, kernel width 2 degrees, and opacity of the heatmap was 75%. Focus maps also showed the accumulated time that a participant looked at the different areas of the stimulus. The opacity of the dark areas was 85%. A data range of 0–103 ms was used and the kernel width was 2 degrees. For the videos, the same settings were used, and videos were exported at 10 fps.

Participants would see the heatmap, scanpath, or focus map stimulus of the same learner in both the static and the dynamic versions, the assignment of learners to visualization type was counterbalanced across participants. Stimuli were blocked according to movement type (static or dynamic). Within movement type, stimuli were blocked according to visualization type (scanpath, focus map, heatmap, and order randomized), and the order of stimuli within a block was randomized. Each block consisted of five gaze visualizations: One learner from each of the five match types. Thus, in total, six blocks were presented, with five visualizations per block. Static displays were presented in a self-paced manner. Dynamic displays were played once and after that, participants could replay it two more times or provide an answer immediately.

Instructions Participants received the following instruction regarding the full session: “In this experiment, you will be asked to watch gaze visualizations. Gaze visualizations show where a learner has looked while watching an instruction video. During this time, they saw a single image and heard the voice of an instructor. The instructor explained how to recognize a fast heartbeat on an ECG (electrocardiogram). In the picture below, you can see the image that those participants saw. The gaze visualizations that you will see show the gaze (i.e., viewing location) of a person during 8–10 seconds. The black box shows which part of the picture the instructor was talking about during those fragments. The visualizations that you will see

are representative of the participant's looking behavior over the whole video. After watching the instruction video, participants were asked to explain the content of the movie in their own words. You will be asked to estimate how well each participant did this (below average, average, or above average) and report how confident you are in your answer. You will see different types of gaze visualizations. Before each block of visualizations, you will receive information about the gaze visualization that you will see. After the experiment is finished, you will be asked which of the visualizations you preferred."

Before each visualization type, a short instruction with an example was presented. This instruction only featured information about how the visualization represents gaze. For a scan-path, participants were informed that: The circles are locations where a learner looks, larger circles mean longer looking times, and the lines represent jumps between those locations. For focus maps: Light areas are places that the person looked at, and dark areas were not looked at. For heatmaps: The heatmaps show where and how long a person looked somewhere, warmer colors denote longer looking times. No information was given about how longer looking times relate to posttest performance.

Rating scales

Performance: After each gaze visualization, participants were required to judge the performance of the participant on the posttest (below average, average, or above average). Participants' performance was the proportion of correct judgments.

Response time: The response time was the time until the performance judgment was made. For dynamic visualizations, this was recorded from the time after the video was played once. For static visualizations, it was recorded from the time at which the image was first presented.

Confidence: After each performance judgment, participants rated their certainty in their estimation on a five-point scale ranging from 1 (not certain at all) to 5 (very certain).

Mental effort: After each block of gaze visualizations, participants were asked to report how much mental effort they invested to estimate the posttest performance. Participants rated their invested mental effort on a Likert scale from 1 (very, very low mental effort) to 9 (very, very high mental effort) (cf. Paas, 1992).

Preference: After working through all blocks, participants were presented with thumbnails of all six visualization types and asked to rate how useful they found each visualization type for estimating task performance, on a nine-point scale (1–9).

Cue utilization: After each block, participants were asked to report what information they had used to interpret the gaze visualization (open question). We coded those open answers to investigate whether participants indeed used the similarity between teacher and student, and attention to relevant information as cues to interpret the gaze displays. Additionally, we used open coding to categorize any other cues that were mentioned and

developed a code book. Based on the codebook, two researchers individually coded 50% of the data. Krippendorff's alpha (Hayes & Krippendorff, 2007) was 0.716, which can be considered acceptable. The other half of the data was coded by one of the researchers. The final codebook can be found in Table 3.

3.1.3. Procedure

Participants first received a short demographics questionnaire and received instructions about the task. Next, they were presented with the six blocks of five trials. Each trial consisted of a gaze visualization with the question to indicate the participant's posttest performance. Subsequently, participants reported their certainty in that estimate. After each block of gaze visualizations, participants were asked to report what information they used to interpret the gaze visualizations and how much mental effort they invested to estimate the posttest performance. They were also given the opportunity to take some rest before continuing to the next block. After working through all blocks, participants reported their preference. Finally, they were thanked for their participation.

3.1.4. Measures and analyses

To analyze whether participants can interpret gaze visualizations in terms of learners' posttest performance and which information they use to do so, the Bayesian informative hypotheses approach was used (Hojtink et al., 2019) in R using the *bain* package. Four informative hypotheses were formulated and the Bayes factors for those hypotheses were used to quantify the evidence for each of them (see Table 4). The dependent variable for this analysis is judgment accuracy, which was operationalized as the proportion of gaze visualizations for which the correct bin was selected.

We assume that when participants cannot use any information from the gaze visualizations, they would perform at a chance level (0.33), see H_4 . When they do use information from the gaze displays, however, there are two options. First of all, they can use the information that we found to relate to performance, that is, the gaze coupling and proportion of fixation time on relevant information, to predict the learners' posttest performance. Second, they can use other information from the gaze displays. Therefore, we distinguish between match and mismatch displays. For match displays, the gaze coupling and proportion fixation time on relevant information are related to performance, and for mismatch displays, they are not. Thus, if this information is the sole source of the prediction, it would be expected that performance is above chance for the match but not the mismatch displays (H_1). On the other hand, if other information is used than the information that we used to distinguish between match and mismatch displays, we would not expect a difference in accuracy between match and mismatch conditions, and we would expect performance in both conditions to be higher than chance level (H_3). Finally, if both gaze coupling and proportion fixation time on relevant information, and other information are used, we would expect both conditions to have an accuracy above chance level, but the match condition would score higher than the mismatch condition (H_2), since the gaze coupling and proportion fixation time information are only predictive of performance in the match but not the mismatch condition.

Table 3
Codebook for coding cue utilization

Code	Description	Example
Comparison with teacher	If the participant mentions a comparison between teacher's gaze and student gaze (e.g., similarity, similar, compare, and the same). Also, if the participant mentions that the student gazes at the information/area/part that the teacher is looking at.	"The similarities in both the lecturer and learner's focus maps"
Relevant information	If the participant mentions using whether relevant/important information (or: the ECG, fast heartbeat, trace, and graph) was looked at or not. Also: if the participant refers to the relevant information as "the part/area that the teacher was talking about, pointing to."	"Observed how long the student's look where focusing at the information needed"
Both comparison and relevant information	If the participants' answer could both be coded as "comparison with teacher" and "relevant information."	"How long they spent on each relevant area, how well their map matched that of the teacher"
Coverage	If the participant mentions using how much of the image is looked at.	"How much white coverage there was on the screen"
General remark about duration	If the participant mentions looking at the duration of gazes without specifying that this is about relevant/irrelevant or similar/dissimilar to teacher.	"How long they focused on certain areas"
General remark about location	If participants mention looking at the location of gaze data without specifying that this is about relevant/irrelevant or similar/dissimilar to teacher.	"I tracked the scanpaths to see where the students had been focusing."
General remark about duration and location	If general remarks about both duration and location are made.	"Took into consideration the amount of time spent looking at particular parts of the graph"
Other characteristic of eye-tracking data	If the participant mentions another characteristic of the eye-tracking data.	"The number of big circles"
General remark about figure	If the participant mentions using the figure (e.g., heatmap and graph) without further specifying the cue.	"The heatmaps and the darker areas of them."
Other/cannot be coded	For remarks that cannot be coded in other categories.	"Different gazes"

Table 4
 Overview of the four informative hypotheses and their theoretical interpretation

Hypothesis	Theoretical interpretation
$H_1: \mu_{\text{mismatch}} = .33 \ \& \ \mu_{\text{match}} > .33$	Participants base their estimate of posttest performance on gaze coupling and proportion of fixation time on relevant information.
$H_2: \mu_{\text{mismatch}} > .33 \ \& \ \mu_{\text{match}} > .33 \ \& \ \mu_{\text{match}} > \mu_{\text{mismatch}}$	Participants base their estimate of posttest performance on gaze coupling and proportion of fixation time on relevant information and on other information.
$H_3: \mu_{\text{mismatch}} > .33 \ \& \ \mu_{\text{mismatch}} = \mu_{\text{match}}$	Participants base their estimate of posttest performance on other information.
$H_4: \mu_{\text{match}} = .33 \ \& \ \mu_{\text{mismatch}} = .33$	Participants use no information/cannot interpret gaze visualizations in terms of posttest performance.

Differences between gaze visualization types were explored using 2×3 Bayesian ANOVAs (van den Bergh et al., 2020) executed in JASP (JASP Team, 2021), with movement type (static vs. dynamic) and visualization type (focus map, heatmap, or scanpath) as factors and average accuracy, time, confidence, mental effort, and preference as dependent variables. We used default settings. We report the Inclusion Bayes factors ($BF_{inclusion}$) for the two factors and the interaction effect. The Inclusion Bayes factors can be interpreted as evidence in the data for including the predictor (van den Bergh et al., 2020). Higher Bayes factors reflect stronger evidence for including the predictor. A $BF_{inclusion}$ of three, for example, means that the data are about three times more likely under models that include this predictor than under models without the predictor. If the $BF_{inclusion}$ is smaller than 1, smaller Bayes factors reflect stronger evidence for excluding the predictor. The $BF_{exclusion} = 1/BF_{inclusion}$, so a $BF_{inclusion}$ of $1/3$, for example, means that the data are about three times more likely under models that exclude this predictor than under models that include the predictor. We flag predictors with $BF_{inclusion}$ higher than 3 or lower than $1/3$, but invite the reader to interpret the Bayes factors in a continuous manner.

3.2. Results

3.2.1. Can laypeople predict students' comprehension from gaze visualizations?

Table 5 provides an overview of the number of times an answer was selected for each of the match types and stimulus types. Match items were those for which the gaze measure and the posttest performance were in the same bin (both were either below average, average, or above average), and mismatch items were those for which the gaze measure and the posttest performance were not in the same bin. There were two types of mismatches: learners for whom the bin of the gaze measures was higher than that of the score (underestimation), and those for whom the bin of the gaze measures was lower than that of the score (overestimation).

The table shows that for match items, the answer that was selected by most participants was often the correct answer. This was true if below average or average was the correct answer, but not if the above average was the correct answer. For mismatch items, the answer that was selected by most participants was often not the correct answer. Note that all the answer options (below average, average, and above average) are equally likely to be the correct answer, so the slightly different response patterns to match and mismatch items are unlikely to explain the differences between match and mismatch trials. Table 6 shows that the average accuracy is indeed above chance level for match but not for mismatch items. Fig. 6 provides violin plots for match and mismatch items.

The posterior probability of each hypothesis quantifies the support for that hypothesis given the data (see Table 7), whereas the Bayes factor BF_{0u} can be used to compare hypotheses, because it quantifies how much more likely the data are to be observed under H_0 versus H_u (Hoijtink et al., 2019). For static visualizations, the posterior probability of H_2 is highest, and this hypothesis is 1.6 times as likely as H_1 given these data ($BF_{21} = 1.59$), almost four times as likely as H_3 ($BF_{23} = 3.89$), and 2183 times as likely as H_4 ($BF_{24} = 2182.88$). The Bayes factor of 1.59 can be interpreted as some uncertainty as to whether H_1 or H_2 is the best description of the data (Hoijtink et al., 2019). However, both hypotheses state that participants can interpret

Table 5
Number of selected answers by movement type for each of the match types

Selected answer	Match type and participant score in Study 1								Total	
	Match		Mismatch—underestimation		Mismatch—overestimation		Mismatch—overestimation			
	below average	above average	below average	above average	below average	above average	below average	above average		
Static										
Below average	86*	48	48	38*	11	24	26	281		
Average	51	75*	72	46	23*	23*	49	339		
Above average	13	27	30*	16	16	3	25*	130		
Dynamic										
Below average	89*	29	40	36*	18	43	37	292		
Average	44	85*	78	50	16*	7*	49	329		
Above average	17	36	32*	14	16	0	14*	129		
Total	300	300	300	200	100	100	200			

Note. *correct responses. Note that all the answer options (below average, average, and above average) are equally likely to be the correct answer.

Table 6
Average proportion correct for the match and mismatch trials

	Accuracy			
	Static		Dynamic	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Match	0.42	0.16	0.46	0.13
Mismatch	0.36	0.19	0.24	0.16

Note. Chance level is 0.33.

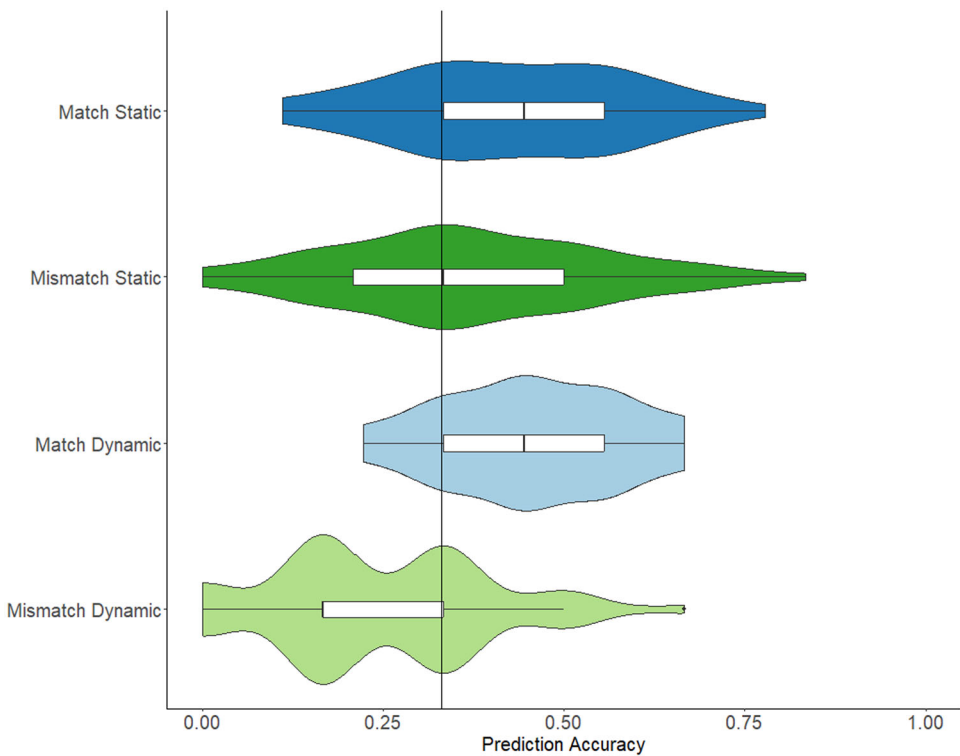


Fig. 6. Violin plots with boxplots for the match and mismatch items, for static and dynamic stimuli.
Note. The line denotes chance level (0.33).

gaze visualizations, so the data provide ample support that participants can indeed interpret gaze visualizations in terms of posttest performance.

For dynamic visualizations, the posterior probability of H_1 is highest, and this hypothesis is 14 times as likely as H_2 , ($BF_{12} = 13.98$), and much more likely than H_3 and H_4 (both BF 's > 1000) given the data. Thus, for dynamic visualizations, the pattern discussed in H_1 is very

Table 7
Posterior probabilities for each informative hypothesis

Meaningful hypotheses	Static visualizations	Dynamic visualizations
$H_1: \mu_{\text{mismatch}} = .33 \ \& \ \mu_{\text{match}} > 33\%$	0.33	0.93
$H_2: \mu_{\text{mismatch}} > .33 \ \& \ \mu_{\text{match}} > .33 \ \& \ \mu_{\text{match}} > \mu_{\text{mismatch}}$	0.53	0.07
$H_3: \mu_{\text{mismatch}} > .33 \ \& \ \mu_{\text{mismatch}} = \mu_{\text{match}}$	0.14	0.00
$H_4: \mu_{\text{match}} = .33 \ \& \ \mu_{\text{mismatch}} = .33$	0.00	0.00

Note. The posterior probabilities include the posterior probabilities of only the four meaningful hypotheses, ignoring the hypothesis that means are unconstrained ($H_u = \mu_{\text{mismatch}}, \mu_{\text{match}}$).

likely. This means that for dynamic visualizations, people use the information we provided, and probably no other information.

3.2.2. How do different visualization techniques impact performance predictions?

Table 8 shows descriptives for the five dependent variables by visualization type. Absolute accuracy, response time, and confidence are averaged for the match items only. The effort was expressed after a full block and preference after the full study.

Differences between gaze visualization types were explored using 2×3 Bayesian ANOVAs with movement type (static vs. dynamic) and visualization type (focus map, heatmap, or scanpath) as factors and average accuracy, time, confidence, mental effort, and preference as dependent variables. Table 9 reports the Inclusion Bayes factors ($BF_{\text{inclusion}}$) for the two factors and the interaction effect.

Response times were longer for static visualizations because for dynamic visualizations the first time playing the visualization was not included, so the main effect of movement type should not be interpreted. There was very strong evidence for including movement type to predict effort: Participants invested more mental effort in interpreting dynamic gaze visualizations than static gaze visualizations. There was substantial evidence for including movement type to predict confidence: Participants felt more confidence in their answers for static versus dynamic displays. Apart from that, there was either uncertainty regarding the predictive value of the factors, or substantial evidence that the factor or interaction did not predict the dependent variable. Thus, there are only minor differences between the displays.

As an exploratory analysis, we report which cues were used by participants to interpret the gaze visualizations. Frequencies of cues use can be found in Table 10. For all visualization types, the majority of the cues reported were the ones that we provided. Note that for the cues “duration” and “location,” participants did not explicitly report which areas correspond to high versus low performance. However, it seems likely that they mean that looking at relevant locations is related with higher performance and looking at irrelevant locations is related with lower performance, so those participants might have still used the information that we provided. Likewise, the answer “the learners that looked everywhere for a short period of time, probably didn’t fully understand” was coded as “coverage,” but using this cue still indicates an understanding that looking at irrelevant information (i.e., everywhere, without distinguishing between relevant and irrelevant information) is related to low performance.

Table 8
Average absolute accuracy, response time, confidence, effort, and preference by visualization type for match items only

	Absolute accuracy		Time (s)		Confidence		Effort		Preference	
	M	SD	M	SD	M	SD	M	SD	M	SD
<i>Static</i>										
Focus map	0.37	0.29	5.61	4.36	3.69	0.89	6.00	2.05	6.28	2.01
Heatmap	0.46	0.24	7.17	5.90	3.54	0.81	6.06	1.80	6.58	1.67
Scanpath	0.45	0.25	7.89	6.33	3.60	0.86	6.00	1.99	6.02	1.92
<i>Dynamic</i>										
Focus map	0.42	0.22	3.33	2.95	3.55	0.91	6.36	1.77	6.46	2.15
Heatmap	0.51	0.26	4.13	4.74	3.45	0.83	6.56	1.52	6.78	1.96
Scanpath	0.45	0.26	5.15	8.00	3.57	0.81	6.68	1.67	6.90	2.00

Table 9
 Inclusion Bayes factors for the main effect of movement type, the main effect of visualization, and the interaction effect for average absolute accuracy, response time, confidence, effort, and preference

	Absolute accuracy		Time (s)		Confidence		Effort		Preference	
	$BF_{inclusion}$		$BF_{inclusion}$		$BF_{inclusion}$		$BF_{inclusion}$		$BF_{inclusion}$	
Movement type	0.17	E	> 1000	I	3.30	I	> 1000	I	0.72	U
Visualization type	0.61	U	2.54	U	0.30	E	0.09	E	0.06	E
Interaction	0.04	E	0.22	E	0.10	E	0.05	E	0.03	E

Note. E denotes factors for which the $BF_{inclusion} < 1/3$ (substantial evidence for exclusion of this predictor), I denotes factors for which the $BF_{inclusion} > 3$ (substantial evidence for inclusion of this predictor), and U denotes factors with $1/3 > BF_{inclusion} > 3$ (uncertainty regarding inclusion or exclusion of this predictor).

Table 10
 Frequency of reported cue use by movement type and visualization type

	Focus map				Heatmap				Scanpath			
	Dynamic		Static		Dynamic		Static		Dynamic		Static	
Comparison with teacher	24	48%	27	54%	24	48%	27	54%	20	40%	27	54%
Remark about relevant information	5	10%	7	14%	5	10%	1	2%	2	4%	1	2%
Both comparison and relevant information	4	8%	2	4%	2	4%	3	6%	5	10%	4	8%
Total teacher and relevant information	33	66%	36	72%	31	62%	31	62%	27	54%	32	64%
Coverage	1	2%	1	2%	1	2%	0	0%	1	2%	0	0%
General remark about duration	1	2%	1	2%	0	0%	2	4%	4	8%	2	4%
General remark about location	5	10%	6	12%	7	14%	10	20%	5	10%	6	12%
General remark about duration and location	3	6%	0	0%	6	12%	1	2%	5	10%	2	4%
Other characteristic of eye-tracking data	0	0%	0	0%	0	0%	0	0%	2	4%	3	6%
General remark about figure	3	6%	3	6%	2	4%	4	8%	4	8%	3	6%
Other	4	8%	3	6%	3	6%	2	4%	2	4%	2	4%

3.3. Discussion

In Study 2, we investigated whether people use the information displayed in gaze visualizations to predict learners' posttest performance. Indeed, we found that laypeople can use the information displayed in gaze visualizations to predict learners' posttest performance, even though they had not received any instructions on how gaze data relate to posttest performance. We found only minor differences between visualization techniques.

4. General discussion

Visualization of students' gaze data might inform teachers about their level of comprehension of online lectures. In the current study, we investigated and found support for three prerequisites of this idea: (1) Gaze measures of with-me-ness are related to posttest performance, (2) Laypeople can predict students' posttest performance from their gaze visualizations, (3) We understand whether and how different gaze visualization techniques impact this prediction.

4.1. Are gaze measures related to comprehension in instruction videos?

An important prerequisite for using gaze visualizations is that there is a correlation between gaze measures and posttest performance, because otherwise there is no information in the gaze data for teachers to base their assessment of student comprehension on. As expected, we found strong correlations between gaze measures of with-me-ness (proportion fixation duration on relevant information and gaze coupling) and the score on the open question (which reflects comprehension) about video lectures, which means there is indeed a pattern for participants to pick up. The correlations were relatively high ($r = .54$ and $r = .48$), and even slightly higher than the Spearman correlation of 0.36 that Sharma and colleagues found (Sharma et al., 2014). Investigating this prerequisite also made it possible to investigate which information was used by participants, and indeed participants mostly used the measures that we found to relate to posttest performance.

The proportion fixation duration on relevant information and the measure of gaze coupling were highly correlated ($r = .88$). The measure of gaze coupling required that the teacher's eyes were tracked during the development of the lecture, and thus has practical problems. The data show that the simpler measure of proportion fixation duration on relevant information may suffice in the current situation. Note though, that this measure is not that "simple" to compute in dynamic stimuli either, because what is relevant changes from moment to moment in the video, as it is associated with what the teacher is talking about. However, because what teachers are talking about is typically what they are focusing on in these materials, this measure will come close to the gaze coupling measure, yet without requiring the teacher being eye tracked.

Whereas we found strong correlations between the gaze measures and the score on the open question, we did not find strong correlations between the gaze measures and the score on the true/false questions, and inconclusive evidence for the "did the speaker say..." questions.

Power might have been too small to detect correlations with those measures. However, an alternative explanation lies in the content: Both sets of multiple-choice questions might have relied more on the simple recall of verbal information; for the comprehension questions, taking in the right visual information at the right time might have been more important to attain understanding.

4.2. *Can laypeople interpret gaze visualizations in terms of learners' posttest performance?*

Given that there is indeed a correlation between gaze measures and comprehension, the next prerequisite is that people with limited eye-tracking experience (i.e., laypeople) are able to interpret the visualization in terms of learners' posttest performance. Previous research had already shown that participants can interpret gaze visualizations in terms of perceptual or cognitive processes (Bahle et al., 2017; Emhardt et al., 2020; Foulsham & Lock, 2015; Haider & Frensch, 1996; Van Wermeskerken et al., 2018; Zelinsky et al., 2013). The current study adds evidence that is highly relevant for education, showing that laypeople could use the information from gaze visualizations to predict students' posttest performance of a lecture video, even without instructions on how gaze data relate to posttest performance. When asked which information was used to predict posttest performance, the majority of participants reported using the eye-tracking measures that we found to correlate with posttest performance.

In our study and in the other studies discussed earlier, participants' performance on gaze interpretation tasks is often above chance level yet far from perfect. Two things can explain the imperfect prediction of performance based on gaze visualizations. Either the gaze visualizations do not contain predictive information, or the participants do not possess the ability to extract and interpret the information. Greene's work (2012) showed an example of the first situation. In their study, participants failed to interpret visualizations of viewing behavior in terms of the viewer's task. A pattern classifier that used measures from the scanpaths could also not predict viewers' tasks. Borji and Itti (2014) argued that this was caused by very similar gaze patterns between the different tasks, and showed that only advanced classifiers can classify those visualizations slightly above chance levels. An inability to interpret gaze visualizations can thus be caused by a lack of information in the visualizations, but the differences between gaze visualizations can also be too subtle for participants to pick up. In our study, the correlations between gaze measures and posttest performance were high ($r = .54$ and $r = .48$), and the explained variance (R^2) in the posttest performance was, respectively, 29% and 23%, leaving at least 71% of the variance in performance unexplained, and thus this variance could not be predicted based on the gaze visualizations.

Regarding the ability of participants to interpret gaze visualizations, it is important to note that in most of the existing literature, participants are not told how visual information relates to the process they have to predict. All studies (including the current studies), thus investigate how much information participants pick up intuitively, which might be considered a lower boundary for interpretation performance. For example, in the study by Zelinsky et al. (2013), participants were told that they were presented with information about the longest-fixated distractor and the first-fixated distractor and used both in their inference. However, only focusing on the longest-fixated distractor would have led to improved performance over using both. We

asked participants to report which information they used to predict performance. Many participants in our study reported using the information that we found to be related to posttest performance. However, participants also reported other cues such as the distance between subsequent fixations (i.e., jumping quickly over a larger distance relates to a lower posttest score) or more fuzzy cues such as randomness of the viewing pattern. It is thus very relevant to investigate to what extent participants do better if they receive information about which cues they could extract from the visualizations and how these cues relate to posttest performance. Indeed, many of the participants remarked that this information was missing and that they felt they were only guessing as to what the displayed gaze means. Furthermore, it is yet unknown how domain knowledge (i.e., about ECGs rather than eye tracking) and teaching experience impact interpretation performance, and those are important avenues for further research.

4.3. *How do different visualization techniques impact performance predictions?*

Finally, given that laypeople can predict students' posttest performance from their gaze visualizations, it is important to understand whether and how different gaze visualization techniques impact the prediction of students' posttest performance from their gaze visualizations, so that optimal gaze visualization techniques can be used. We found only minor differences between visualization techniques in how they were interpreted. Participants invested somewhat more effort in dynamic than in static visualizations, and were somewhat more confident in their answers regarding static versus dynamic visualizations. Overall, however, differences in self-reported interpretation ease and preference and actual differences in performance and speed were small. This seems to contrast to the literature stating that different visualization techniques have different affordances for interpretations (Blascheck et al., 2014; Kurzhals et al., 2015).

Bahle, Mills & Dodd (2017) found the effects of visualization techniques on interpretation performance, and Van Wermeskerken et al. (2018) found that dynamic visualizations were easier to interpret in terms of the viewer's instruction than static visualizations. In both studies, participants had to judge viewer's viewing behavior in terms of a process (i.e., what instruction did the viewer get that resulted in this viewing behavior?). In our study, the participant had to predict students' later performance based on the viewing behavior of a lecture. It seems that for this prediction, visualization techniques do not differ much in their affordances for interpretation. However, further research (on lectures and other tasks) would be required to understand under which circumstances visualization techniques do or do not impact interpretation performance.

4.4. *Limitations*

This study extended prior research on gaze interpretation to an educationally relevant context, that is, learning from multimedia video lectures. This is common in online and blended education and students' gaze data could potentially have added value in this context, as teachers typically lack information that allows them to monitor students' comprehension of video lectures. A limitation of the present study, however, is that by focusing on video lectures, and on this particular lecture, it remains unclear how task-specific or content-specific

our findings are. The relationship between gaze and comprehension or posttest performance might depend on the specific task or content (prerequisite 1), and peoples' ability to interpret gaze visualizations might depend on the particulars of a task (prerequisite 2). It is, therefore, important to investigate those prerequisites for different videos and other educational situations. For example, in video lectures, the speaker is sometimes visible and sometimes not. Because faces attract attention (e.g., Cerf, Frady, & Koch, 2009), the presence of a teacher means that less attention is paid to the lecture slides and that participants are slower to look at the information that the teacher verbally refers to (van Wermeskerken, Ravensbergen, & van Gog, 2018). However, with their learning task (solving probability calculation problems), van Wermeskerken et al. did not find a significant difference in posttest performance between conditions with and without the teacher visible. Yet, this could be due to the fact that remembering the problem-solving steps was most important for learning that task, and the appearance of those steps would draw attention in their videos. Seeing the teacher and, therefore, being slower to look at referenced information might have a different impact with a more conceptual learning task in which the interpretation of visual information is central, like our video about ECG interpretation. Indeed, our data suggest that when allocating attention to the teacher disrupts with-me-ness, it could be detrimental for learning this task. Finally, the effects of teacher visibility might depend on whether an online lecture is live or prerecorded, as shown by De Felice, Vigliocco, and Hamilton (2021). Thus, further research could investigate whether the relation between gaze measures and posttest performance is impacted if a speaker is visible (prerequisite 1), and whether observers of gaze visualizations are aware of how the presence of a speaker impacts the relation between gaze and posttest performance (prerequisite 2).

As another example, the findings regarding the gaze coupling between student and teacher might depend on the cognitive processes of the teacher. In this experiment, the teacher could focus on the presentation. However, in real live, a teacher might worry about their presentation, or otherwise be distracted from the content and might not have a viewing pattern that would be best for the comparison. Another option might be to compare the students' gaze to the best-performing student or the average of students (e.g., Madsen, Júlio, Gucik, Steinberg, & Parra, 2021). Further research could investigate the generalizability of our findings to classroom situations.

Another potential limitation is inherent to the use of the Bayesian informative hypotheses approach. In this approach, informative hypotheses are formulated that each have a theoretical implication, and subsequently, Bayesian analyses quantify support for each of those hypotheses given the collected data (Hojtink et al., 2019). This is a powerful approach to compare support for different sets of hypotheses. However, in our situation, the restrictions formulated by the data are not supported by the data for the dynamic visualizations, and if a hypothesis were included in the analysis that means were unconstrained (i.e., anything can be going on), this hypothesis was highly preferred. However, this hypothesis has no theoretical meaning (it does not provide any information about the pattern of means) and was thus not included in our analyses.

Finally, we manually selected a subset of the data collected in Study 1 for use in Study 2. Our selection process was set up to ensure that the data selected were representative of the full data set, and we found that the proportion of time spent on relevant information was

indeed very similar between the selected fragments and the full data set. Even so, we still had to select data and could not control all potential differences between the selected fragments and the full data.

4.5. Conclusion

Overall, the findings from our two studies support the idea that gaze visualizations can inform teachers about students' comprehension of a video lecture. We investigated three prerequisites for this idea. First, we found evidence that gaze visualizations indeed contain relevant information: The measures of with-me-ness (the proportion fixation duration on relevant information and gaze coupling) were both found to correlate with posttest performance. Second, we found that laypeople indeed use this information to predict posttest performance based on gaze visualizations. Third, we found only minor differences between visualization techniques. Sample sizes in the experiments were somewhat small and replication in larger samples is necessary. However, those findings provide promising input for the idea that gaze visualizations can support teachers in settings where they have limited cues about learners' comprehension, for example, in online lectures. Further research could investigate whether visualizations that summarize gaze of groups of learners could also be interpreted in a similar manner, and if teachers can act on the information (e.g., provide additional explanations or slow down) and thereby help students to learn more efficiently.

Acknowledgments

The authors would like to thank Marja Erisman for help with data collection, Stan Linders en Jan Bouwman for scoring the open questions, and Jos Jaspers for help with data preprocessing. We would like to thank Herbert Hoijsink for his help with the BAIN analyses. This research was funded by an NRO PROO grant (project number 405-17-301).

Funding

This research was funded by an NRO PROO grant (project number 405-17-301).

Additional supporting information

Supplementary Material.

References

Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education*, 125, 413–428. <https://doi.org/10.1016/j.compedu.2018.06.023>

- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. <https://doi.org/10.1101/438242>
- Ayres, P., & Paas, F. (2007). Can the cognitive load approach make instructional animations more effective? *Applied Cognitive Psychology*, *21*(6), 811–820. <https://doi.org/10.1002/acp.1351>
- Bahle, B., Mills, M., & Dodd, M. D. (2017). Human classifier: Observers can deduce task solely from eye movements. *Attention Perception & Psychophysics*, *79*(5), 1415–1425. <https://doi.org/10.3758/s13414-017-1324-7>
- Blascheck, T., Kurzhals, K., Raschke, M., Burch, M., Weiskopf, D., & Ertl, T. (2014). State-of-the-Art of Visualization for Eye Tracking Data. In R. Borgo, R. Maciejewski, & I. Viola (Eds.), *Eurographics Conference on Visualization (EuroVis)*. (pp. 63–82). Eurographics Association <https://doi.org/10.2312/eurovisstar.20141173>
- Bojko, A. (2009). Informative or misleading? Heatmaps deconstructed. In J. A. Jacko (Ed.), *Human-Computer Interaction. New Trends* (pp. 30–39). https://doi.org/10.1007/978-3-642-02574-7_4
- Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision*, *14*(3), 29–29. <https://doi.org/10.1167/14.3.29>
- Cakir, M. P., & Uzunosmanoğlu, S. D. (2014). An Interactional Analysis of Gaze Coordination during Online Collaborative Problem Solving Activities. In J. L. Polman, E. A. Kyza, D. K. O'Neill, I. Tabak, W. R. Penuel, A. S. Jurow, K. O'Connor, T. Lee, & L. D'Amico (Eds.), *International Conference of the Learning Sciences*. (pp. 1112–1116). Boulder, CO: International Society of the Learning Sciences <https://doi.org/10.22318/icls2014.1112>
- Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision*, *9*(12), 10–10. <https://doi.org/10.1167/9.12.10>
- Chisari, L. B., Mockevičiūtė, A., Ruitenburg, S. K., van Vemde, L., Kok, E. M., & van Gog, T. (2020). Effects of prior knowledge and joint attention on learning from eye movement modelling examples. *Journal of Computer Assisted Learning*, *36*(4), 569–579. <https://doi.org/10.1111/jcal.12428>
- Coco, M. I., & Dale, R. (2014). Cross-recurrence quantification analysis of categorical and continuous time series: an R package. *Frontiers in Psychology*, *5*, 510. <https://doi.org/10.3389/fpsyg.2014.00510>
- De Felice, S., Vigliocco, G., & Hamilton, A. F. d. C. (2021). Social interaction is a catalyst for adult human learning in online contexts. *Current biology*, *31*(21), 4853–4859.e4853. <https://doi.org/10.1016/j.cub.2021.08.045>
- de Koning, B. B., & Jarodzka, H. (2017). Attention guidance strategies for supporting learning from dynamic visualizations. In R. Lowe & R. Ploetzner (Eds.), *Learning from Dynamic Visualization: Innovations in Research and Application* (pp. 255–278). Springer International Publishing. https://doi.org/10.1007/978-3-319-56204-9_11
- Donovan, T., Manning, D. J., & Crawford, T. (2008). Performance changes in lung nodule detection following perceptual feedback of eye movements. In *Proc. SPIE 6917, Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment* (Vol. 6917, pp. 9). <https://doi.org/10.1117/12.768503>
- Eder, T. F., Richter, J., Scheiter, K., Keutel, C., Castner, N., Kasneci, E., & Huettig, F. (2020). How to support dental students in reading radiographs: effects of a gaze-based compare-and-contrast intervention. *Advances in Health Sciences Education: Theory and Practice*, *26*, 159–181. <https://doi.org/10.1007/s10459-020-09975-w>
- Emhardt, S. N., van Wermeskerken, M., Scheiter, K., & van Gog, T. (2020). Inferring task performance and confidence from displays of eye movements. *Applied Cognitive Psychology*, *34*(6), 1430–1443. <https://doi.org/10.1002/acp.3721>
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods*, *39*(2), 175–191. <https://doi.org/10.3758/bf03193146>
- Foulsham, T., & Lock, M. (2015). How the eyes tell lies: Social gaze during a preference task. *Cognitive Science*, *39*(7), 1704–1726. <https://doi.org/10.1111/cogs.12211>
- Greene, M. R., Liu, T., & Wolfe, J. M. (2012). Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns. *Vision Research*, *62*, 1–8. <https://doi.org/10.1016/j.visres.2012.03.019>
- Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11*(4), 274–279. <https://doi.org/10.1111/1467-9280.00255>

- Haider, H., & Frensch, P. A. (1996). The role of information reduction in skill acquisition. *Cognitive Psychology*, 30(3), 304–337. <https://doi.org/10.1006/cogp.1996.0009>
- Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1), 77–89. <https://doi.org/10.1080/19312450709336664>
- Henneman, E. A., Cunningham, H., Fisher, D. L., Plotkin, K., Nathanson, B. H., Roche, J. P., Marquard, J. L., Reilly, C. A., & Henneman, P. L. (2014). Eye tracking as a debriefing mechanism in the simulated setting improves patient safety practices. *Dimensions of Critical Care Nursing*, 33(3), 129–135. <https://doi.org/10.1097/DCC.0000000000000041>
- Hew, K. F., & Cheung, W. S. (2014). Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges. *Educational Research Review*, 12, 45–58. <https://doi.org/10.1016/j.edurev.2014.05.001>
- Hojitink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, 24(5), 539. <https://doi.org/10.1037/met0000201>
- Holmqvist, K., & Anderson, R. (2017). *Eye tracking, A comprehensive guide to methods, paradigms and measures*. Lund Eye-Tracking Research Institute.
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press. <http://books.google.nl/books?id=CjeGZwEACAAJ>
- Ishihara, I. (2017). *Ishihara's tests for color-blindness* (Concise ed.). Tokyo, Japan: Kanehara & Co.
- Jarodzka, H., Balslev, T., Holmqvist, K., Nyström, M., Scheiter, K., Gerjets, P., & Eika, B. (2012). Conveying clinical reasoning based on visual observation via eye-movement modelling examples [*]. *Instructional Science*, 40(5), 813–827. <https://doi.org/10.1007/s11251-012-9218-5>
- Jarodzka, H., Holmqvist, K., & Gruber, H. (2017). Eye tracking in educational science: Theoretical frameworks and research agendas. *Journal of Eye Movement Research*, 10(1), 1–18. <https://doi.org/10.16910/jemr.10.1.3>
- Jarodzka, H., van Gog, T., Dorr, M., Scheiter, K., & Gerjets, P. (2013). Learning to see: Guiding students' attention via a model's eye movements fosters learning. *Learning and Instruction*, 25(0), 62–70. <https://doi.org/10.1016/j.learninstruc.2012.11.004>
- JASP Team. (2021). *JASP (Version 0.16) [Computer software]*
- Jermann, P., & Nüssli, M.-A. (2012). Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task. In *ACM 2012 conference on Computer Supported Cooperative Work* (pp. 1125–1134). <https://doi.org/10.1145/2145204.2145371>
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795.
- Knoop-van Campen, C. A. N., Kok, E., van Doornik, R., de Vries, P., Immink, M., Jarodzka, H., & van Gog, T. (2021). How teachers interpret displays of students' gaze in reading comprehension assignments. *Frontline Learning Research*, 9(4), 116–140. <https://doi.org/10.14786/flr.v9i4.881>
- Kok, E. M., Aizenman, A. M., Vö, M. L.-H., & Wolfe, J. M. (2017). Even if I showed you where you looked, remembering where you just looked is hard. *Journal of Vision*, 17(12), 1–11. <https://doi.org/10.1167/17.12.2>
- Kok, E. M., Hormann, O., Rou, J., van Saase, E., van der Schaaf, M., Kester, L., & Van Gog, T. (2022). Re-viewing Performance: Showing Eye-tracking Data as Feedback to Improve Performance Monitoring in a Complex Visual Task. *Journal of Computer Assisted Learning*, 1087–1101. <https://doi.org/10.1111/jcal.12666>
- Kok, E. M., & Jarodzka, H. (2017a). Before your very eyes: The value and limitations of eye tracking in medical education. *Medical Education*, 51(1), 114–122. <https://doi.org/10.1111/medu.13066>
- Kok, E. M., & Jarodzka, H. (2017b). Beyond your very eyes: eye movements are necessary, not sufficient. *Medical Education*, 51(11), 1190. <https://doi.org/10.1111/medu.13384>
- Kostons, D., van Gog, T., & Paas, F. (2009). How do I do? Investigating effects of expertise and performance-process records on self-assessment. *Applied Cognitive Psychology*, 23(9), 1256–1265. <https://doi.org/10.1002/acp.1528>
- Kurzahls, K., Burch, M., Blascheck, T., Andrienko, G., Andrienko, N., & Weiskopf, D. (2015). A task-based view on the visual analysis of eye tracking data. In Burch, M., Chuang, L., Fisher, B., Schmidt, A., & Weiskopf, D.

- (Eds.), *Eye Tracking and Visualization. ETVIS 2015. Mathematics and Visualization*. https://doi.org/10.1007/978-3-319-47024-5_1
- Lindner, M. A., Eitel, A., Thoma, G. B., Dalehefte, I. M., Ihme, J. M., & Köller, O. (2014). Tracking the decision-making process in multiple-choice assessment: Evidence from eye movements. *Applied Cognitive Psychology*, 28(5), 738–752. <https://doi.org/10.1002/acp.3060>
- Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, 4(1), 6–14. [https://doi.org/10.1016/S1364-6613\(99\)01418-7](https://doi.org/10.1016/S1364-6613(99)01418-7)
- Madsen, J., Júlio, S. U., Gucik, P. J., Steinberg, R., & Parra, L. C. (2021). Synchronized eye movements predict test scores in online video education. *Proceedings of the National Academy of Sciences*, 118(5), e2016980118. <https://doi.org/10.1073/pnas.2016980118>
- Marsman, M., & Wagenmakers, E.-J. (2017). Bayesian benefits with JASP. *European Journal of Developmental Psychology*, 14(5), 545–555. <https://doi.org/10.1080/17405629.2016.1259614>
- Mason, L., Pluchino, P., & Tornatora, M. C. (2015). Eye-movement modeling of integrative reading of an illustrated text: Effects on processing and learning. *Contemporary Educational Psychology*, 41, 172–187. <https://doi.org/10.1016/j.cedpsych.2015.01.004>
- Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R.E. Mayer (ed.), *The Cambridge handbook of multimedia learning*, 2nd ed. (pp. 43–71). Cambridge University Press. <https://doi.org/10.1017/CBO9781139547369.005>
- Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, 84(4), 429. <https://doi.org/10.1037/0022-0663.84.4.429>
- Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). Webgazer: scalable webcam eye tracking using user interactions. In S. Kambhampati (Ed.), *International Joint Conference on Artificial Intelligence* (pp. 3839–3845). AAAI Press.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372.
- Reingold, E. M., & Sheridan, H. (2011). Eye movements and visual expertise in chess and medicine. In S. P. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 528–550). Oxford University Press.
- Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, 29(6), 1045–1060. https://doi.org/10.1207/s15516709cog0000_29
- Richter, J., Scheiter, K., & Eitel, A. (2016). Signaling text-picture relations in multimedia learning: A comprehensive meta-analysis. *Educational Research Review*, 17, 19–36. <https://doi.org/10.1016/j.edurev.2015.12.003>
- Scheiter, K., Schubert, C., & Schüler, A. (2018). Self-regulated learning from illustrated text: Eye movement modelling to support use and regulation of cognitive processes during learning from multimedia. *British Journal of Educational Psychology*, 88(1), 80–94. <https://doi.org/10.1111/bjep.12175>
- Scheiter, K., & van Gog, T. (2009). Using eye tracking in applied research to study and stimulate the processing of information from multi-representational sources. *Applied Cognitive Psychology*, 23(9), 1209–1214. <https://doi.org/10.1002/acp.1524>
- Sharma, K., Jermann, P., & Dillenbourg, P. (2014). “With-me-ness”: A gaze-measure for students' attention in MOOCs. In J. L. Polman, E. A. Kyza, D. K. O'Neill, I. Tabak, W. R. Penuel, A. S. Jurow, K. O'Connor, T. Lee, & L. D'Amico (Eds.), *International Conference of the Learning Sciences* (pp. 1017-1022-1017-1022). ISLS. <https://doi.org/10.22318/icls2014.1017>
- Sharma, K., Leftheriotis, I., Noor, J., & Giannakos, M. (2017). Dual Gaze as a Proxy for Collaboration in Informal Learning. In B. K. Smith, M. Borge, E. Mercier, & K. Y. Lim (Eds.), *International Conference on Computer Supported Collaborative Learning*. International Society of the Learning Sciences. <https://doi.org/10.22318/csl2017.27>
- Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience*, 6(12), 1317. <https://doi.org/10.1038/nn1150>

- Špakov, O., Siirtola, H., Istance, H., & Riih , K. (2017). Visualizing the reading activity of people learning to read. *Journal of Eye Movement Research*, 10(5), 1–12. <https://doi.org/10.16910/jemr.10.5.5>
- S dkamp, A., Kaiser, J., & M ller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, 104(3), 743–762. <https://doi.org/10.1037/a0027627>
- Sweller, J., Van Merri nboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251–296. <https://doi.org/10.1023/A:1022193728205>
- TCL. (2021). <http://www.tcl.tk/>
- van den Bergh, D., Van Doorn, J., Marsman, M., Draws, T., Van Kesteren, E.-J., Derks, K., Dablander, F., Gronau, Q. F., Kucharsk ,  ., & Gupta, A. R. K. N. (2020). A Tutorial on Conducting and Interpreting a Bayesian ANOVA in JASP. *L'Ann e Psychologique*, 120(1), 73–96. <https://doi.org/10.3917/anpsy1.201.0073>
- van Gog, T., & Jarodzka, H. (2013). Eye tracking as a tool to study and enhance cognitive and metacognitive processes in computer-based learning environments. In R. Azevedo & V. Aleven (Eds.), *International Handbook of metacognition and learning technologies* (pp. 143–156). Springer Science+ Business media.
- Van Gog, T., Jarodzka, H., Scheiter, K., Gerjets, P., & Paas, F. (2009). Attention guidance during example study via the model's eye movements. *Computers in Human Behavior*, 25(3), 785–791. <https://doi.org/10.1016/j.chb.2009.02.007>
- van Gog, T., Kester, L., Nieuvelstein, F., Giesbers, B., & Paas, F. (2009). Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction. *Computers in Human Behavior*, 25(2), 325–331. <https://doi.org/10.1016/j.chb.2008.12.021>
- Van Gog, T., & Scheiter, K. (2010). Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction*, 20(2), 95–99. <https://doi.org/10.1016/j.learninstruc.2009.02.009>
- Van Marlen, T., van Wermeskerken, M., & van Gog, T. (2019). Effects of visual complexity and ambiguity of verbal instructions on target identification. *Journal of Cognitive Psychology*, 31(2), 206–214.
- Van Wermeskerken, M., Litchfield, D., & van Gog, T. (2018). What am I looking at? Interpreting dynamic and static gaze displays. *Cognitive Science*, 42(1), 220–252. <https://doi.org/10.1111/cogs.12484>
- van Wermeskerken, M., Ravensbergen, S., & van Gog, T. (2018). Effects of instructor presence in video modeling examples on attention and learning. *Computers in Human Behavior*, 89, 430–438. <https://doi.org/10.1016/j.chb.2017.11.038>
- Villamor, M., & Rodrigo, M. M. (2018). Predicting successful collaboration in a pair programming eye tracking experiment. In *Conference on User Modeling, Adaptation and Personalization* (pp. 263–268). Singapore. <https://doi.org/10.1145/3213586.3225234>
- Zelinsky, G. J., Peng, Y., & Samaras, D. (2013). Eye can read your mind: Decoding gaze fixations to reveal categorical search targets. *Journal of Vision*, 13(14), 1–13. <https://doi.org/10.1167/13.14.10>

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary Material