Received: 26 August 2021 | Revised: 18 January 2022 | Accepted: 27 February 2022

DOI: 10.1111/jcal.12666

ARTICLE

Journal of Computer Assisted Learning WILEY

Re-viewing performance: Showing eye-tracking data as feedback to improve performance monitoring in a complex visual task

Ellen Kok[1] | Olle Hormann[1] | Jeroen Rou[1] | Evi van Saase[1] | Marieke van der Schaaf[1,2] | Liesbeth Kester[1] | Tamara van Gog[1]

[1]Department of Education, Utrecht University, Utrecht, The Netherlands

[2]Center for Research and Development of Education, University Medical Center Utrecht, Utrecht, The Netherlands

Correspondence
Ellen Kok, Department of Education, Utrecht University, Heidelberglaan 1, 3584 CS Utrecht, The Netherlands.
Email: e.m.kok@uu.nl

Abstract

Background: Performance monitoring plays a key role in self-regulated learning, but is difficult, especially for complex visual tasks such as navigational map reading. Gaze displays (i.e. visualizations of participants' eye movements during a task) might serve as feedback to improve students' performance monitoring.

Objectives: We hypothesized that participants who review their performance based on screen recordings that also display their gaze would have a higher monitoring accuracy and increase in post-test performance and would remember more executed actions than participants who review based on a screen recording only (i.e. control condition).

Methods: Sixty-four higher education students were randomly assigned to a gaze-display or control condition. After watching an instruction video, they practiced five navigational map-reading tasks and then reviewed their performance while thinking aloud, either prompted by a screen recording with gaze display or a screen recording only. Before and after reviewing, participants estimated the number of correctly solved tasks and finally made a five-item post-test.

Results and conclusions: Analyses with frequentist and Bayesian statistics showed that gaze displays did not improve monitoring accuracy (i.e. estimated minus actual performance), post-test performance, or the number of reported actions. It is concluded that scanpath gaze displays do not provide useful cues to improve monitoring accuracy in this task.

Takeaways: Gaze displays are a promising tool for education, but scanpath gaze displays did not help to enhance monitoring accuracy in a navigational map-reading task.

KEYWORDS

eye tracking, gaze display, metacognition, monitoring, navigational map reading

1 | INTRODUCTION

When learning a new task, how much do you remember of how you performed on a practice task? Do you remember enough of the actions you

Olle Hormann, Jeroen Rou and Evi van Saase contributed equally to this work.

performed and how effective they were to guide further study or additional practice? Accurately monitoring performance and learning is crucial for effective regulation of subsequent study behaviour but is also very difficult (Nelson & Nahrens, 1990; Zimmerman, 2002). Monitoring accuracy is often operationalized as bias and absolute accuracy (Griffin et al., 2019). Bias is the singed difference between the estimated and actual performance, which gives an impression of whether there is an overall pattern of overestimation or underestimation. However, if both underestimation and overestimation occur, they cancel each other out and one might erroneously conclude that monitoring accuracy is high. Therefore, many researchers analyse absolute accuracy, which is the absolute (i.e. unsigned) difference between the estimated and actual performance. Bias and absolute accuracy values closer to zero reflect greater monitoring accuracy. That is, a learner is aware how well a task was performed and whether or not it needs to be practiced further (Dunlosky & Rawson, 2012).

Accurate performance monitoring upon task completion requires that students remember what actions they performed and how effective those actions were (Kostons et al., 2009). This requires that students monitor their actions and the consequences of those actions while they are engaged in the task. This, however, is difficult, as novel learning tasks typically impose a high working memory load (van Merriënboer & Sweller, 2005). Given that working memory is limited in capacity and duration, there may not be sufficient capacity left for keeping track of what you are doing while you are doing it (Kostons et al., 2009; Van Gog et al., 2005). Therefore, monitoring accuracy might benefit from giving people cues about their performance process. However, especially on visual tasks, there are little if any overt traces of the performance process that could be logged and offered to the learner for review.

Eye-tracking technology could generate process cues for complex visual tasks. Eye tracking is a technique to measure the movements of the eyes to see what a person is looking at, for how long, and in which order (Holmqvist et al., 2011; Kok & Jarodzka, 2017). Whereas eye tracking has been widely used to *investigate* visual behaviour, it could also be used as an educational tool to *improve* task performance and foster learning (Jarodzka et al., 2017; van Gog, Kester, et al., 2009; Van & Scheiter, 2010). Eye-tracking data can be visualized in a gaze display, which is a visualization (image or video) of where a person was looking while executing a task (see Figure 1). Since differences in eye-tracking data are found to be related to differences in task performance (e.g. Dong et al., 2018; Putto et al., 2014), it seems promising to investigate whether a gaze display provides students with process cues for improving learners' monitoring accuracy in a complex visual task. In the current study, we investigate the effect of gaze-display feedback on monitoring accuracy in a navigational map-reading task.

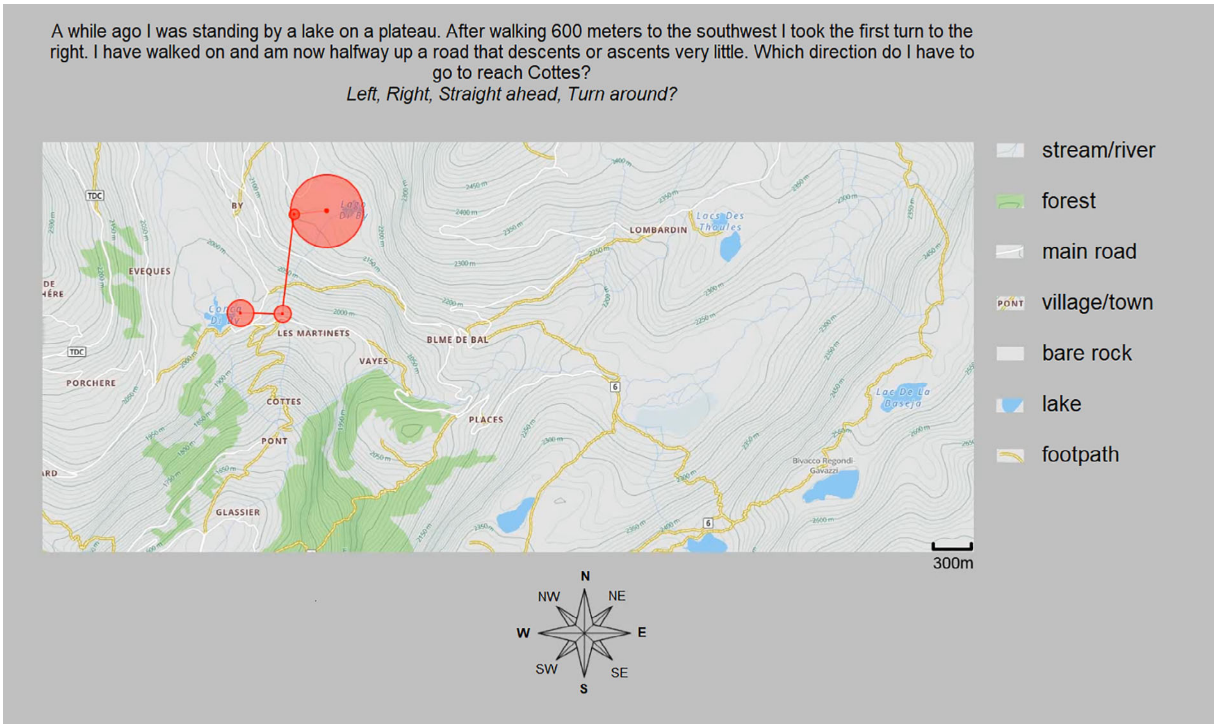
1.1 | Monitoring of performance and learning when training complex visual tasks

What should be monitored when learning a complex visual task? Training complex visual tasks, such as navigational map reading, radiograph interpretation, luggage inspection, or military threat detection typically includes three aspects: the use of task-specific technology or equipment, the object identification (i.e. what are the visual features of targets and non-targets), and search strategy (Kramer et al., 2019). The training of task-specific technology or equipment (i.e. knobology) refers to technical knowledge of the specific features of technology, such as how to change settings. Learning the visual features of targets and non-targets to be able to identify objects is the next central aspect of learning complex visual tasks. In navigational map reading, an example is learning how contour lines represent the shape of the landscape (e.g. small circles denote a mountaintop). Finally, training search strategies is important for developing the skills to execute visual tasks, and it arguably has a larger potential for generalization to untrained stimuli, as it can help people to recognize objects that were not explicitly trained. While search pattern training is a central component of learning a visual task, search pattern training alone is rarely effective to improve performance (Kok et al., 2016; Kramer et al., 2019; van Geel et al., 2017) and a combination with learning visual features and knobology is required.

What makes monitoring of performance on those three aspects especially difficult is precisely the fact that they are highly visual: There are often no overt actions that can serve as anchors for memorizing the perceptual processes while doing the task. Indeed, it has been shown that people have trouble remembering their own viewing behaviour in visual tasks. For example, people can hardly remember where they have searched for objects in scenes (Clarke et al., 2016; Kok et al., 2017; Marti et al., 2015; Võ et al., 2016). When people are presented with displays of their gaze and the gaze of others, they even have trouble recognizing which displays show their own gaze, and which displays show that of others (Clarke et al., 2016; van Wermeskerken et al., 2018). This might become problematic in critical visual tasks such as interpreting radiological images. For instance, Aizenman et al. (2017) found that radiologists report a very different search strategy from the one they actually executed. This inaccurate performance monitoring makes it difficult to adapt an ineffective search strategy and use more effective alternatives.

Whereas monitoring of search strategies is known to be difficult, little is known about monitoring the identification of visual features. It was found that people have some awareness as to what characteristics of targets and distractors make a search task difficult (Green & Redford, 2016; Redford et al., 2011). At the same time, phenomena such as inattentional blindness (Simons & Chabris, 1999) suggest incomplete metacognitive awareness of visual processes.

Not knowing where you have looked (longer) could be problematic because differences in viewing behaviour exist between correct and incorrect feature identification, for example in radiology (Manning et al., 2006). Manning and colleagues found that decisions were likely to be incorrect if participants looked at a visual feature (potential tumour) for a long time, and they suggest that gaze durations might provide cues about the effectiveness of object recognition, which can be fed back to the participant. Thus, information about viewing behaviour could provide learners with important cues to monitor their performance on a complex visual task. For example, it could show which areas of a stimulus are not inspected, and which areas are looked at very long (possibly because object recognition was difficult there).



A while ago I was standing by a lake on a plateau. After walking 600 meters to the southwest I took the first turn to the right. I have walked on and am now halfway up a road that descents or ascents very little. Which direction do I have to go to reach Cottes?

Left, Right, Straight ahead, Turn around?

FIGURE 1 Screenshot from a dynamic gaze display. The gaze location is shown in red with a two-second trail. Circles denote fixations (i.e. moments when the eye is relatively still and takes in information) with the size of the fixation reflecting its duration. Lines denote jumps between fixations (saccades). The text was translated from Dutch

2 | GAZE DISPLAYS

Eye-tracking technology can make covert viewing behaviour visible by visualizing people's viewing behaviour. Such visualizations, also known as gaze displays (see Figure 1) have been suggested to provide a promising tool for education (Jarodzka et al., 2017; van Gog, Kester, et al., 2009; Van & Scheiter, 2010). For instance, it has been shown that displaying the gaze of a teacher to learners in eye movement modelling examples enhances learning on a variety of tasks compared to regular video modelling examples without the gaze displayed (e.g., Chisari et al., 2020; Jarodzka et al., 2012; Mason et al., 2015; Van Gog et al., 2009; for a meta-analysis see Xie et al., 2021). It has also been proposed that displaying the gaze of a learner back to them as a cue could improve monitoring accuracy (van Gog et al., 2009). However, this has not yet been widely investigated.

3 | GAZE DISPLAYS AS FEEDBACK

Displaying the gaze of a learner back to them could be considered a form of feedback since it provides information regarding aspects of one's performance of a task (Hattie & Timperley, 2007). Only a few studies about gaze displays as feedback have focused on improving monitoring accuracy, and most studies have only focused on task performance. It could be assumed that improved task performance is (at least partly) caused by increased monitoring accuracy, but this is often not directly measured.

Positive effects of the use of gaze-display feedback on task performance were found in several studies, but these studies either did not include a no-gaze display control condition (Sommer et al., 2016; Tsai et al., 2019) or the gaze display was part of a larger intervention (Eder et al., 2020; Henneman et al., 2014; O'Meara et al., 2015; Wilson et al., 2011; Zhai et al., 2018). In more controlled experiments with the gaze display being the only difference between conditions, the effectiveness of gaze-display feedback to improve performance seemed to depend on which aspect of learning the visual task is supported. If learning a search strategy is supported, gaze-display feedback often does not result in increased performance (Dickinson & Zelinsky, 2005; Donovan et al., 2005; Drew & Williams, 2017; Peltier & Becker, 2017), although small and selective effects of gaze-display feedback on search performance were also found (Otto et al., 2018; Qvarfordt et al., 2010). In those studies, a complete search (i.e. looking at all parts of the stimulus) was required, and the display conveyed which information was looked at (even for a short time), for example by colouring areas that were previously ignored, or colouring or blurring areas that were looked at. It seems that providing cues on the use of a search strategy does not improve task performance.

What does seem to result in increased performance, however, is feedback in the form of a gaze display that shows areas of prolonged attention (Donovan et al., 2008; Kundel et al., 1990). Donovan and colleagues showed participants their fixation locations as circles, the size of which depended on viewing duration. Clusters of large circles thus showed areas of prolonged attention. Kundel and colleagues showed circles around areas that were looked at for at least 1000 ms. It can be argued that the gaze-display feedback does not so much assist the search strategy training but rather shows areas where the process of recognizing visual features is less than optimal (as reflected in prolonged attention). This could provide a useful cue for monitoring performance and learning processes.

Whereas some positive effects of gaze-display feedback on *performance* were found (but mostly when it shows areas of prolonged attention), only two studies have investigated the effect of gaze-display feedback on *monitoring accuracy*. Kok et al. (2017) showed participants their eye movements *while* performing a visual task (search for objects in where-is-Waldo stimuli). The gaze display in this study was a spotlight display, in which the location looked-at was lighter and the background darker. Afterward, participants were asked to report where they had looked. Participants in the gaze-display feedback condition were indeed more correct in reporting their viewing locations than participants in a control condition without gaze-display feedback. However, their monitoring accuracy was still rather low. It can be argued that simultaneously monitoring which locations were viewed and conducting a challenging task is simply too cognitively taxing, and gaze visualizations should be presented after task performance. Many self-regulated learning theories include an 'appraisal' phase, in which a learner looks back on task performance, reflects on the quality, and adapts for future performance (Panadero, 2017). It could be that inspecting a gaze display after task performance (in an 'appraisal phase') is more effective in improving monitoring accuracy.

The only study that investigated the effects of gaze displays after task performance (i.e. during appraisal) on metacognitive accuracy was executed by Kostons et al. (2009). In this study, adult participants executed a problem-solving task (about the laws of heredity). Afterwards, they were asked to judge their performance, either based on a screen recording (control condition) or based on a screen recording with gaze display (i.e. gaze-display feedback). It was found that the gaze display helped participants with a lower level of expertise to report more performed actions. This suggests that without the display they indeed did not remember the specifics of their performance. This was presumably due to high cognitive load as this effect was not found for students with a higher level of expertise. Furthermore, the display helped participants with a higher level of expertise to evaluate their task performance (e.g. "I think I did that wrong"). However, those findings were based on analyses of verbal protocols; participants did not make monitoring judgements. Thus, it is as yet unclear whether gaze displays would indeed improve monitoring accuracy.

4 | THE PRESENT STUDY

Learners often have trouble remembering their viewing behaviour in visual tasks (Clarke et al., 2016; Kok et al., 2017; Marti et al., 2015; Võ et al., 2016), which makes it difficult to monitor and regulate their learning of complex visual tasks. Gaze displays show learners where they looked, which might provide them with process cues on which they can base their monitoring. Improved monitoring, in turn, might improve post-test performance. However, this has not been investigated yet. Therefore, our research question is: Does reviewing performance based on replays of their eye movements (i.e. gaze displays) result in a higher increase in monitoring accuracy and a higher increase in performance compared to a control condition in which participants review performance based on a screen recording without gaze visible?

In the present study, we investigate the effect of gaze-display feedback during appraisal (i.e. the review phase) on monitoring accuracy and post-test performance in a navigational map-reading task. While it is relatively easy to train laypeople to execute this task, the processes involved are comparable to those in other complex visual tasks. For example, the interpretation of contour lines is considered difficult because it requires that elevation information be extracted from the 2D lines on paper, and interpreted in terms of for example mountains and valleys (Putto et al., 2014). This is very similar to the difficulty of interpreting 2D information in a radiograph into a 3D representation of the human body (van der Gijp et al., 2015).

Thus far, little is known about the effects of gaze-display feedback on monitoring accuracy. However, two studies discussed earlier suggest that gaze-display feedback might support process monitoring during an appraisal phase. Kok et al. (2017) found that displaying gaze improved process monitoring during a task. Kostons et al., 2009 found that high-scoring participants made more metacognitive remarks during the review phase as a result of gaze displays (but they did not measure monitoring accuracy). Therefore, we hypothesize that reviewing performance would result in a higher increase in monitoring accuracy for participants in the gaze-display condition compared to the control condition. Monitoring accuracy will be operationalized as bias, which is the signed difference between estimated and actual performance, and absolute accuracy, which is the absolute (i.e. unsigned) difference between estimated and actual performance (Griffin et al., 2019).

As for the effects of gaze displays on performance, the studies discussed earlier found some evidence that feedback in the form of gaze-displays could support performance (Donovan et al., 2008; Kundel et al., 1990), but this applied mainly when the displays provided information about the effectiveness of interpreting visual features, and not when they provided cues on the use of a search strategy (Dickinson & Zelinsky, 2005; Donovan et al., 2005; Drew & Williams, 2017; Peltier & Becker, 2017). Since we provide learners with gaze-display feedback regarding their interpretation of visual features, we hypothesize that participants in the gaze-display condition will show a higher post-test performance than participants in the control condition.

In sum, we hypothesize that reviewing performance would result in a higher increase in monitoring accuracy (less bias and absolute accuracy) and a higher increase in performance for participants in the gaze-display condition compared to the control condition (i.e. participants review performance based on a screen recording without gaze visible). To explore how gaze displays might affect monitoring accuracy, verbal protocols of the review phase from a subsample of participants will be analysed. In line with the findings by Kostons et al. (2009), we expect that participants in the gaze-display condition would report more (cognitive) actions than participants in the control (screen-recording only) condition. Furthermore, we explore in more detail what actions would be reported, in which we distinguished between problem-solving steps (cf. search strategy training), recognition of visual features, and use of map features (cf. use of tools).

5 | MATERIALS AND METHODS

5.1 | Participants and design

The independent variables were the condition (screen recording vs. gaze display) and time (for monitoring accuracy: before and after review, for test performance: practice and post-test). The experiment had a 2 × 2 mixed factorial design with condition as between-subjects factor and time as within-subjects factor. Dependent variables were absolute accuracy, degree of bias, and test performance (number of correctly solved tasks).

Participants were 75 students from Dutch universities or universities of applied sciences. The majority ($n = 60$) were social sciences students, the others studied in health professions education ($n = 5$) and other programs ($n = 10$). Participants reported limited expertise with navigational map reading and participated in the experiment to learn more about navigational map reading. Two participants who were initially included but turned out to have a different educational background (vocational education or high school) were excluded from the sample. Three participants were excluded due to technical problems during data collection. Finally, eight participants were excluded because they received incorrect instructions during the think-aloud phase. The final sample included 15 male and 49 female participants (M_{age} 22.8 years, $SD = 3.4$, range = 18.2–40.8 years), who were randomly assigned to the gaze display ($n = 34$) or control condition ($n = 30$). The study was approved by the research ethics committee of the institute where this study was conducted; all participants provided written informed consent. The data that support the findings of this study are openly available in Dataverse at https://doi.org/10.34894/9KCFB6.

5.2 | Materials

5.2.1 | Prior knowledge test

The written prior knowledge test consisted of five images of landscape features that were shown one by one (mountaintop, plateau, mountain pass, brook valley, ridge). The images showed how those features were represented in contour lines. Participants were required to write the name of that landscape feature if they knew it. The total score was the total number of correct answers out of five.

5.2.2 | Instruction video

The instruction video consisted of 27 PowerPoint slides with an audio explanation (see Figure 2). In this instruction video, the visual features of five landscape features were shown and explained (mountain top, plateau, mountain pass, brook valley, ridge), as well as the interpretation of contour lines (contour lines connect points with the same height, the closer together they are, the steeper the area) and the use of the legend, rose, and scale. Furthermore, a navigation strategy was introduced that stressed that the participant should first use all available information to decide on the location, and only after that decide on the direction to head next. It was further stressed that all possible locations should be scrutinized before settling on a location. All of this was repeated and demonstrated in a worked example. The length of the video was 9 min and 46 s, and the video could not be paused or replayed.

5.2.3 | Maps and tasks for the practice phase and post-test

Maps were screenshots from geocaching.com which are developed by OpenStreetMap.org (© OpenStreetMap contributors). Maps are licensed as CC BY-SA, see https://www.openstreetmap.org/copyright. This source was selected because the maps do not use shaded relief, so contour line interpretation is required to recognize landscape features. All maps had a scale of 1:200,000 and showed mountainous terrain. Maps were accompanied by written descriptions of location and destination, a legend, a scale, and a rose. For each task, participants were required to establish their location based on the written description, and decide where to head next (left, right, straight ahead, or turn around). Once they had decided on their answer, they hit the space bar and selected the correct answer from the four options on the next page. The difficulty of 12 tasks (See Figure 2,3 for an example task) was pilot tested with 12 participants, who watched the instruction video and solved all 12 tasks, without being eye-tracked and without appraisal phase. Two tasks were removed for being too easy or unclear. The remaining 10 tasks (five for the practice phase, five for the post-test) were used in this study. The total score of both the practice phase and the post-test was the number of correct answers (out of five).

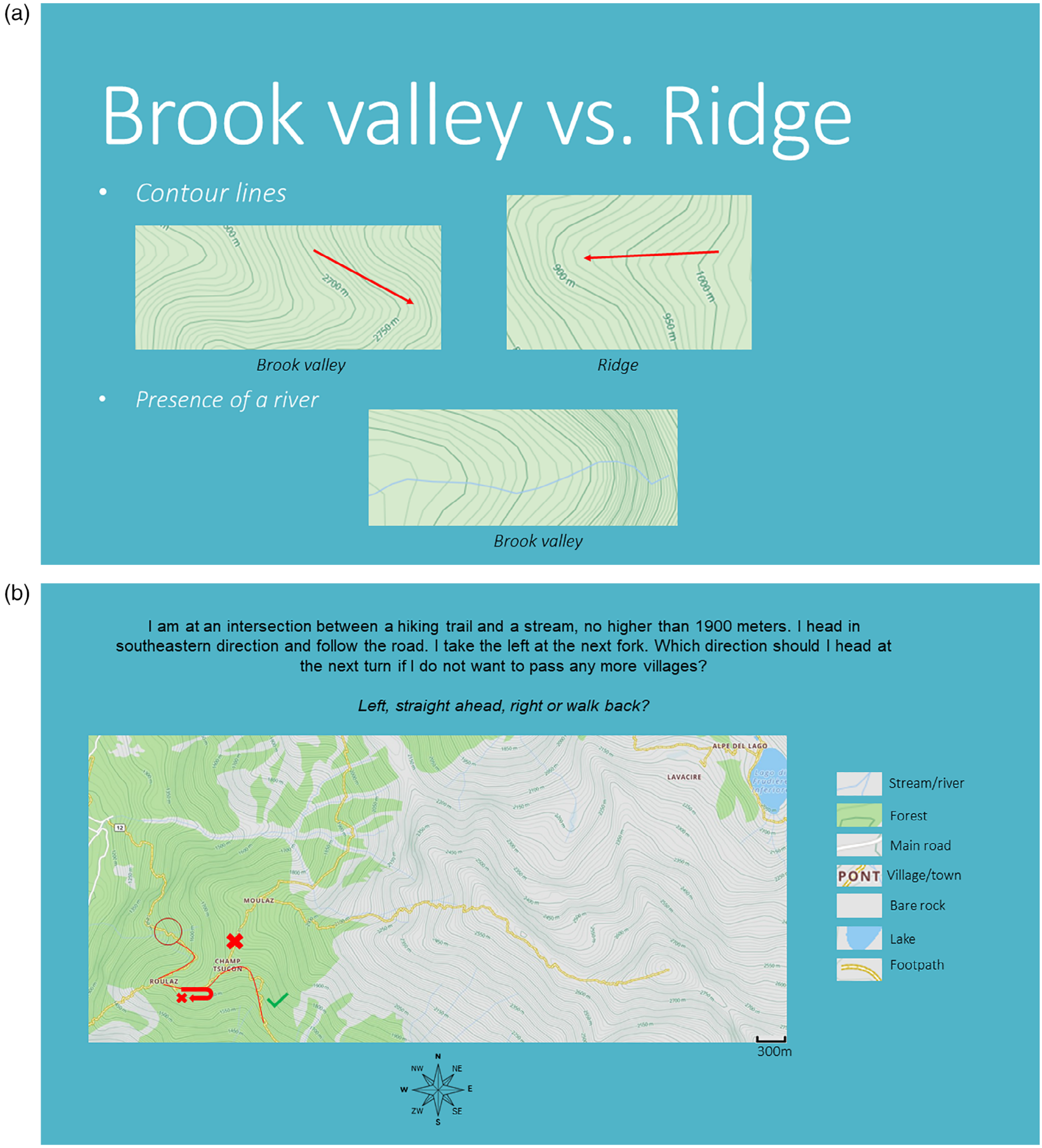
5.2.4 | Monitoring judgement

Directly after the practice phase, participants were asked to indicate on a multiple-choice scale (answer options 0–5) how many out of the five tasks they think they had solved correctly. The same question was asked directly after the performance review.

5.2.5 | Performance review

Control condition

In the control condition, the performance review phase consisted of observing a full-screen replay of the practice phase (i.e. the original stimulus was presented for the same amount of time that participants took during the practice phase) in which their chosen answer



Brook valley vs. Ridge

- *Contour lines*

Brook valley

Ridge

- *Presence of a river*

Brook valley

I am at an intersection between a hiking trail and a stream, no higher than 1900 meters. I head in southeastern direction and follow the road. I take the left at the next fork. Which direction should I head at the next turn if I do not want to pass any more villages?

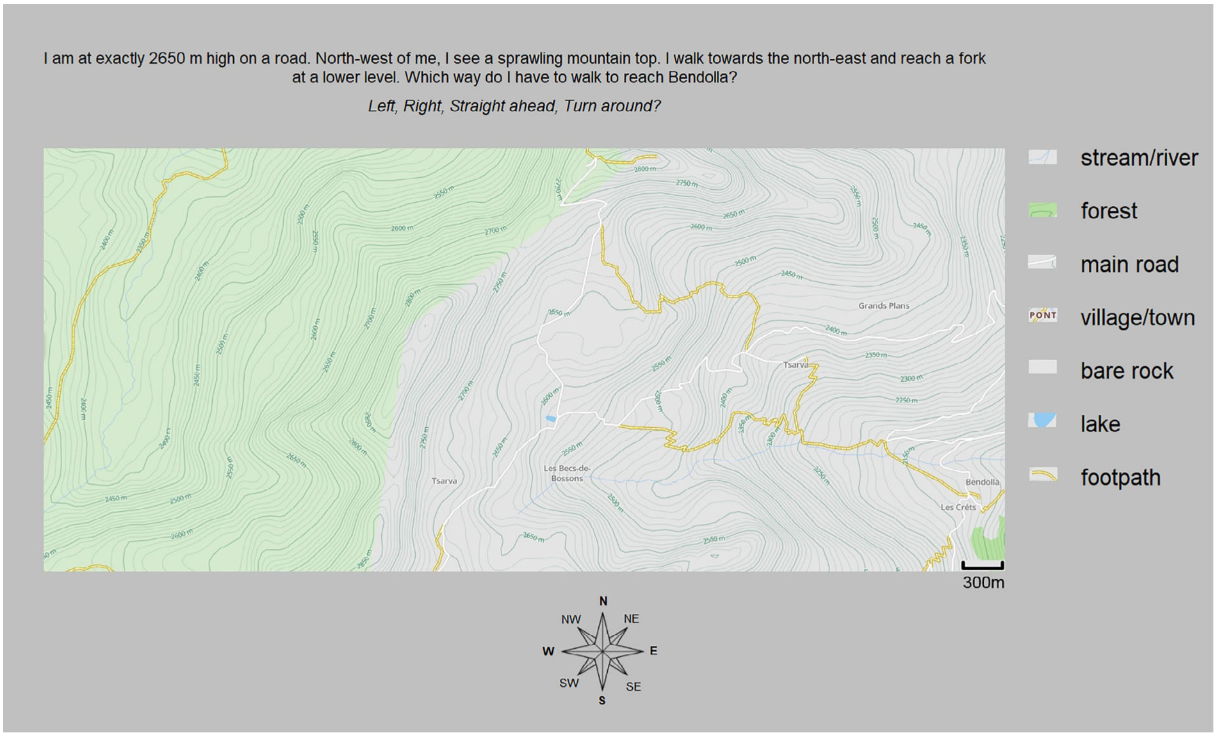
Left, straight ahead, right or walk back?

FIGURE 2 Two screenshots from the instruction video. (a) Taken from the first half of the instruction video. On this slide, an explanation is given about the differences between a brook valley and a ridge, since they are visually quite similar. (b) Taken from the worked example. On this slide, an example is given on how to choose the right direction after knowing your location. Text translated from Dutch

(i.e. mouse click on one of the answers) on a task was depicted as a red diamond. The replay was presented using the scanpath utility of the BeGaze software (Version 3.7; SensoMotoric Instruments GmbH, 2017), but with the gaze data hidden.

Gaze-display condition

In the gaze-display condition, the performance review phase consisted of observing a screen recording with an overlaid dynamic visualization of the participant's gaze during the practice phase, using the scanpath utility of the BeGaze software (Version 3.7; SensoMotoric Instruments GmbH, 2017; cf. screenshot in Figure 1). The SMI high-velocity event detection algorithm was used to detect fixations, with a velocity threshold of 40°/s and a minimum fixation duration of 50 ms. Fixations were then displayed as red circles with the size displaying the duration of fixation (100 ms = 50 pixels) and were connected with 2 px lines. A trail of 2 s was shown, that is,



I am at exactly 2650 m high on a road. North-west of me, I see a sprawling mountain top. I walk towards the north-east and reach a fork at a lower level. Which way do I have to walk to reach Bendolla?

Left, Right, Straight ahead, Turn around?

FIGURE 3 Example task from the post-test. Text translated from Dutch. This map had the highest average score ($M = 0.96$), the correct answer to this map is 'Right'

the displays showed the participant's current gaze location as a circle as well as the fixations (circles) and connecting lines in the 2 s preceding that.

5.2.6 | Colour blindness test

We used the Ishihara colour blindness test (Ishihara, 2017) to screen for colour blindness. This test consists of 11 plates (and three additional plates for detailed screening) with embedded numbers that participants are required to read aloud. Participants with normal colour vision can easily see the numbers, but participants with different colour deficiencies cannot distinguish them correctly. Participants with more than one error might have colour deficiencies. None of the participants had to be excluded for making more than one error.

5.3 | Apparatus

The experiment was conducted using SMI Experiment Center (Version 3.7; SensoMotoric Instruments GmbH, 2017) and presented on a 22-inch monitor (1680 × 1050 pixels), the screen subtended 44° horizontally and 28° vertically at a viewing distance of 59 cm. Eye movements were recorded using the SMI RED250 eye tracker at a sampling rate of 250 Hz (SensoMotoric Instruments GmbH, 2017). Replays were presented using the BeGaze software (version 3.7). A headrest was used to stabilize the participant's head position to increase data quality during the practice phase and post-test but not during the video and the replay.

5.4 | Procedure

Participants were individually tested in soundproof rooms. After providing written informed consent, participants were screened for colour blindness and conducted the written pre-test. Subsequently, participants were placed comfortably behind a computer. They first watched the instruction video. After that, a nine-point calibration procedure with a 4-point validation procedure was completed a maximum of three times while striving for deviations less than 0.5° of visual angle, but accepting calibration values below 1.5° of visual angle after three tries. The calibration with the lowest average deviation was accepted. In the experimental condition, the average deviation was $M_x = 0.53$, $SD_x = 0.21$, $Max_x = 1.0$, $M_y = 0.51$, $SD_y = 0.22$, $Max_y = 1.1$. In the control condition, the average deviation was $M_x = 0.81$, $SD_x = 1.24$, $Max_x = 7.2$, $M_y = 0.63$, $SD_y = 0.37$, $Max_y = 1.8$. Participants in the control condition were not excluded if they did not meet acceptable calibration values.

Participants practiced on five maps (without performance feedback). Next, participants were instructed to watch the recording (of the screen/their eye movements) and evaluate their performance while thinking aloud. When participants stopped thinking aloud for 3 s or more, the experimenter would prompt them to continue talking by saying: "Could you please continue to talk?"

Before and after the performance review, participants estimated the number of correctly solved tasks. Finally, they completed a five-item post-test. The entire procedure took approximately 1 h.

5.5 | Data analysis

5.5.1 | Performance and monitoring accuracy

Performance was scored by counting the number of correctly solved tasks (0–5). Monitoring accuracy was operationalized in terms of bias and absolute accuracy. Bias is the difference between estimated and actual performance (number of correctly solved practice tasks) and can range from -5 (complete underestimation) to $+5$ (complete overestimation), with 0 being fully accurate. Absolute accuracy is the absolute (i.e. unsigned) difference between the estimated and actual performance and can range from 0 (fully accurate estimation) to 5 (fully inaccurate estimation).

5.5.2 | Coding the think-aloud protocols

For the content analysis, the verbal protocols from the first 15 participants in each condition were transcribed and analysed (i.e. 30 in total). The transcribed protocols were first segmented into meaningful units before coding took place, in line with recommendations of Strijbos et al. (2006). A fine-grained segmentation scheme was developed, in which segments generally hold a single action or thought (i.e. a single finite verb and subject). For example, the following fragment contains three segments: "The first thing I did was reading the story, line by line (1). I saw that it contained the word 'lake' (2) so I looked at all lakes on the map (3)."

Then the units were coded according to a coding scheme that was an adapted version of the scheme used by Kostons et al. (2009), see Table 1. The three main codes (survey, action, and monitoring/assessment) were taken from their coding scheme. We added the main codes 'visualization' (for remarks about the gaze display) and 'other'. In line with the earlier distinction between training in the use of task-specific technology, object identification training, and search strategy training (Kramer et al., 2019), we subdivided 'action' codes into the use of map features (rose, legend, and scale) to reflect the use of task-specific technology, recognizing visual features (object identification), and problem-solving steps (search strategy). For monitoring/assessment, we used Kostons et al. (2009) subdivision into adequacy, efficiency, affect, and difficulty, and further subdivided them into positive and negative comments.

Three of the protocols were used to practice the segmentation and coding process: They were individually segmented and coded by two individuals. Differences between coders were discussed and if necessary, clarifications to the segmentation and coding rules were added. After that, four of the protocols were independently segmented and coded by two individuals. The average upper and lower boundary for the estimated interrater reliability (proportion overlap) in segmentation were 0.78 and 0.80 (calculated in line with recommendations by Strijbos et al., 2006), which is considered acceptable. Krippendorff's alpha was calculated as a measure of inter-rater reliability (using the KALPHA macro in SPSS; Hayes & Krippendorff, 2007) because it can handle a large number of categories and is thus well suited for content analysis. Alpha was 0.87, which is considered good (Hayes & Krippendorff, 2007). All other protocols were segmented and coded by one of the coders.

5.5.3 | Statistical analyses

Mixed ANOVAs on monitoring accuracy (bias and absolute accuracy) and performance with the condition (gaze-display/control) as between-subjects and time as a within-subjects factor (monitoring accuracy: before/after review; performance: practice/post-test) were conducted in IBM SPSS 24. Since all dependent variables showed substantial non-normality, additional Mann–Whitney *U* tests were conducted with the condition as independent variable and bias and absolute accuracy after review and performance on the test as dependent variables. For non-significant results, Bayesian analyses were executed to quantify the evidence for the null-hypothesis (i.e. no differences between conditions) versus a difference between conditions using Bain Welch's t-tests using the Bain package (Hoijtink et al., 2019) in JASP 0.12.2. Higher Bayes factors reflect stronger evidence, where for example a Bayes factor of 3 means that the support in the observed data is three times larger for the null-hypothesis than the alternative hypothesis. Guidelines for the interpretation of Bayes factors differ widely and many statisticians argue against any cut-off values as they are arbitrary (cf. p <0.05). Even so, most guidelines are similar in that Bayes factors between 1 and 3 are considered ignorable evidence (e.g., Kass & Raftery, 1995), that is, it is not clear whether the null hypothesis or the alternative hypothesis describes the data better. Therefore, we consider Bayes factors higher than 3 as substantial evidence and refrain from further interpretation. For the verbal data, substantial non-normality was found for all variables, so Mann–Whitney U Tests were used to analyse differences between conditions in numbers of codes.

6 | RESULTS

6.1 | Monitoring accuracy

Descriptive statistics for the prior knowledge test, monitoring judgements (estimated score), and actual score can be found in Table 2, and for bias and absolute accuracy in Table 3. Figure 4 shows the violin plots of bias and absolute accuracy.

For bias, the mixed ANOVA showed no main effect of condition, $F(1,62) = 0.16$, $p = 0.69$, $\eta^2_p = 0.01$, no main effect of time, $F(1,62)$ <0.001, $p = 0.99$, η^2_p <0.01, and no significant interaction effect,

TABLE 1 Definition and example of each code

Code	Definition	Example
1. Survey	Comments relating to surveying information in the task and task characteristics, orientation on the task. If the participant read the description aloud, this was also coded as 'Survey'	The first thing I did was reading the story, line by line. (P03T01)
2. Action	Comments relating to performing task-related actions (i.e. what did I do). A separation was made between the use of map features, recognizing visual features, and performing problem-solving steps	I'm looking for a walking path at twenty-one hundred and fifty meters (P13T5)
2.1 Use of map features	Comments related to the use of the legend, scale and rose	Then I checked what southwest was, that was bottom-left. (P03T1)
2.2 Recognizing visual features	Comments related to recognizing visual features	I looked at the contour lines, how much they would be apart because the path descends very little (P18T1)
2.3 Problem-solving steps	Comments related to performing problem-solving steps	And then I looked for a brook valley (P25T3)
3. Monitoring/Assessment	Comments related to monitoring task performance. For each code, positive and negative are coded separately if possible	
3.1 Adequacy	Evaluations of the adequacy/effectiveness of problem-solving steps, the overall process, or knowledge/ ability	Um so I did not quite get it (P19T5)
3.2 Efficiency	Evaluations including a time component	Also found one fairly quickly (P25T3)
3.3 Affect	Evaluations of emotional and motivational states	Well, then I was a little bit stressed (P24T1)
3.4 Difficulty	Evaluations concerning the difficulty of the task	This question was very difficult for me (P16T2)
4. Visualization	Remarks about the visualization or about looking behaviour	But I cannot determine where I read because it is not a very good calibration (P09T2)
5. Other	Everything that cannot be coded otherwise	

TABLE 2 Descriptive statistics for the prior knowledge test, monitoring judgement, and actual score ($n = 64$)

			Monitoring judgement (estimated score)				Actual score			
	Prior knowledge test score		Before review		After review		Practice phase		Post-test	
Condition	*M*	SD	*M*	SD	*M*	SD	*M*	SD	*M*	SD
Gaze display	0.53	0.56	2.53	0.79	2.50	1.02	2.79	1.15	4.03	0.87
Control	0.53	0.57	2.43	0.77	2.47	1.22	2.83	1.05	3.60	0.97

Note: The maximum score for all tests was 5.

$F(1,62) = 0.07$ $p = 0.79$, η^2_p <0.01. The Mann–Whitney *U* test on bias after performance review did not show a significant difference between conditions either, $U = 490.0$, $p = 0.78$. A Bayes factor of 7.78 was found, which indicates substantial evidence that there is no difference between conditions.

For absolute accuracy, the mixed ANOVA showed no main effect of condition, $F(1,62) = 1.12$, $p = 0.29$, $\eta^2_p = 0.02$, no main effect of time, $F(1,62) = 0.77$, $p = 0.38$, $\eta^2_p = 0.01$, and no significant interaction effect, $F(1,62) = 0.38$, $p = 0.54$, $\eta^2_p = 0.01$. The Mann–Whitney U test on absolute accuracy after performance review did not show a significant difference between conditions either, $U = 487.0$, $p = 0.74$. A Bayes factor of 7.02 in favour of the null-hypothesis was found, which indicates substantial evidence that there is no difference between conditions.

An explorative analysis of the direction in which participants changed their estimates also showed no differences between conditions. In the control condition, 11 (36.7%) participants did not change their estimate, 9 (30.0%) participants changed upwards and

TABLE 3 Average bias and absolute accuracy before and after the review phase ($n = 64$)

	Bias				Absolute accuracy			
	Before review		After review		Before review		After review	
Condition	*M*	SD	*M*	SD	*M*	SD	*M*	SD
Gaze-display	−0.26	1.19	−0.29	1.27	0.91**	0.79	0.94**	0.89
Control	−0.40*	0.89	−0.37	1.10	0.67**	0.71	0.83**	0.79

Note: Bias and absolute accuracy were significantly different from zero.
*p <0.05. **p <0.001.

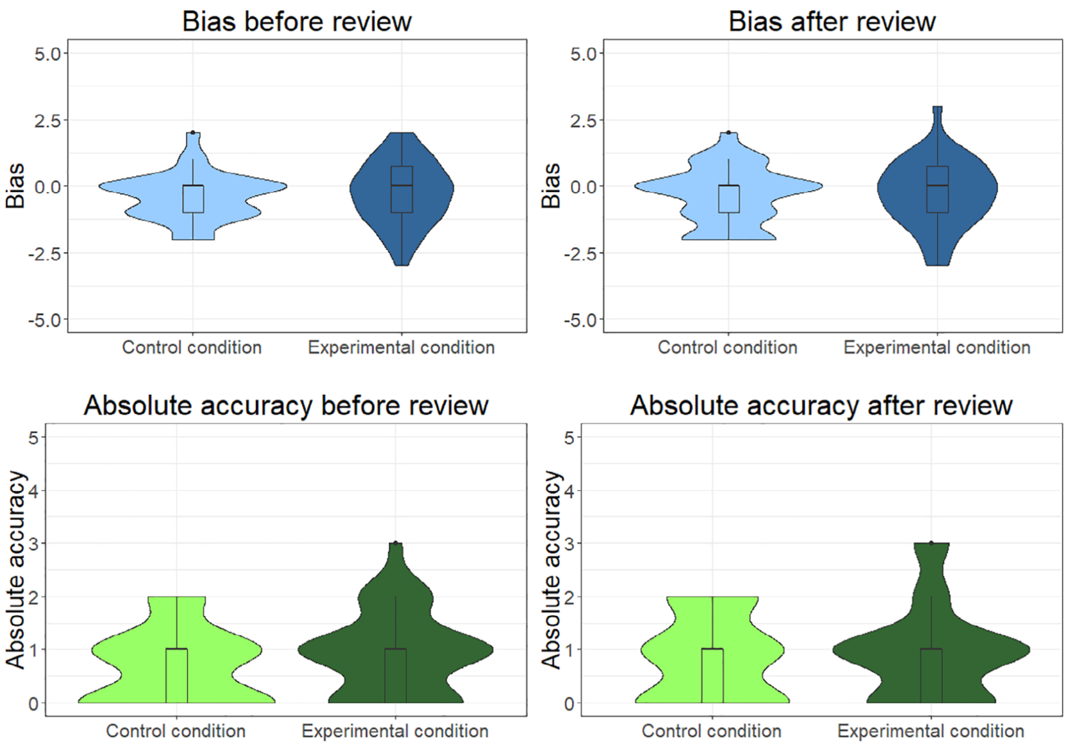


FIGURE 4 Violin plots with embedded boxplots for variables bias (before and after the review phase) and absolute accuracy (before and after review phase)

10 (33.3%) participants changed downwards. In the experiment condition, 14 (32.4%) participants did not change their estimate, 9 (26.2%) participants changed upwards and 11 (32.4%) participants changed downwards.

6.2 | Task performance

The mixed ANOVA on performance showed a main effect of time, $F(1,62) = 36.85$, p <0.001, $\eta^2_p = 0.37$, indicating that participants in both conditions did better on the post-test tasks compared to the practice tasks. However, there was no main effect of condition on performance, $F(1,62) = 1.01$, $p = 0.32$, $\eta^2_p = 0.02$, and no significant interaction effect, $F(1,62) = 2.02$, $p = 0.16$, $\eta^2_p = 0.03$. Although numerically, the gaze display condition seemed to outperform the control condition on the post-test tasks, The Mann–Whitney *U* test did not show a significant effect of condition on post-test performance either, $U = 384{,}5$, $p = 0.07$. A Bayes factor of 1.42 in favour of the null hypothesis was found, which can be considered ignorable evidence in either direction.

6.3 | Verbal data

The average number of segments per task was 20.66 (SD = 4.57). The average number of codes per task can be found in Table 4. No effects of condition were found on the number of comments in categories Survey ($U = 103.0$, $p = 0.69$), Action ($U = 86.5$, $p = 0.28$), and Monitoring ($U = 112$, $p = 0.98$). Bayes factors showed substantial evidence for the null-hypothesis that there was no difference between conditions, $BF_{survey} = 5.14$, $BF_{action} = 5.11$, $BF_{monitoring} = 5.41$.

Looking at the three different types of actions, Mann Whitney U tests showed no effects of condition on the number of problem-solving steps reported ($U = 99$, $p = 0.58$) and the number of comments on visual feature recognition ($U = 74.5$, $p = 0.10$). Bayes factors showed substantial evidence for the null hypothesis that there

TABLE 4 Average number of codes per task for survey, action, and monitoring remarks

	Gaze display		Control	
	M	SD	*M*	SD
Survey	3.60	1.65	3.43	0.87
Action				
Total	10.57	2.16	10.16	3.72
Problem-solving steps	8.96	2.14	9.17	3.47
Recognizing visual features	0.35	0.57	0.44	0.34
Use of map features	1.27	0.63	0.55	0.39
Monitoring				
Total	3.32	1.71	3.43	1.87
Positive	0.48	0.41	0.72	0.49
Negative	2.28	1.41	2.35	1.48
Visualization	0.20	0.25	0.03	0.07
Other	3.45	1.59	3.13	1.28

Note: The total number of monitoring statements also includes monitoring statements without a clear direction.

was no difference between conditions, $BF_{problem-solving\ \ steps} = 5.37$, $BF_{recognizing\ \ visual\ \ features} = 4.72$. However, participants in the gaze-display condition reported using map features significantly more often ($U = 32$, p <0.001). The Bayes factor in favour of the hypothesis that conditions are not equal is very large, $BF_{use\ of\ map\ features} = 233.1$.

As can be seen in Table 4, participants made more than three times as many negative monitoring remarks as positive remarks. Although numerically, more positive and negative monitoring remarks were made in the control condition, those differences between conditions were neither significant for the number of positive remarks ($U = 76.5$, $p = 0.13$), nor for the number of negative remarks ($U = 109.5$, $p = 0.90$). The Bayes factors provide substantial support that no differences exist for the number of negative remarks, $BF_{negative} = 5.43$, but ignorable evidence that there are no differences in the number of positive remarks, $BF_{positive} = 1.93$.

7 | DISCUSSION

The present experiment aimed to investigate whether reviewing your performance utilizing a screen recording with gaze overlaid (gaze-display feedback) would improve monitoring accuracy and post-test performance on a navigational map-reading task, compared to reviewing based on a screen recording only. The gaze display, which showed learners where they looked while executing the task, was expected to provide cues to learners to inform their monitoring judgement.

In contrast to our expectations, however, we found substantial evidence that gaze displays did not increase monitoring accuracy or performance in a navigational map-reading task. Absolute monitoring accuracy and performance scores did not change significantly from before to after the review phase in either condition. Yet, in both conditions and both phases, absolute accuracy was significantly different from zero, meaning participants were significantly inaccurate in monitoring. The fact that the average bias was close to zero suggests that some participants overestimated and others underestimated their performance. Likewise, equal numbers of participants changed their estimations upwards and downwards after the review phase, suggesting that the displays did not bias estimates in different directions depending on the condition. In line with this, the verbal data show limited differences between conditions. No differences between conditions were found in the number of survey statements, the number of actions reported and the number of monitoring statements made. Furthermore, Bayesian analyses provide substantial support for a lack of differences. Only one difference between conditions was found, which was that students in the gaze-display condition reported more use of map features than the control condition.

A potential explanation for why the gaze displays did not improve monitoring and performance, might be that perhaps the specific gaze display that we used did not provide the information that participants needed to improve their monitoring accuracy and performance on this particular task. Earlier research shows that gaze displays are effective for improving performance if participants can extract information regarding the effectiveness of recognizing visual features from them (Donovan et al., 2008; Kundel et al., 1990), but not if they convey information about search strategy only (Dickinson & Zelinsky, 2005; Donovan et al., 2005; Drew & Williams, 2017; Peltier & Becker, 2017). Information about the effectiveness of recognizing visual features can be conveyed, for example, by scaling the size of fixations to their duration (cf. Donovan et al., 2008). We expected the gaze display to help participants gain insight in particular into the effectiveness of their recognition of visual features. However, far more remarks in the verbal reports related to the search strategy (85%–90% of all reported actions) than to the recognition of the visual features. Thus, participants probably interpreted the displays as mostly showing the search strategy instead of the recognition of visual features, and therefore the gaze display did not improve performance.

This could explain the difference between our findings and those of Kostons et al. (2009). They found that lower-expertise participants in the gaze-display condition reported more than twice as many actions than participants in the control condition, and higher-expertise participants made more monitoring/assessment statements in the gaze-display versus the control condition. A marked difference between their work (Kostons et al., 2009) and this study is the type of tasks. The information conveyed in the gaze display probably provided input on how problem-solving steps were executed, which is predictive for actual task performance in their study, but not in ours (as, additionally, the recognition of visual features is important).

Another explanation for the lack of benefits from the gaze displays could be that interpreting gaze displays in terms of cognitive/learning processes is too difficult for learners. Anecdotally, many students remarked on the difficulty of either remembering what they did or remembering why they did what they did. So what makes it difficult for learners to interpret gaze displays in terms of cognitive/learning processes? On the one hand, it might be that participants have trouble extracting information from the display that informs their monitoring judgement. Participants might not have a good idea of correct and wrong gaze behaviour, or how the gaze displays reflect correct and wrong problem-solving behaviour. Whereas several studies have shown that people can interpret gaze displays in terms of (other persons') cognitive or attentional processes (Bahle et al., 2017; Foulsham & Lock, 2015; van Wermeskerken et al., 2018; Zelinsky et al., 2013), performance in those studies is often still far from perfect and research by Greene et al. (2012) showed that people were unable to interpret a gaze display in terms of task performance.

On the other hand, it might be the case that information that could inform the monitoring judgement is not present in the gaze display (or not salient enough to be detected). For example, gaze displays were found to be useful when areas of prolonged attention were flagged (Donovan et al., 2008; Kundel et al., 1990). This was based on the finding that differences in fixation durations were found between correct and incorrect feature recognition (Kundel et al., 1978; Manning et al., 2006). In this study, we do not have information on effective and ineffective feature recognition on the level of individual features, so we cannot check that longer fixation durations are indeed associated with problems in feature recognition. Therefore, it could be the case that participants did indeed use this heuristic to evaluate their performance, but that on an individual level, this relationship was not present or not salient enough for participants to use it.

Further research could investigate how participants interpret gaze displays (i.e. what cues people think they convey), and if this information indeed relates to task performance. Such information could be used to predict when gaze displays do and do not support monitoring and task performance.

It has to be noted that Bayesian analyses were used to quantify evidence in favour of the null hypothesis. Those analyses showed substantial evidence that there was in fact no difference in monitoring accuracy between conditions. Likewise, there is substantial evidence that there is no difference between conditions in the number of comments in each of the categories, except for the use of map features and the number of positive monitoring remarks. For task performance, however, there is uncertainty (i.e. ignorable evidence in favour of the null hypothesis) as to whether there was an effect of the gaze displays on performance, and thus follow-up research with a larger sample size is required to understand whether or not a gaze display impacts post-test performance.

7.1 | Limitations

An important limitation of this study is that we only used one type of visualization, the dynamic version of the scanpath visualization of SMI. This visualization was chosen because it allows for seeing the sequence of fixations (Blascheck et al., 2014), as well as locations that received prolonged attention. It is very similar to the visualizations used in earlier studies (Kostons et al., 2009; Van Gog et al., 2005). However, it has been argued that scanpath visualizations result in too much visual clutter to find patterns (Blascheck et al., 2014). This was avoided by using a trail of only 2 s. However, we cannot rule out that another type of visualization such as an attention map (a visualization that shows areas that received prolonged attention without showing separate fixations, see for example, Blascheck et al., 2014) could have had better affordances for supporting the review of performance. It can also not be ruled out that the dynamic scanpath visualization did not provide the information necessary for the participants. Further research could investigate the effects of different types of visualizations.

Moreover, we should note that the choice for this particular task and this particular sample might have impacted the outcomes of the study, and thus it cannot be concluded that gaze displays will be ineffective in general. They might be beneficial for other tasks and groups. For example, the participants in our sample were self-selected, as we recruited participants who were interested in learning about map reading. A theoretical understanding of which types of visualizations can support which learners for which tasks is currently lacking, so this is an important avenue for further research. Another limitation that is inherent to the task we used, is that we cannot verify whether all remarks made in the verbal protocols were correct. Remarks could refer to cognitive actions that happened during task performance, or could have been fabricated by participants during the review process. We provided a non-directive prompt in the review phase, which is thought to avoid the fabrication of thoughts as much as possible (Ericsson & Simon, 1993). The analysis of verbal protocols, however, should not be interpreted as a complete overview of all cognitive processes that participants executed, but rather as an overview of what they reported about their viewing behaviour.

Finally, in line with Kostons et al. (2009), the instruction that participants received before the review phase was minimal and did not provide information on how the gaze displays should be used to evaluate performance. This was done to avoid confounding the current findings with the effects of different instructions. However, it might be that instruction on how to use the gaze displays to review performance make them more effective, and this is another relevant direction for further research. In the same vein, it would be relevant to collect data on how participants interpret gaze displays (cf. Knoop-van Campen et al., 2021), because those interpretations are likely to influence the impact of the gaze displays on performance.

8 | CONCLUSION

In conclusion, reviewing performance on a navigational map-reading task with a gaze display did not result in better monitoring accuracy than reviewing it without a gaze display. No effects were found on monitoring accuracy and performance, and the verbal data analysis did not reveal differences in the number of action and monitoring statements. Earlier research found that gaze displays impact performance mostly when information about the interpretation of visual features is shown (Donovan et al., 2008; Kundel et al., 1990). Our findings suggest that the same may be true for gaze displays that aim to support process monitoring. Participants were found to extract mostly information about their search strategy from the display and did not extract that much information about the effectiveness of their interpretation of visual features. However, the interpretation of visual features is often the main bottleneck for task performance (Kok et al., 2016; Kramer et al., 2019; van Geel et al., 2017), and thus this information was needed to monitor task performance. The finding that participants did not extract much information about the interpretation of visual features could explain why participants' monitoring did not improve.

New technological tools such as eye tracking provide promising options to support students and teachers in educational practice. In the present task, gaze displays offered the potential to provide information on learners' attentional processes that would otherwise have remained covert. This did not seem helpful in the present study, yet gaze displays might be effective when implemented in other tasks. For example, Kostons et al. (2009) found that participants made more monitoring remarks in the context of gaze displays that showed how they executed problem-solving steps, and the number of correctly executed problem-solving steps was predictive of learning. Therefore, future research should continue to investigate the usefulness of different types of gaze displays for improving monitoring and performance in different visual tasks.

ACKNOWLEDGMENTS

The authors would like to thank Marja Erisman and Jolanda Scholman for help with transcribing, segmenting, and coding the protocols. During the realization of this study, Ellen Kok was funded by an NRO PROO grant (Project 405-17-301).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/jcal.12666.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Dataverse at https://doi.org/10.34894/9KCFB6.

ORCID

Ellen Kok https://orcid.org/0000-0001-9752-2531

REFERENCES

Aizenman, A., Drew, T., Ehinger, K. A., Georgian-Smith, D., & Wolfe, J. M. (2017). Comparing search patterns in digital breast tomosynthesis and full-field digital mammography: An eye tracking study. *Journal of Medical Imaging*, *4*(4), 045501. https://doi.org/10.1117/1.JMI.4.4.045501

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 (R. Borgo, R. Maciejewski, & I. Viola, Eds.), *Proceedings of the Eurographics conference on visualization (EuroVis)*. Eurographics Association. https://doi.org/10.2312/eurovisstar.20141173

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

Clarke, A. D. F., Mahon, A., Irvine, A., & Hunt, A. R. (2016). People are unable to recognize or report on their own eye movements. *The Quarterly Journal of Experimental Psychology*, *70*(11), 2251–2270. https://doi.org/10.1080/17470218.2016.1231208

Dickinson, C. A., & Zelinsky, G. J. (2005). Marking rejected distractors: A gaze-contingent technique for measuring memory during search. *Psychonomic Bulletin & Review*, *12*(6), 1120–1126. https://doi.org/10.3758/bf03206453

Dong, W., Jiang, Y., Zheng, L., Liu, B., & Meng, L. (2018). Assessing map-reading skills using eye tracking and Bayesian structural equation modelling. *Sustainability*, *10*(9), 3050. https://doi.org/10.3390/su10093050

Donovan, T., Manning, D. J., & Crawford, T. (2008). Performance changes in lung nodule detection following perceptual feedback of eye movements. *Proceedings of the SPIE 6917, Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment* (Vol. *6917*, pp. 9). https://doi.org/10.1117/12.768503

Donovan, T., Manning, D. J., Phillips, P. W., Higham, S., & Crawford, T. (2005). The effect of feedback on performance in a fracture detection task. *Proceedings of SPIE* The International Society for Optical Engineering. https://doi.org/10.1117/12.593294

Drew, T., & Williams, L. H. (2017). Simple eye-movement feedback during visual search is not helpful. *Cognitive Research: Principles and Implications*, *2*(1), 44. https://doi.org/10.1186/s41235-017-0082-3

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003

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*, *26*, 159–181. https://doi.org/10.1007/s10459-020-09975-w

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge: MIT Press.

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

Green, S. R., & Redford, J. S. (2016). Metasearch accuracy for letters and symbols: Do our intuitions match empirical reality? *Metacognition and Learning*, *11*(2), 237–256. https://doi.org/10.1007/s11409-015-9143-5

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, T., Mielicki, M., & Wiley, J. (2019). Improving students' metacomprehension accuracy. In J. Dunlosky & K. A. Rawson (Eds.), *Cambridge handbook of cognition and education* (pp. 619–646). Cambridge University Press.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

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

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, *24*(5), 539–556. https://doi.org/10.1037/met0000201

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*. 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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

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., Võ, 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., & Jarodzka, H. (2017). 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., de Bruin, A. B. H., BinAmir, H. A. N., Robben, S. G. F., & van Merriënboer, J. J. G. (2016). Systematic viewing in radiology: Seeing more, missing less? *Advances in Health Sciences Education*, *21*(1), 189–205. https://doi.org/10.1007/s10459-015-9624-y

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

Kramer, M. R., Porfido, C. L., & Mitroff, S. R. (2019). Evaluation of strategies to train visual search performance in professional populations. *Current Opinion in Psychology*, *29*, 113–118. https://doi.org/10.1016/j.copsyc.2019.01.001

Kundel, H. L., Nodine, C. F., & Carmody, D. (1978). Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative Radiology*, *13*(3), 175–181.

Kundel, H. L., Nodine, C. F., & Krupinski, E. A. (1990). Computer-displayed eye position as a visual aid to pulmonary nodule interpretation. *Investigative Radiology*, *25*(8), 890–896.

Manning, D., Barker-Mill, S. C., Donovan, T., & Crawford, T. (2006). Time-dependent observer errors in pulmonary nodule detection. *The British Journal of Radiology*, *79*(940), 342–346. https://doi.org/10.1259/bjr/13453920

Marti, S., Bayet, L., & Dehaene, S. (2015). Subjective report of eye fixations during serial search. *Consciousness and Cognition*, *33*, 1–15. https://doi.org/10.1016/j.concog.2014.11.007

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

Nelson, T. O., & Nahrens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, *26*, 125–173. https://doi.org/10.1016/S0079-7421(08)60053-5

O'Meara, P., Munro, G., Williams, B., Cooper, S., Bogossian, F., Ross, L., Sparkes, L., Browning, M., & McClounan, M. (2015). Developing situation awareness amongst nursing and paramedicine students utilizing eye tracking technology and video debriefing techniques: A proof of concept paper. *International Emergency Nursing*, *23*(2), 94–99. https://doi.org/10.1016/j.ienj.2014.11.001

Otto, K., Castner, N., Geisler, D., & Kasneci, E. (2018). Development and evaluation of a gaze feedback system integrated into eyetrace. *ACM Symposium on Eye Tracking Research & Applications* (pp. 1–5). https://doi.org/10.1145/3204493.3204561

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*(422). https://doi.org/10.3389/fpsyg.2017.00422

Peltier, C., & Becker, M. W. (2017). Eye movement feedback fails to improve visual search performance. *Cognitive Research: Principles and Implications*, *2*(1), 47. https://doi.org/10.1186/s41235-017-0083-2

Putto, K., Kettunen, P., Torniainen, J., Krause, C. M., & Tiina Sarjakoski, L. (2014). Effects of cartographic elevation visualizations and map-reading tasks on eye movements. *The Cartographic Journal*, *51*(3), 225–236. https://doi.org/10.1179/1743277414Y.0000000087

Qvarfordt, P., Biehl, J. T., Golovchinsky, G., & Dunningan, T. (2010). Understanding the benefits of gaze enhanced visual search. *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (pp. 283–290). https://doi.org/10.1145/1743666.1743733

Redford, J. S., Green, S. R., Geer, M., Humphrey, M., & Thiede, K. W. (2011). Exploring metacognitive accuracy in visual search. *Memory & Cognition*, *39*(8), 1534–1545. https://doi.org/10.3758/s13421-011-0123-y

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059–1074. https://doi.org/10.1068/p281059

Sommer, S., Hinojosa, L. & Polman, J. L. (2016). Utilizing eye tracking technology to promote Students' metacognitive awareness of visual STEM literacy. (C. K. Looi, J. Polman, U. Cress, & P. Reimann, Eds.), Transforming learning, empowering learners: The International Conference of the Learning Sciences (pp. 1231-1232).

Strijbos, J.-W., Martens, R. L., Prins, F. J., & Jochems, W. M. (2006). Content analysis: What are they talking about? *Computers & Education*, *46*(1), 29–48. https://doi.org/10.1016/j.compedu/2005.04.002

Tsai, P.-Y., Yang, T.-T., She, H.-C., & Chen, S.-C. (2019). Leveraging college students' scientific evidence-based reasoning performance with eye-tracking-supported metacognition. *Journal of Science Education and Technology*, *28*(6), 613–627. https://doi.org/10.1007/s10956-019-09791-x

van der Gijp, A., Ravesloot, C. J., van der Schaaf, M. F., van der Schaaf, I. C., Huige, J., Vincken, K. L., Ten Cate, O. T. J., & van Schaik, J. P. J. (2015). Volumetric and two-dimensional image interpretation show different cognitive processes in learners. *Academic Radiology*, *22*(5), 632–639. https://doi.org/10.1016/j.acra.2015.01.001

van Geel, K., Kok, E. M., Dijkstra, J., van Merriënboer, J. J. G., & Robben, S. G. F. (2017). Teaching systematic viewing to final-year medical students improves systematicity but not coverage or detection of radiologic abnormalities. *Journal of the American College of Radiology*, *14*(2), 235–241. https://doi.org/10.1016/j.jacr.2016.10.001

van Gog, T., Kester, L., Nievelstein, 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 Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*(2), 147–177. https://doi.org/10.1007/s10648-005-3951-0

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., Paas, F., Van Merriënboer, J. J., & Witte, P. (2005). Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology-Applied*, *11*(4), 237–244. https://doi.org/10.1037/1076-898x.11.4.237

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 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

Võ, M. L. H., Aizenman, A. M., & Wolfe, J. M. (2016). You think you know where you looked? You better look again. *Journal ofExperimental Psychology: Human Perception and Performance*, *42*(10), 1477–1481. https://doi.org/10.1037/xhp0000264

Wilson, M. R., Vine, S. J., Bright, E., Masters, R. S. W., Defriend, D., & McGrath, J. S. (2011). Gaze training enhances laparoscopic technical skill acquisition and multi-tasking performance: A randomized, controlled study. *Surgical Endoscopy and Other Interventional Techniques*, *25*(12), 3731–3739. https://doi.org/10.1007/s00464-011-1802-2

Xie, H., Zhao, T., Deng, S., Peng, J., Wang, F., & Zhou, Z. (2021). Using eye movement modelling examples to guide visual attention and foster cognitive performance: A meta-analysis. *Journal of Computer Assisted Learning*, *37*(4), 1194–1206. https://doi.org/10.1111/jcal.12568

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

Zhai, X., Fang, Q., Dong, Y., Wei, Z., Yuan, J., Cacciolatti, L., & Yang, Y. (2018). The effects of biofeedback-based stimulated recall on self-regulated online learning: A gender and cognitive taxonomy perspective. *Journal of Computer Assisted Learning*, *34*(6), 775–786. https://doi.org/10.1111/jcal.12284

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, *41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

How to cite this article: Kok, E., 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*, *38*(4), 1087–1101. https://doi.org/10.1111/jcal.12666