Received: 26 June 2019 | Revised: 26 June 2020 | Accepted: 5 July 2020

DOI: 10.1002/acp.3721

WILEY

RESEARCH ARTICLE

Inferring task performance and confidence from displays of eye movements

Selina N. Emhardt[1] | Margot van Wermeskerken[2] | Katharina Scheiter[3,4] | Tamara van Gog[5]

[1]Department of Educational Sciences, Open Universiteit Nederland (Open University of the Netherlands), Heerlen, The Netherlands

[2]Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands

[3]Leibniz-Institut für Wissensmedien, Tübingen, Germany

[4]University of Tübingen, Tübingen, Germany

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

Correspondence
Selina N. Emhardt, Faculty of Educational Sciences, Open Universiteit Nederland (Open University of the Netherlands), 6419 AT Heerlen, The Netherlands.
Email: selina.emhardt@ou.nl

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: 452-11-006

Summary

Eye movements reveal what is at the center of people's attention, which is assumed to coincide with what they are thinking about. Eye-movement displays (visualizations of a person's fixations superimposed onto the stimulus, for example, as dots or circles) might provide useful information for diagnosing that person's performance. However, making inferences about a person's task performance based on eye-movement displays requires substantial interpretation. Using graph-comprehension tasks, we investigated to what extent observers (N = 46) could make accurate inferences about a performer's multiple-choice task performance (i.e., chosen answer), confidence, and competence from displays of that person's eye movements. Observers' accuracy when judging which answer the performer chose was above chance level and was higher for displays reflecting confident performance. Observers were also able to infer performers' confidence from the eye-movement displays; moreover, their own task performance and perceived similarity with the performer affected their judgments of the other's competence.

KEYWORDS

eye tracking, gaze interpretation, instructional design, performance assessment

1 | INTRODUCTION

Eye movements reveal what is at the center of a person's visual attention and thereby have the potential to provide valuable information about that person's ongoing cognitive processes during task performance. The idea that cognitive processes are associated with eye movements is widely accepted, and is based on two assumptions. First, the eye-mind assumption states that what is processed at a perceptual level, is also processed at a cognitive level. Second, the immediacy assumption states that what we are looking at is immediately processed (Just & Carpenter, 1980).

Because attention and cognition are so tightly linked, displays of eye movements might be useful tools for diagnosing task performance. Modern eye-tracking technology (see for example, Holmqvist et al., 2011) is becoming increasingly affordable and easy to use. Moreover, the technology allows not only for recording but also for visualizing a person's eye movements, for instance, by displaying the person's current focus of attention superimposed onto the processed stimulus. Technically, the focus of attention is determined by assessing which information is fixated (i.e., attended to) and hence processed at a given moment. In eye-movement visualizations, fixations can be represented as circles or dots overlaid on the area of the stimulus that is currently attended. The size of the circles or dots can be varied depending on the duration of the fixation (i.e., the larger the diameter, the longer the fixation). The fixations can either be displayed all at once (statically) or dynamically in the order of appearance, which makes the temporal

aspects of information processing visible. Such eye-movement displays are already being used in education research to visualize an expert's gaze patterns to guide learners' attention (i.e., Eye-Movement Modeling Examples, see for example, Jarodzka et al., 2012, Jarodzka, van Gog, Dorr, Scheiter, & Gerjets, 2013; Mason, Pluchino, & Tornatora, 2015; Scheiter, Schubert, & Schüler, 2018; Van Marlen, van Wermeskerken, Jarodzka, & van Gog, 2018).

However, inferring cognitive processes from eye-movement displays requires substantial interpretation. It amounts to assigning meaning to a pattern of circles or dots (representing fixations) that are overlaid statically or dynamically on the original task material. If observers would be able to interpret eye-movement displays in terms of the cognitive processes the task performer is engaging in, then it could be a useful tool, for instance, for teachers to get more detailed insight into their students' reasoning. Teachers often rate students' performance based on their answer choice. The students' processes of decision-making are, in contrast, not directly observable for the teacher and it has been long known that verbal self-reports of task performers (here students) are far from perfect (Ericsson & Simon, 1980). Eye-movement displays might provide a teacher with additional information about the students' performance that would otherwise remain inaccessible to them, such as which options they considered and for how long, prior to answering the question (which could be indicative of the student's knowledge and confidence). As such, providing teachers with this information on students' eye movements could allow them to give more adaptive feedback to their students—provided they are able to make meaningful inferences from those displays.

In this study, we investigated if observers can infer other people's task performance (i.e., what answer they chose in a multiple-choice task) and their confidence in task performance from dynamic video displays of the performer's eye movements (eye-movement displays). This investigation also contributes to our fundamental knowledge about the human ability to infer cognitive processes from eye-movement displays.

1.1 | Evidence that eye-movement displays reflect performance

One prerequisite for being able to interpret eye-movement displays in terms of a person's task performance is that such displays differ as a function of the type of task and quality of task performance (such as confidence during performance). Yarbus (1967) provided early evidence for this assumption by showing that different task instructions yielded different eye-movement displays. In his seminal study, a subject's eye movements were recorded while inspecting a painting ("The Unexpected Visitor" by Ilya Repin, 1883) under different viewing instructions (e.g., "Give the ages of the people", or "Remember the position of the people and objects in the room"). Eye-movement displays differed substantially between the different instruction conditions; the displays of eye movements reflected the cognitive process the person was engaged in (see also for example, Bahle, Beck, & Hollingworth, 2018; Borji & Itti, 2014; Castelhano, Mack, & Henderson, 2009; DeAngelus & Pelz, 2009; Tatler, Wade, Kwan, Findlay, & Velichovsky, 2010).

As for the quality of performance (e.g., confident task performance), eye-movement patterns of experts or people with high prior knowledge have been found to systematically differ from novices and people with lower prior knowledge. As such, eye-movement displays may also contain information about the task performer's competence and confidence (e.g., a more confident performance of more competent performers). For instance, people with higher levels of expertise have been found to attend to task-relevant information more often and longer (e.g., Cooper, Gale, Darker, Toms, & Saada, 2009; Haider & Frensch, 1999; Jaarsma, Jarodzka, Nap, van Merrienboer, & Boshuizen, 2014), and to show longer saccade lengths (e.g., Charness, Reingold, Pomplun, & Stampe, 2001; Reingold, Charness, Pomplun, & Stampe, 2001). In addition, they double-checked their answers less often (Jaarsma et al., 2014).

Regarding multiple-choice task performance, it has been shown that when deciding between several possible answer options, an attentional bias (i.e., gaze bias; Shimojo, Simion, Shimojo, & Scheier, 2003) toward the preferred option is observed (Lindner et al., 2014; see also Foulsham & Lock, 2015; Glaholt, Wu, & Reingold, 2009). Hence, eye-movement displays can be expected to reflect several aspects of task performance.

1.2 | The ability to interpret eye-movement displays

The question is, however, whether observers can infer different aspects of task performance from eye-movement displays, which is the prerequisite for later using eye-movement displays as a diagnostic tool in education or other training situations. To do so, an observer needs to make assumptions about why the other person is looking at a specific location at a specific time and what this performer is thinking while doing so. So far, findings regarding the ability to make such inferences from eye-movement displays are promising. However, studies have mainly focused on relatively simple aspects of task performance. We present these studies in the following sections. Moreover, as will become clear from the descriptions below, these studies used different ways of designing eye-movement displays.

A study by Greene, Liu, and Wolfe (2012) suggested that observers are not able to interpret eye-movement displays in terms of the underlying task. In their study, participants had to inspect various photographs under different viewing instructions while their eye movements were recorded. The static eye-movement displays (images containing 10 or 60 s of eye-movement data) were then presented to a group of observers. Fixations were visualized as colored dots (size irrespective of fixation duration) that were connected with lines (i.e., representing saccades). Unexpectedly, the observers were not able to infer which viewing instruction was reflected in the eye-movement displays. However, it might be that the specific task instructions that guided generation of the eye-movement displays yielded eye-movement patterns that were not sufficiently distinct from each other when visualized (Borji & Itti, 2014).

In contrast to Greene et al. (2012), Zelinsky, Peng, and Samaras (2013) found that participants *could* make valid inferences

about the underlying viewing instruction based on another person's eye-movement displays. In their study, observers had to identify the target of other people's visual search based on their static eye-movement displays. A first group of participants had to search for a specific target (a bear or a butterfly) among three distractor objects that varied in their similarity (high, medium, or low) to the target. Then, a second group of participants saw static eye-movement displays of this search process and had to determine for which target category the other person was searching. The object fixated first was marked by a green circle (all others were red), and the circles had four different sizes, depending on the time the performers spent looking at each object (the largest circle indicated the longest fixation). Blue trail lines connected the circles and indicated the order of saccades between the fixations. Based on the eye-movement displays of target-absent trials the observers could correctly judge with high accuracy for which targets the performers had searched. Presumably, the observers made use of the visual similarity between the objects that were fixated more often during search and the target objects (target-distractor similarity) for making their judgments.

In the study by Van Wermeskerken, Litchfield, and van Gog (2018) participants observed the painting "The Unexpected Visitor" under three different instructions (i.e., estimate the ages of the people in the painting, remember the positions of the objects in the room, and estimate how long the unexpected visitor had been away from the family; cf. Yarbus, 1967). While doing so, their eye movements were recorded. Afterwards, participants were shown either static or dynamic eye-movement displays of themselves or another person. In the first two experiments, fixations were displayed by yellow circles of constant size with consecutive fixations being connected by a yellow line. Dynamic eye-movement displays showed a moving yellow circle of constant size without trails (connection lines). Observers were able to recognize the instruction reflected in eye-movement displays above chance level (regardless of whether it was their own or someone else's, Experiment 1: $ds \geq 1.38$; Experiment 2: $ds \geq 2.04$). Furthermore, instruction recognition performance was higher for dynamic than for static eye-movement displays (Experiment 1: $d = 0.63$; Experiment 2: $\eta_p^2 = 0.08$). In a third experiment, the authors investigated the role of order information for instruction recognition performance by comparing dynamic displays (i.e., full order information); static displays with lines between consecutive fixations (i.e., limited order information); and static displays without lines between consecutive fixations (i.e., without order information). Temporal information was provided in this third experiment by displaying smaller circles for shorter fixations and larger circles for longer fixations. The dynamic condition with full order information led to higher instruction recognition performance than the static conditions, but the two static conditions did not differ.

Furthermore, there is some evidence that observers can interpret dynamic eye-movement displays with respect to answer preference. For instance, Foulsham and Lock (2015) presented participants with four colorful patterns and they had to select the pattern they preferred most while their eye movements were recorded ("truth" trials). Subsequently, participants were shown dynamic eye-movement displays of another person allegedly performing the same task (i.e., selecting a preferred pattern) and had to indicate which of the four patterns that person preferred. This study used a red dot of constant size that moved across the answer options to indicate the performers' fixations at each point in time. Finally, participants were shown similar displays with four patterns, but were instructed to deliberately hide their preference from future observers of their eye-movement displays ("lie" trials). In the "lie" trials, the fixations were more evenly distributed across the four patterns, whereas in the "truth" trials, fixations were directed relatively more at the preferred pattern. Results indicated that participants could infer which pattern another person preferred from eye-movement displays of "truth" trials above chance-level (all $ds > 1.2$), but not from eye-movement displays of "lie" trials. This finding suggests that the above-mentioned gaze bias plays a role in the interpretability of eye-movement displays, with accuracy declining when fixations are more evenly distributed among options.

A study by Van Wermeskerken, Litchfield, and van Gog (submitted) took this line of inquiry a step further using a more complex task. Observers were shown short (i.e., 10 s) dynamic and static eye-movement displays of performers who were solving relational reasoning tasks (multiple-choice, see Alexander, Dumas, Grossnickle, List, & Firetto, 2016) and had to decide which answer option the performer had chosen. Fixations were displayed as red circles and their diameter depended on the duration of the fixation (e.g., for a fixation of 500 ms it was 80 px). In the static displays, the fixations were connected with a line. In the dynamic displays, no trail (connection line) was visible. Eye-movement displays in this study were furthermore either high or low in *distinctiveness*. A higher distinctiveness means that the performer focused relatively more on the chosen answer option than on the other answer options (and this would be visible in the eye-movement displays). In contrast, a lower distinctiveness means that the performer's fixation patterns were more diffuse, that is, more evenly distributed across the possible answer options (and this would be visible in the eye-movement displays). In general, observers' accuracy of judging which answer option was chosen by the performer was above chance level. Observers' judgment accuracy was higher for high-distinctive eye-movement displays than for low-distinctive displays (Experiment 1: $\eta_p^2 \geq .492$; Experiment 2: $\eta_p^2 = .749$; Experiment 3: $\eta_p^2 \geq .681$).

The fact that distinctiveness seems to play a role in accuracy is interesting, especially because distinctiveness is (presumably) also affected by the performer's confidence. High confidence would lead to stronger focus on the chosen answer and less double-checking, whereas low confidence would result in consideration of various answer options and more double-checking and, hence, more evenly distributed fixations (cf. Jaarsma et al., 2014; Lindner et al., 2014). Thus, an open question is whether observers can detect this information from the gaze displays and, in turn, can make inferences about another person's confidence during performance. Therefore, the present study aims to partially replicate but also extend the findings of Van Wermeskerken et al. (submitted) by investigating not only how accurately the chosen answer can be inferred from dynamic eye-movement displays, but also whether confidence can be derived from it.

Moreover, we extend prior research by investigating whether the observers' own performance influences their interpretation of another person's performance. There are indications from social psychology that perceived similarity can bias proficiency ratings. In a study by Bates (2002) the competence of employees of several companies was rated on a questionnaire with items like "this individual and I are alike in terms of coming up with a similar solution for a work problem" or "this individual and I handle work problems in similar ways". Judgments about the performance of an employee were positively biased when the employee was perceived as being more similar to oneself. Furthermore, research on the *myside bias* shows that people evaluate other people in a manner biased toward their own opinions and attitudes (Stanovich, West, & Toplak, 2013), with the strength of the opinion predicting the degree of bias (Stanovich & West, 2008). Most studies that confirm the existence of the myside bias used tasks that require complex argument evaluation on controversial topics like abortion or gun control (e.g., Stanovich & West, 2008; Taber & Lodge, 2006; Wolfe, 2012). Until now, the influence of observers' own beliefs about the correct task solution and confidence on the interpretation of eye-movement displays has not been investigated. If a kind of myside bias would also apply to the interpretation of perceived performance similarity based on eye-movement displays (with the observer judging the performer as being more competent when they choose the same answer), this would be relevant for future applications in education. For instance, it might imply that a teacher could reliably judge a student's competence from an eye-movement display, but other students might erroneously judge a student to be competent just because they choose the same answer, when in fact both demonstrated incompetent performance.

1.3 | The present study

We used multiple-choice graph interpretation tasks, each consisting of a bar or line graph, a problem statement, and four answer options. Prior studies mostly focused on people's ability to make inferences from eye-movement displays of performers who solved short multiple-choice tasks (e.g., Van Wermeskerken et al., submitted; Foulsham & Lock, 2015; Zelinsky et al., 2013). We aimed to extend this line of research by using more complex and more educationally relevant multiple-choice tasks. Graphical illustrations of data are ubiquitous in everyday life. However, graphical literacy skills are, even for adults, often not well developed. This makes graph comprehension a challenging task (Shah & Hoeffner, 2002). Graph comprehension consists of several processes such as visual pattern recognition, pattern interpretation, and information integration (e.g., labels and titles and the graphical illustration; Carpenter & Shah, 1998). Eye-movement displays of performers might reveal these processes. After solving each graph item themselves, the observers were asked to infer from dynamic displays of another performer's eye movements what answer option the other person selected, how confident the other person was in their performance (certainty about the correctness of the given answer choice), and how competent the other person was. Observers were provided with both correct and incorrect performances and high and low confidence performances.

In order to replicate and generalize previous findings (e.g., Van Wermeskerken et al., submitted; Foulsham & Lock, 2015; Van Wermeskerken et al., 2018), we first hypothesized that observers would demonstrate an answer judgment accuracy above chance level (Hypothesis 1). Second, prior research showed answer judgment accuracy to be affected by the distribution of attention over the answer options (e.g., Van Wermeskerken et al., submitted; Foulsham & Lock, 2015). We therefore expected that the observers' judgment accuracy of which answer was chosen would be higher when the eye-movement displays reflected a confident performance compared to an unconfident performance (Hypothesis 2). Furthermore, we explored whether answer judgment accuracy of the observers would differ as a function of correctness of the performance.

Third, it was hypothesized that the observers would be able to pick up on distinctiveness of eye-movement patterns as a cue regarding the performers' confidence (cf. Jaarsma et al., 2014; Lindner et al., 2014) and therefore rate the performers' confidence as higher when eye-movement displays reflected a confident compared to an unconfident performance (Hypothesis 3). The question whether confidence can be inferred from eye-movement displays has not been investigated in previous studies. Additionally, it was explored whether confidence ratings would differ as a function of correctness of the performance.

Fourth, based on findings regarding similarity and myside bias (e.g., Stanovich et al., 2013), we hypothesized that observers' competence judgments about a performer would be higher when the observers perceive a similarity in answering behavior between themselves and the performer. That is, when the observer chose answer option A and they also inferred that the performer chose answer option A, we expected them to rate the performer as being more competent. We expected that this effect would be stronger when the observers are confident about the correctness of their given answer (Hypothesis 4, see Stanovich & West, 2008).

2 | METHODS

2.1 | Participants and design

Participants were 46 undergraduates from a Dutch university (M_{age} = 22.59 years, SD = 5.15; 40 females). All participants had normal or corrected-to-normal vision, and received EUR 7.50 for their participation.

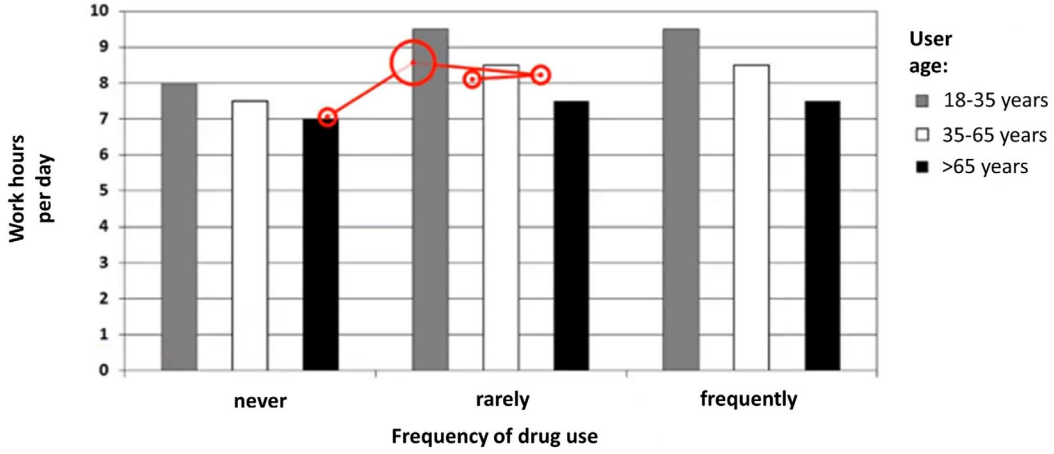
This study had a within-subjects design, so all participants viewed eye-movement displays with correct and incorrect task performances displaying confident as well as unconfident task performances. Post hoc power calculation with G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007) revealed that with a sample size of N = 46, we would be able to detect small to medium effect sizes of d = 0.34 and η^2 = .028 with a power of .80.

2.2 | Materials and procedure

Generation of eye-movement displays. Prior to the main study, eye movements of 17 performers (employees from a Dutch university; M_{age} = 27.47 years; SD $_{age}$ = 4.21; 11 females) were recorded to create the 32 eye-movement displays used in the main study.

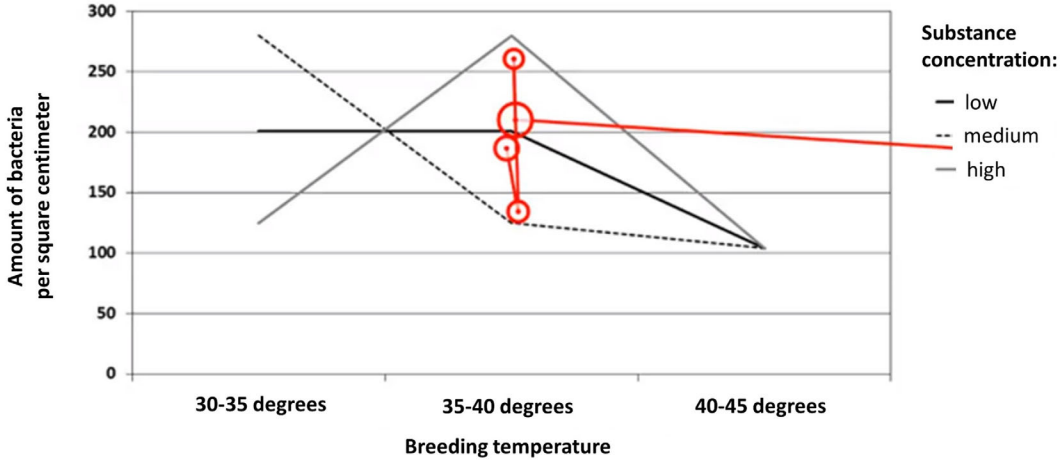
To this end, 24 graphs (12 bar graphs, 12 line graphs) with multiple-choice questions were designed in Microsoft Office Excel. Underneath the graph, we presented a multiple-choice question with four answer options (e.g., see Figure 1 for sample items). Each graph depicted the effects of two independent variables with three levels on one dependent variable. For instance, the top item of Figure 1 depicts the effects of the frequency of a drug usage (never, rarely, often) and the user age (18–35 years, 35–65 years and >65 years) on subjects' work hours per day.

The relationship between the two independent variables depicted in the graph was either no interaction (eight graphs), an ordinal interaction (eight graphs), or a disordinal interaction (eight graphs). Since prior knowledge could help viewers to keep track of the information depicted in graphs (Shah & Freedman, 2011), fictitious relationships between variables were used. The question was placed directly below the graph. In half of the questions, participants needed to calculate



For people who use drugs, the following applies regardless of their frequency of use: With increasing age of the users, the work hours per day...

a) decrease.	b) increase.	a) The work hours don't change.	d) Non of the options a), b), or c) are correct



If we average across all breeding temperatures until 40 degrees, we can conclude that the highest substance concentration leads, in comparison to both other concentrations, to...

a) more bacteria.	b) less bacteria.	a) the same amount of bacteria.	d) Non of the options a), b), or c) are correct

FIGURE 1 Examples of a bar graph (top) and line graph (bottom) with superimposed eye-movement displays in red

the average across different factor levels (e.g., bottom graph item in Figure 1). In the other half of the questions, they needed to compare the levels of a factor without averaging (e.g., top graph item in Figure 1). The four answer options A, B, C, and D (one option was correct) were presented in juxtaposition underneath the question. The images consisted of 1087 × 1050 pixels and were presented in the center of a 22″ monitor with a resolution of 1680 × 1050 pixels.

To record the eye movements, the 24 graphs (out of which eight recordings were later selected for the main experiment, see below) were presented to the group of performers using SMI Experiment Center software (version 3.7; SensoMotoric Instruments GmbH, Teltow, Germany). Eye movements were recorded binocularly at 250 Hz using a SMI RED250 infrared eye tracker (SensoMotoric Instruments GmbH, Teltow, Germany). The performers were instructed to first inspect the graph item and multiple-choice questions on the presented slide with a time limit of 40 s (based on completion times from a previous study with the same materials). To avoid the recording of meaningless eye movements, participants could proceed manually to the next slide as soon as they knew the answer. After inspecting each item, the participants had to indicate their answer by selecting one of the four options (A-D) on the next slide and rate how confident (or certain) they were that their answer was correct on a Likert scale from 1 (absolutely unconfident) to 7 (absolutely confident). The performance was then categorized in terms of correctness (correct vs. incorrect answer) and confidence (unconfident performance with confidence ratings between 1 and 3 and confident performance with ratings of 5-7). From these categories, suitable items for the main study were selected (i.e., using only graphs for which eye-movement recordings had sufficient calibration accuracy and for which eye-movement recordings were available for all). Based on these criteria, we selected eight of the 24 piloted graph items for the main study. For each of these items, we created four videos, one from each of the four conditions of the 2 (correct/incorrect) × 2 (unconfident/confident) design of the main study. As a result, each participant viewed 32 eye-movement display videos during the experiment.

The dynamic eye-movement displays were generated using SMI BeGaze software (Version 3.7; SensoMotoric Instruments), with the performers' fixations (i.e., lasting ≥50 ms and speed ≤40.0°/s) overlaid on the original material with the "Scan Path" option from SMI BeGaze. The fixations were displayed by a red circle with a line width of 5 px. The size of the red circle depended on the duration of each fixation. As long as the location was fixated, the circle gradually increased with a constant speed (e.g., the diameter of a fixation of 500 ms was 80 px). Consecutive fixations were connected through a red line (5 px). This trail faded out after 1.2 s. The frame rate of the output for the dynamic displays was set to 250 Hz. Figure 1 shows example screenshots of an eye-movement display for a bar and line graph item. Taken together, we decided on using settings for the eye-movement displays that were likely to facilitate gaze following and gaze interpretation. When using a moving circle, all non-fixated and fixated information is easily visible to the observers (in contrast to a spotlight or solid circle visualization). The size of the circle increased dynamically to make the information regarding the fixation durations very explicit. The trails highlighted the fixation order and caused a smoother eye-movement visualization than a circle that moves quickly from one fixation location to the next fixation location.

Table 1 shows some descriptive statistics from the 32 selected eye-movement displays (e.g., mean duration of the eye-movement display videos, mean confidence ratings of the performers). As mentioned in the Introduction, the distinctiveness of the eye-movement displays (i.e., the extent to which fixations are spread across answer options) could be an influential factor for the observers' answering behavior. Distinctiveness was determined by generating an area of interest around each answer option, calculating fixation times on each of the answer options, and then calculating the relative fixation time on the chosen answer option in relation to the other answer options. Distinctiveness values above zero indicate that fixations were more frequently directed at the chosen answer option than at the other answer options, and the higher the value, the more attention the chosen answer received compared to the other answers (see Table 1).

Procedure of the main experiment. The experiment was conducted using the online assessment tool Qualtrics (Qualtrics, Provo, UT). Participants were seated in front of a screen with an approximate viewing distance of 50 cm. The resolution of the screen was 1920 × 1,080 pixels. After signing the consent form, the participants answered personal questions about age, study field, and gender, and completed an example trial to become familiar with the procedure.

TABLE 1 Means and SD of the video durations, the confidence ratings of the performers and the distinctiveness of the eye-movement patterns in each condition

	Incorrect		Correct	
	Unconfident	Confident	Unconfident	Confident
Duration of performers' videos in seconds	35.62 (7.97)	27.61 (9.76)	39.09 (2.57)	32.69 (8.49)
Confidence of performers	2.43 (0.98)	6.13 (0.83)	2.63 (0.74)	5.75 (0.89)
Distinctiveness[a]	0.14 (0.41)	0.29 (0.27)	0.05 (0.35)	0.24 (0.39)

[a]To calculate the distinctiveness, the relative fixation duration that was spent on each of the four answer options (in %) was calculated for each item. From the answer option that was finally chosen by the participant, the other three relative durations of fixation (in %) were subtracted separately. The mean of these differences can be considered the distinctiveness (see also Van Wermeskerken et al. (submitted). The higher the distinctiveness measurement, the more relative time was spent fixating the *chosen* answer option. The distinctiveness could range from −0.33 to 1.0 with lower values indicating fewer fixations on the finally chosen answer option.

For each of the eight graph items, participants first completed the item themselves. Then they were asked to judge the dynamic eye-movement displays of four performers for the same item. More specifically, the participants always had to first inspect one graph (with a size of 1,299 × 811 pixels) with the corresponding multiple-choice question. After 40 s, the next slide was shown automatically on which participants selected their answer (A-D) and rated how confident they were that their answer was correct on a horizontally presented Likert scale ranging from 1 (absolutely unconfident) to 7 (absolutely confident). If participants knew the answer within less than 40 s, they could also proceed manually to the next slide. After solving an item, the participants were informed that they would now observe the eye-movement displays of four other people who performed the same graph task. One of four dynamic eye-movement displays corresponding to the previously solved item was then presented in the center of the subsequent page with a size of 800 × 600 pixels. There was no option to pause the video and the presentation proceeded automatically to the next slide after each video. Immediately after each eye-movement display video, participants had to indicate which multiple-choice option they thought the performer had chosen (A-D), how confident they thought the performer was in their answer, and how competent they perceived the performer to be (both on 7-point Likert scales from 1 [absolutely unconfident/incompetent] to 7 [absolutely confident /competent]). In total, the participants observed and judged four eye-movement display videos per graph item in a randomized order (correct-confident, correct-unconfident, incorrect-confident, incorrect-unconfident). Thus, they judged 32 eye-movement displays in total.

We designed four lists with randomly generated orders of the eight graph items. Within each list, bar and line graphs alternated and each participant was randomly assigned to one of the four lists and completed the experiment in approximately 45 minutes.

2.3 | Data analysis

The data collected from one of the eye-movement displays had to be excluded because of an error in classification. As a result, the data in the unconfident-incorrect condition are based on 7 eye-movement displays (instead of 8). The observers' *answer judgment accuracy* was determined by counting the correctly inferred answer options (A-D) per observer and condition and dividing that number by the total number of eye-movement displays in that condition (i.e., 7, in the unconfident-incorrect condition, or 8, in all other conditions). *Confidence and competence judgment scores* were determined by averaging each observer's ratings of the perceived confidence/perceived competence of the performers per condition.

To compute perceived similarity, the answer that the observer thought that the performer had chosen (A-D) was compared to the observer's own answer (A-D). *Perceived similarity* was scored as present (yes) when those answers matched and as absent (no) when the two answers did not match. For the analysis of variance (ANOVA) calculations, we used the *ez-package* for R (Lawrence & Lawrence, 2016) and for plotting the results we used the *ggplot2-package* (Wickham, 2009).

To explore whether *observers' confidence* affected inference making, observers' confidence was categorized in the same way as the performers' confidence: trials on which the confidence ratings about their own answers were between 1 and 3 on a 7-point Likert scale were categorized as unconfident and trials with confidence ratings between 5 and 7 were categorized as confident. Trials in which the observers provided confidence ratings of 4 (not confident/unconfident) were excluded (164 out of 1,426 trials, 11.5%), since the observers in those trials could not unambiguously be categorized. In total, 1,262 trial observations were used for the analysis of whether observers' competence ratings were affected by perceived similarity and own confidence (Hypothesis 4). Since not every participant (observer) showed each of the combinations of confidence (confident vs. unconfident) and perceived similarity (yes vs. no), a linear mixed effect model (lme model) was used. Such models are flexible in processing data sets with unbalanced cells and do not require averaging across participants (Baayen, Davidson, & Bates, 2008). The model was analyzed by using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) for R (R Core Team, 2015).

The following model was used to predict the perceived competence rating of the performer as judged by the observer *i* on item *j* (Y_{ij}):

$$Y_{ij} = \beta 0 + \beta 1 * \text{observers}'\text{confidence}_{ij} + \beta 2 * \text{perceived similarity}_{ij} + \beta 3 * \left(\text{observers}'\text{confidence}_{ij} \times \text{perceived similarity}_{ij}\right) + \upsilon 0 + \upsilon 1_j + \varepsilon_{ij}.$$

The model contained an intercept of the regression model (β0), the regression coefficient of the main effect of the *observer's confidence* (β1, confident vs. unconfident) about the correctness of their answer and the regression coefficient of the main effect of *perceived similarity* between the observer's and performer's answer (β2, yes vs. no). β3 was defined as the regression coefficient of the interaction effect between the variables *observers' confidence* and *perceived similarity*. Subjects ($\upsilon 0 \sim N(0, \sigma^2)$) and items ($\upsilon 1_j \sim N(0, \sigma^2)$) were specified as random effects. ε was defined as the error term of the equation ($\varepsilon \sim N[0, \sigma^2]$).

3 | RESULTS

All analyses were conducted in R (R Core Team, 2015) with a significance level of $\alpha = .05$.

For the following analysis, Cohen's *d* and the generalized η^2 and are reported as a measure of effect size, with $d = .20$ and $\eta^2 = .02$, $d = .50$ and $\eta^2 = .13$, and $d = .80$ and $\eta^2 = .26$ corresponding to small, medium, and large effects, respectively (Cohen, 1988). For the lme model R^2 was reported, as described in Hox, Moerbeek, and van de Schoot (2017). The data that support the findings of this study are available from the corresponding author upon reasonable request.

3.1 | Answer judgment accuracy

Our first hypothesis was that observers would be able to infer which answer option was chosen by the performers above the chance level of 25% (1 out of 4 answer options). The overall mean answer judgment accuracy was 54.28% (SD = 11.30). A one-sided one-sample *t*-test showed that this answer judgment accuracy was indeed significantly above chance level, $t(45) = 17.57$, $p < .001$, CI = [51.48; 100], $d = 2.59$. To test whether the above-chance accuracy would be independent of confidence in or correctness of the performance in the display, it was tested whether answer judgment accuracy was above chance in all conditions. Figure 2a shows the judgment accuracies. Separate one-sided one sample *t* tests (Bonferroni corrected alpha level of $\alpha = .0125$ for four comparisons) revealed that this was the case for all conditions (all *p*s < .001).

Our second hypothesis stated that the observers' answer judgment accuracy would be higher when eye-movement displays reflect a high confidence performance compared to a low confidence performance. Furthermore, we explored whether the correctness of the performance affected judgment accuracy. A 2 (correct vs. incorrect) × 2 (high vs. low confidence) repeated-measures ANOVA on the observers' answer judgment accuracy revealed a significant main effect of confidence, $F(1,45) = 108.68$, $p < .001$, $\eta^2 = 0.31$, indicating that the observers' judgment was more accurate when the displays showed high confidence than low confidence performance. There was no significant main effect of the correctness, $F(1,45) = 3.10$, $p = .085$, $\eta^2 = 0.01$, nor a significant interaction between confidence and performers' correctness, $F(1,45) = 2.60$, $p < .114$, $\eta^2 = 0.01$. Figure 2a illustrates these results.

3.2 | Inferring confidence of the performers

Our third hypothesis was that observers would be able to infer the performers' confidence and, hence, that the observers' ratings of the performers' confidence would be higher when eye-movement displays reflected high rather than low confidence performance. Again, we explored whether the correctness of the performers' answer would

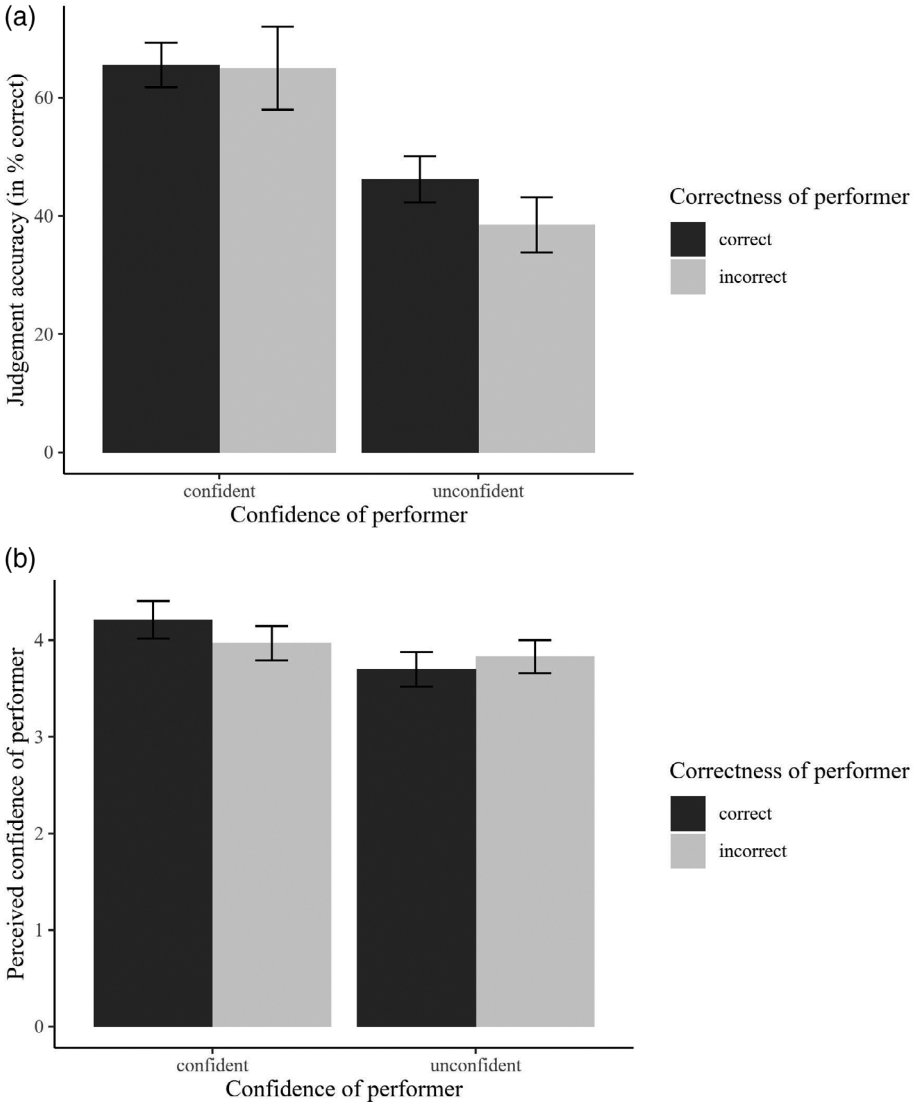


FIGURE 2 Observers' judgment accuracy (a) and estimates of performers' confidence (b) as a function of performers' actual confidence ratings (confident/unconfident) and correctness (correct/incorrect). The error bars display the standard errors of the means

affect confidence inferences. A 2 (correct vs. incorrect) × 2 (high vs. low confidence) repeated-measures ANOVA on the observers' mean confidence ratings revealed a main effect of confidence. Eye-movement displays of performers with high confidence ratings received higher confidence ratings than eye-movement displays of performers with low confidence, $F(1,45) = 22.43$, $p < .001$, $\eta^2 = 0.06$. There was no significant main effect of correctness, $F(1,45) < 1$, $p = .468$, $\eta^2 = 0.002$. However, we found a significant interaction effect between confidence and correctness, $F(1,45) = 6.03$, $p = .018$, $\eta^2 = 0.02$. Figure 2b illustrates these results.

Follow-up pairwise t tests (Bonferroni adjusted alpha = .025, given two pairwise comparisons) revealed an effect of confidence in the conditions with correct performances: confidence ratings in the correct-confident condition were significantly higher than in the correct-unconfident condition, $t(45) = 5.10$, $p < .001$, $d = 0.75$. However, ratings in the incorrect-confident condition did not differ significantly from the confidence ratings in the incorrect-unconfident condition, $t(45) = 1.36$, $p = .179$, $d = 0.20$.

3.3 | Perceived similarity: judging the other performers' competence

Our fourth hypothesis stated that the observers' competence judgments of the performers would be higher when the observers would perceive a similarity between themselves and the performers (i.e., when the observers thought the performer selected the same answer as they themselves had given). We expected that this effect would be stronger if the observers were confident about the correctness of their own answer.

In the final model ($R^2 = 0.23$), there was a significant main effect of the observers' confidence, indicating that observers with low confidence in their own performance perceived the performers' competence as being higher than observers with high confidence in their own performance, $\beta 1 = 0.47$, SE = 0.12, $t = 4.45$, $p < .001$. There was also a significant main effect of perceived similarity, indicating that when the observers thought the performers gave the same answer as they themselves had given (high perceived similarity), observers rated the competence of the performers as higher than when they thought the performers selected a different answer (low perceived similarity), $\beta 1 = 1.22$, SE = 0.10, $t = 12.74$, $p < .001$. These main effects were qualified by a significant interaction effect, indicating that the effect of perceived similarity was stronger, when the observers were confident about the correctness of their final choice, $\beta 3 = -0.88$, SE = 0.17, $t = -5.08$, $p < .001$. Table 2 provides the exact results of the model analysis and Figure 3 displays these findings.

TABLE 2 Summary of the model that describes the influence of similarity and own confidence on the competence rating

	Competence		
	B	**CI**	***p***
Fixed parts			
(Intercept)	3.58	3.35–3.80	<.001
Observers'confidence (unconfident)	0.47	0.26–0.68	<.001
Similarity (yes)	1.22	1.03–1.41	<.001
Observers' confidence × similarity	−0.88	−1.22 – −0.54	<.001
Random parts			
σ^2	1.734		
$\tau_{00,\ PPnum}$	0.134		
$\tau_{00,\ item}$	0.048		
N_{PPnum}	46		
N_{item}	8		
ICC_{PPnum}	0.070		
ICC_{item}	0.025		
Number of observations	1,262		
$R^2/\Omega_0{}^2$	.225/.222		

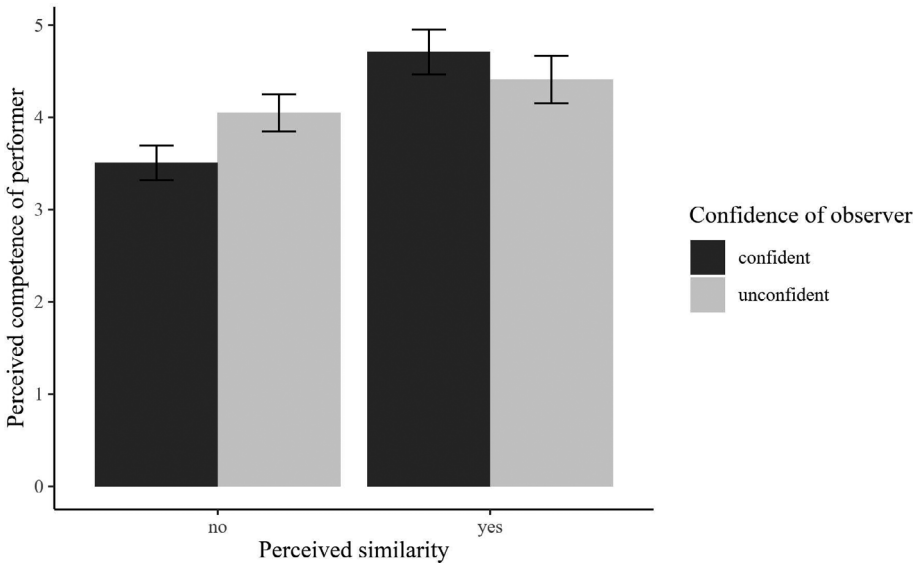


FIGURE 3 Observers' estimates of performers' competence as a function of observers' actual confidence ratings (confident/unconfident) and perceived similarity with the performers. The error bars display the standard errors of the means

4 | DISCUSSION

This study aimed to investigate to what extent observers can make inferences from other people's eye-movement displays. Furthermore, we explored which factors influence this inference making. After first solving each line or bar graph task themselves, the observers were asked to infer from dynamic displays of another performer's eye movements which multiple-choice answer option the performer had selected, and how confident and competent the performer was. Observers judged eye-movement displays of correct and incorrect performances in which performers had high and low confidence (i.e., all combinations were present in the set of displays).

4.1 | Can observers infer from eye-movement displays which answer option was chosen?

The first hypothesis was that observers would be able to infer what answer option (A-D) the performers chose above chance-level (i.e., 25%). In line with this hypothesis and previous findings (e.g., Van Wermeskerken et al., submitted), the overall answer judgment accuracy was above chance level in all conditions.

Prior research had shown that the distribution of attention over answer options affected observers' answer judgment accuracy (e.g., Van Wermeskerken et al., submitted) and that a performer's confidence in an answer affects the attention distribution over answer options (cf. Jaarsma et al., 2014; Lindner et al., 2014). Thus, our second hypothesis was that observers' judgment accuracy would be higher when the eye-movement displays reflected a high confidence performance than a low confidence performance. Indeed, our findings indicated that the chosen answer was more often accurately inferred from displays of confident performances than unconfident performances. This effect of performance confidence did not interact with correctness of the performance.

Looking at the characteristics of the eye-movement displays in the different conditions (Table 1), these findings strongly suggest that observers picked up on the fact that performers tend to gaze more toward the chosen answer (i.e., gaze bias effect; see for example, Foulsham & Lock, 2015; Glaholt et al., 2009; Lindner et al., 2014; Shimojo et al., 2003). Distinctiveness values above zero in Table 1 indicate that performers fixated more often on the chosen answer option than at the other answer options. The higher the distinctiveness value, the more attention the chosen answer received compared to the other answers. It seems safe to assume that observers (either consciously or unconsciously) used this distinctiveness for making their inference about the chosen answer. First, the fact that distinctiveness was above zero in all conditions, could explain why the judgment accuracy was above chance in all conditions. Second, the higher distinctiveness in displays of high confidence performances could explain why observers were better able to infer the chosen answer (i.e., higher judgment accuracy) in high confidence performances. Third, this could also explain why correctness of performance did not affect judgment accuracy, as the chosen answer was fixated most regardless of whether it was correct or not.

4.2 | Can observers infer from eye-movement displays how confident the performer was?

Thus far, prior research mainly focused on the observers' ability to accurately infer what task another person was engaged in (Van Wermeskerken et al., 2018; Zelinsky et al., 2013), or what answer option another person preferred (Foulsham & Lock, 2015) or chose (Van Wermeskerken et al., submitted). We extended this research by investigating whether observers of eye-movement displays would also be able to accurately infer the performers' confidence in their performance. In line with our third hypothesis, observers' rating of performers' confidence was higher when eye-movement displays reflected a confident compared to an unconfident performance. However, this effect was influenced by the correctness of the performance: Confidence ratings were more accurate for displays of correct compared to incorrect performances. This seems to suggest that observers do seem to pick up on confidence differences in correct performances, but not in incorrect performances.

Again, the distinctiveness values per condition (Table 1) can provide a possible explanation for this interaction. Even though there is a substantial difference in distinctiveness in the low and high confidence performances in both the correct and incorrect performance items, this difference is smaller in the incorrect performance items. Consequently, it may have been more difficult for participants to infer confidence (especially low confidence) in the incorrect items.

One potential limitation with regard to the confidence inferences is that Table 1 also shows that there was a time-on-task difference among the conditions, with high confidence performances being shorter than low confidence performances. We tried to prevent large time-on-task variations by imposing a maximum time limit of 40 s per task, but nevertheless participants could proceed earlier when they knew the answer. Note though, that if time-on-task would be used as a cue for confidence, one would expect the correct performances (which had longer durations) on average to have received lower perceived confidence ratings from the observers than incorrect performances (which had shorter durations), which was not the case. This suggests that the observers did use the displayed eye movements (and not solely the time-on-task) in making inferences about the performers' confidence.

4.3 | Does perceived similarity of observer and performer affect competence inferences?

We also investigated inferences about performers' competence, and expected that observers' own performance of the task would affect their rating about the performers' competence. Based on findings regarding similarity and myside bias (Stanovich et al., 2013;

Stanovich & West, 2008), we hypothesized that observers' competence judgments about a performer would be higher when the observers perceived a similarity in answering behavior between themselves and the performer. For instance, when the observer chose answer option A and they also inferred that the performer chose answer option A, they would rate the performer as being more competent. We expected that this effect would be stronger when the observers were confident about the correctness of their given answer. In line with this hypothesis, we found that competence ratings were higher when the observers chose the same answer as the performers (perceived similarity) than when they chose different answers (no perceived similarity) and that this effect was stronger when the observers were confident about the correctness of their own answer. This finding can be linked to the social-psychological effect of the myside bias (Stanovich et al., 2013), which shows that people often evaluate other people in a manner biased positively toward their own prior opinions (or in our case, their own prior answer). Furthermore, stronger own opinions (or in our case, higher confidence in their own answer) cause a greater myside bias (Stanovich & West, 2008). Studies about the myside bias traditionally deal with tasks that require complex argument evaluation on controversial topics like political decisions and involve affective ratings (e.g., Stanovich & West, 2008; Taber & Lodge, 2006; Wolfe, 2012). Our study extended this research and investigated if a kind of myside bias is also observable for observers' competence evaluations. While we generally found that perceived similarity leads to higher competence ratings, we cannot draw any conclusions about the underlying cause of this finding. That is, it is still open for investigation which variables (i.e., own judgment in relation to the performer's judgment, similarity in processing the task at hand, and so forth) drove the observer's competence judgment.

Our findings caution that not only the features of the eye-movement displays, but also observer characteristics may affect the inferences made from eye-movement displays. Previous studies about the inference of cognitive processes from eye-movement displays have not considered observers' own confidence in their judgments of the other person. This is important, however, in light of potential practical implementations. For instance, such biases could be problematic when using eye-movement displays to give teachers insight into students' performance. Thus, future research should further address this issue. Other avenues for future research are outlined in the next section.

4.4 | Limitations and future research

The present study showed that observers are able to make different kinds of inferences about multiple-choice task performance from displays of other performers' eye movements. However, this study cannot yet tell us exactly *how* observers made those inferences, which is a question that should be answered in future research. As discussed above, it is likely that observers picked up on the distinctiveness of the displays. However, we cannot know for sure whether participants consciously used this information as a "cue" (i.e., information source) for their judgment. It is also unclear what other cues from the eye-movement displays (e.g., fixation durations, fixation sequences, et cetera) or tasks (e.g., difficulty) they might have used for making inferences about the chosen answer, confidence, and competence of the performer. One way to investigate the underlying cognitive process is by systematically manipulating the characteristics of the eye-movement displays (e.g., removing temporal or sequence information) to see how this affects their inference making. Another option might be to eye-track observers during the task of inference making and use concurrent think-aloud.

Another issue that future studies should address is the generalization of the present results. In our study, we investigated observers' ability to infer cognitive processes from displays of performers' eye-movements when solving multiple-choice graph-comprehension tasks. These multiple-choice tasks were already more complex and educationally relevant than the materials used in prior research in this area (cf. the multiple-choice task materials of Van Wermeskerken et al., 2018; Foulsham & Lock, 2015; Zelinsky et al., 2013). An important open question for future research is whether these findings would generalize beyond multiple-choice tasks. For instance, would observers (e.g., teachers) also be able to interpret performers' (e.g., students') eye-movement displays while they are answering an open question on a graph task? If this was the case, eye-movement displays could become a relevant tool for teachers, at least for highly visual tasks.

Furthermore, we do not know whether our results generalize not only to different task materials, but also to different types of eye-movement visualizations (e.g., dynamic vs. static displays, fixed vs. expanding fixation visualizations depending on duration, trails of different lengths, moving dot/circle vs. spotlight to show fixation locations). Van Wermeskerken et al. (2018) argued, for instance, that dynamic eye-movement displays provide more temporal information than static displays. This might make the interpretation of dynamic eye-movement displays easier. Other display characteristics could also affect the interpretation of eye-movement displays. The use of trail visualizations, for instance, might be especially helpful for tasks in which the order of fixations is indicative for performance. For example, reading research has shown that regressive eye movements are diagnostic of comprehension difficulties (Rayner, Chace, Slattery, & Ashby, 2006). Furthermore, Jarodzka et al. (2012) and Jarodzka et al. (2013) investigated two types of displays (moving dot/circle vs. spotlight) in the context of learning from eye-movement modeling examples. They showed that for guiding a learner's attention (and improving their learning outcomes), it may matter what display design is used. However, it is an open question if that also applies to the interpretation of eye-movement displays. Moreover, spotlight visualizations might be most suitable to guide an observer's attention through task material, in which only one specific element is relevant at the time (as the rest of the screen is blurred in this type of visualization). In contrast, a moving dot might be more appropriate if observers also require the information of other areas to make sense of the performer's behavior. Thus, investigating effects of different types of eye-movement visualizations on observers' eye-movement interpretation would be an interesting avenue for future research.

Interesting questions for future research also concern what other inferences about the performance process observers would be able to make. For instance, could they infer how difficult a performer experienced the task to be? Or given that eye movements have the potential to reveal (multimodal) graph-comprehension processes, such as information integration difficulties (Acartürk & Habel, 2012; Huestegge & Pötzsch, 2018), would observers be able to pick up on those? In this context, we do not know if task familiarity plays a role in (the accuracy of) inference making. Observers in this study were familiar with the content and difficulty of each graph item (they solved each item themselves before rating the eye-movement displays), and our data regarding our last hypothesis (Section 5.3) suggest that their own experience influenced their inferences about other people's competence. Would observers also be able to make inferences about other people's eye movements when they have not experienced the cognitive processes evoked by the task that was performed? Moreover, research shows that estimating what other people know may depend on one's own expertise (cf. Bromme, Rambow, & Nückles, 2001). As such, would experts and novices differ in the extent to which they rely on their own experience (cf. myside bias) when interpreting other people's performance? Given that patterns of eye movements also differ as a function of expertise (see for example, Charness et al., 2001; Gegenfurtner, Lehtinen, & Säljö, 2011; Haider & Frensch, 1999; Jarodzka, Scheiter, Gerjets, & van Gog, 2010), would novices be able to make accurate inferences about cognitive processes of an expert performer from eye-movement displays? Addressing such questions would shed further light on (boundary conditions of) potential applications of eye-movement displays for performance assessment, for instance, in education or training contexts.

Finally, future research could not only focus on human observers of eye-movement displays, but also on machine learning techniques (artificial intelligence). With such techniques, eye-movement measures can be analyzed to identify aspects of performance such as cognitive load or workload (Appel et al., 2019; Halverson, Estepp, Christensen, & Monnin, 2012; Mussgnug, Singer, Lohmeyer, & Meboldt, 2017), performer expertise (Castner et al., 2018), strategic behavior (Eivazi & Bednarik, 2011), or task interest and curiosity (Baranes, Oudeyer, & Gottlieb, 2015). Future studies could investigate whether machine learning techniques are as good or better than humans at analyzing different aspects of performance (e.g., accuracy, confidence, effort) based on eye-movement measures (or vice versa, whether humans are better than machine learning techniques at interpreting certain aspects of performance). If so, automatic interpretation of eye-movement measures could be applied in learning analytics tools for both teachers (helping them provide feedback to their students) and students (e.g., automatized feedback during task performance).

5 | CONCLUSION

In sum, our study extended prior research on inference making from eye-movement displays. We showed that observers can infer performers' answers on educationally relevant multiple-choice tasks, by showing that the performers' confidence affects the answer judgment accuracy, and by investigating other inferences (confidence, competence) than answer judgments. Although this field of research is still in its infancy, these findings are promising in light of potential applications, for instance, as a tool to diagnose task performance in an education or training context. That is, with eye-tracking technology rapidly becoming more portable and more affordable, it is not unthinkable that in the future, students' eye movements could be recorded and displayed to teachers to give teachers more insight into (potential problems in) students' performance. This would allow them to give more specific scaffolding and feedback already during the task performance process.

ACKNOWLEDGMENTS

This research was funded by a Vidi grant (# 452-11-006) from the Netherlands Organization for Scientific Research (NWO) awarded to Tamara van Gog. The authors would like to thank Susan Ravensbergen for her help with creating the materials.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Selina N. Emhardt https://orcid.org/0000-0001-7585-7031

REFERENCES

Acartürk, C., & Habel, C. (2012). Eye tracking in multimodal comprehension of graphs. In R. Cox & J. P. S. Diego (Eds.), *Proceedings of the Workshop on Technology Enhanced Diagrams Research* (Vol. 887, pp. 11–25). Canterbury, UK. http://ceur-ws.org/Vol-887/.

Alexander, P. A., Dumas, D., Grossnickle, E. M., List, A., & Firetto, C. M. (2016). Measuring relational reasoning. *The Journal of Experimental Education*, *84*, 119–151. https://doi.org/10.1080/00220973.2014.963216

Appel, T., Sevcenko, N., Wortha, F., Tsarava, K., Moeller, K., Ninaus, M., Kasneci, E., & Gerjets, P. (2019, October). *Predicting cognitive load in an emergency simulation based on behavioral and physiological measures*. Paper presented at 2019 International Conference on Multimodal Interaction (pp. 154–163). ACM Digital Library. doi:10.1145/3340555.3353735.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bahle, B., Beck, V. M., & Hollingworth, A. (2018). The architecture of interaction between visual working memory and visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, *58*, 274–283. https://doi.org/10.1016/j.learninstruc.2018.07.005

Baranes, A., Oudeyer, P.-Y., & Gottlieb, J. (2015). Eye movements reveal epistemic curiosity in human observers. *Vision Research*, *117*, 81–90. https://doi.org/10.1016/j.visres.2015.10.009

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Lme4: Linear mixed-effects models using Lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, R. (2002). Liking and similarity as predictors of multi-source ratings. *Personnel Review*, *31*, 540–552. https://doi.org/10.1108/00483480210438753

Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision*, *14*, 1–22. https://doi.org/10.1167/14.3.29

Bromme, R., Rambow, R., & Nückles, M. (2001). Expertise and estimating what other people know: The influence of professional experience and type of knowledge. *Journal of Experimental Psychology: Applied*, *7*, 317–330. https://doi.org/10.1037//1076-898x.7.4.317-330

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, *4*, 75–100. https://doi.org/10.1037/1076-898X.4.2.75

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, *9*, 1–15. https://doi.org/10.1167/9.3.6

Castner, N., Kasneci, E., Kübler, T., Scheiter, K., Richter, J., Eder, T., Hüttig, F., & Keutel, C. (2018). *Scanpath comparison in medical image reading skills of dental students: distinguishing stages of expertise development*. Paper presented at Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications. ACM Digital Library. doi: 10.1145/3204493.3204550.

Charness, N., Reingold, E. M., Pomplun, M., & Stampe, D. M. (2001). The perceptual aspect of skilled performance in chess: Evidence from eye movements. *Memory & Cognition*, *29*, 1146–1152. https://doi.org/10.3758/bf03206384

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cooper, L., Gale, A., Darker, I., Toms, A., & Saada, J. (2009). Radiology image perception and observer performance: How does expertise and clinical information alter interpretation? Stroke detection explored through eye-tracking. In B. Sahiner & D. J. Manning (Eds.), *Medical imaging 2009: Image perception, observer performance, and technology assessment* (Vol. 7263, 72630K–72630K-12). Bellingham, WA: SPIE.

DeAngelus, M., & Pelz, J. B. (2009). Top-down control of eye movements: Yarbus revisited. *Visual Cognition*, *17*, 790–811. https://doi.org/10.1080/13506280902793843

Eivazi, S., & Bednarik, R. (2011). *Predicting problem-solving behavior and performance levels from visual attention data*. Paper presented at Proceedings of the 2nd Workshop on eye gaze in intelligent human machine interaction (pp. 9–16).

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*, 215–251. https://doi.org/10.1037/0033-295X.87.3.215

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. *Behavior Research Methods*, *39*, 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*, 1704–1726. https://doi.org/10.1111/cogs.12211

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review*, *23*, 523–552. https://doi.org/10.1007/s10648-011-9174-7

Glaholt, M. G., Wu, M.-C., & Reingold, E. M. (2009). Predicting preference from fixations. *PsychNology Journal*, *7*, 141–158. https://doi.org/10.1037/e527342012-455

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

Haider, H., & Frensch, P. A. (1999). Eye movement during skill acquisition: More evidence for the information-reduction hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 172–190. https://doi.org/10.1037//0278-7393.25.1.172

Halverson, T., Estepp, J., Christensen, J., & Monnin, J. (2012). Classifying workload with eye movements in a complex task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*, 168–172. https://doi.org/10.1177/1071181312561012

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, England: Oxford University Press.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. New York: Routledge.

Huestegge, L., & Pötzsch, T. H. (2018). Integration processes during frequency graph comprehension: Performance and eye movements while processing tree maps versus pie charts. *Applied Cognitive Psychology*, *32*, 200–216. https://doi.org/10.1002/acp.3396

Jaarsma, T., Jarodzka, H., Nap, M., van Merrienboer, J. J., & Boshuizen, H. (2014). Expertise under the microscope: Processing histopathological slides. *Medical Education*, *48*, 292–300. https://doi.org/10.1111/medu.12385

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*, 813–827. https://doi.org/10.1007/s11251-012-9218-5

Jarodzka, H., Scheiter, K., Gerjets, P., & van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learning and Instruction*, *20*, 146–154. https://doi.org/10.1016/j.learninstruc.2009.02.019

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*, 62–70. https://doi.org/10.1016/j.learninstruc.2012.11.004

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*, 329–354. https://doi.org/10.1037/0033-295x.87.4.329

Lawrence, M. A., & Lawrence, M. M. A. (2016). Package 'ez. R package version 4.4-0. Retrieved from https://cran.r-project.org/web/packages/ez/ez.pdf.

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*, 738–752. https://doi.org/10.1002/acp.3060

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

Mussgnug, M., Singer, D., Lohmeyer, Q., & Meboldt, M. (2017). Automated interpretation of eye–hand coordination in mobile eye tracking recordings. *KI-Künstliche Intelligenz*, *31*, 331–337. https://doi.org/10.1007/s13218-017-0503-y

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from. http://www.R-project.org/

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, *10*, 241–255. https://doi.org/10.1207/s1532799xssr1003_3

Reingold, E. M., Charness, N., Pomplun, M., & Stampe, D. M. (2001). Visual span in expert chess players: Evidence from eye movements. *Psychological Science*, *12*, 48–55. https://doi.org/10.1111/1467-9280.00309

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*, 80–94. https://doi.org/10.1111/bjep.12175

Shah, P., & Freedman, E. G. (2011). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive*

Science, *3*, 560–578, doi: https://doi.org/10.1111/j.1756-8765.2009.01066.x

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, *14*, 47–69. https://doi.org/10.1023/A:1013180410169

Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience*, *6*, 1317–1322. https://doi.org/10.1038/nn1150

Stanovich, K. E., & West, R. F. (2008). On the failure of intelligence to predict myside bias and one-sided bias. *Thinking & Reasoning*, *14*, 129–167. https://doi.org/10.1080/13546780701679764

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, *22*, 259–264. https://doi.org/10.1177/0963721413480174

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*, 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M., & Velichovsky, B. M. (2010). Yarbus, eye movements and vision. *Perception*, *1*, 7–27. https://doi.org/10.1068/i0382

Van Marlen, T., van Wermeskerken, M., Jarodzka, H., & van Gog, T. (2018). Effectiveness of eye movement modeling examples in problem solving: The role of verbal ambiguity and prior knowledge. *Learning and Instruction*, *58*, 274–283. https://doi.org/10.1016/j.learninstruc.2018.07.005

Van Wermeskerken, M., Litchfield, D., & van Gog, T. (2018). What am I looking at? Interpreting dynamic and static gaze displays. *Cognitive Science*, *42*, 220–252. https://doi.org/10.1111/cogs.12484

Van Wermeskerken, M., Litchfield, D., & van Gog, T. (submitted). Eye see what you are doing: Inferring task performance from eye movement data. Manuscript submitted for publication.

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. New York, NY: Springer Publishing Company.

Wolfe, C. R. (2012). Individual differences in the "myside bias" in reasoning and written argumentation. *Written Communication*, *29*, 477–501. https://doi.org/10.1177/0741088312457909

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

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*, 1–13. https://doi.org/10.1167/13.14.10

How to cite this article: Emhardt SN, van Wermeskerken M, Scheiter K, van Gog T. Inferring task performance and confidence from displays of eye movements. *Appl Cognit Psychol*. 2020;1–14. https://doi.org/10.1002/acp.3721