Animation and interactivity in computer-based physics experiments to support the documentation of measured vector quantities in diagrams: An eye tracking study

Christoph Hoyer and Raimund Girwidz

Ludwig-Maximilians-Universität München, Chair of Physics Education, Theresienstraße 37, 80333 München, Germany

(Received 20 June 2020; accepted 17 September 2020; published 19 October 2020)

Simulations and virtual or remote laboratories are increasingly used in schools. The extent to which individual experimental skills can be acquired when experimenting in digital applications is, however, questionable. This paper focuses on finding multimedia features for digital experiments to support the transfer of measured values from the laboratory system to a diagram. Beside physical considerations, spatial translation processes could be crucial for a successful assignment. Therefore, the influence of the subjects' spatial ability is examined. Using a pretest post-test design ($N = 119$), the effects of training with supportive animation (animation group) and training with an interactive task and feedback (interactive group) were tested. The results of both groups were each compared to those of a reference group. Eye tracking data were recorded during training to investigate the origin of different training effects. Hence, fixations and saccades during training were analyzed. For the investigation of the distribution of the saccadic movements, polar diagrams were used in combination with estimated probability density functions. The results show that the score in the pretest is correlated to the score achieved in the card rotation test, which measures the spatial rotation skills of the subjects. Further, the subjects in the interactive group benefited from the training more than the subjects in the reference group did. There were no significant differences in the effect of the training between the animation group and the reference group. Eye tracking data reveal that the training in the interactive group caused the most comparative eye movements between the laboratory system and the diagram. The training in the animation group led to the highest visual attention; however, subjects in this group concentrated on the dynamic elements. These results indicate that especially students with weak spatial skills need additional support when transferring measured values from the laboratory system to the diagram. This assignment can be practiced in computer-based experiments, in particular with an interactive training task and feedback. Additionally, the analysis showed that the training is equally suitable for learners with different spatial abilities. A corresponding task was implemented into a virtual laboratory.

DOI: 10.1103/PhysRevPhysEducRes.16.020124

I. INTRODUCTION

Collecting, organizing, and interpreting measured values are key competences for processing experimental data. The creation and interpretation of graphical representations of measured values plays a particularly important role. The following examines how the learners' skill of transferring measurement results from the laboratory system to the correct points in the diagram can be supported when working with computer-based experiments. This was investigated using a virtual laboratory with which the field of a permanent magnet can be measured and visualized. Supporting multimedia tools are derived from theory and their effectiveness is tested. However, first the relevance of practicing the documentation of measurement results is shown and associated difficulties are described.

Science lessons in school require students to develop a deeper understanding of how science works and to practice relevant skills (e.g., Ref. [1]). Therefore, in schools, planning and performing experiments should play a central role. Simulations and virtual or remote laboratories can be valuable supplements to real experiments as described, among others, by Finkelstein *et al.* [2] for electrical circuits and by Martínez *et al.* [3] for image formation and optical aberration. Experimenting on the computer can also be advantageous if cost and security reasons do not allow all students to carry out a real experiment. When deciding whether to conduct a real or computer-based experiment, the advantages and disadvantages should be taken into account. Also, the expected learning outcomes have to be considered [4,5]. Teaching physics includes imparting

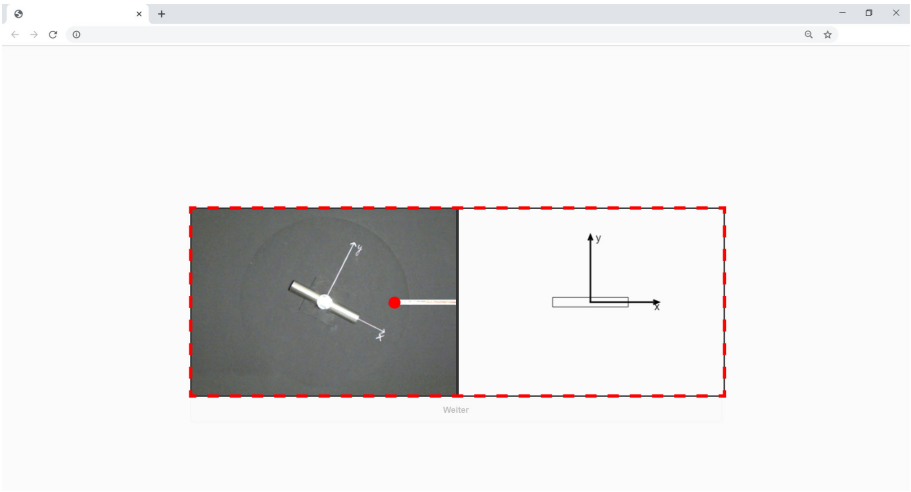

FIG. 1. View of the webpage containing the application. In the middle (within the red dashed line) the experimental setup (left) and the diagram (right) can be seen. The orientation of the laboratory system changes during data acquisition, while the orientation of the coordinate system of the diagram stays the same. The position of the sensor is marked in red. The red markings (dashed line and circle) were added here for illustrative purposes and were not included in the application.

knowledge about physical content and experimental skills. An interesting question is to what extent individual experimental skills can be practiced by using digital applications. To help answer this question, the following presents an investigation into what support in computer-based experiments improves the documentation of measurement results in diagrams.

Both the experimental setup and the diagram are representations that the learner must relate to each other to successfully document measurement results. Therefore, research on learning with multiple representations [6] offers the opportunity to identify potential difficulties in the documentation process.

The DeFT framework [7] describes factors that make it difficult to build relations between representations. The influence of three of these factors on the documentation of measurement results in computer-based experiments is briefly described as an example:

- Level of abstraction: If the axes of the laboratory system of the experiment and the axes of the diagram are coded differently, the translation process between both could be difficult for the learners. This, for example, is the case if a movement is documented in a time-velocity diagram.
- Dimensionality: If measurements are recorded in three-dimensional space and measurement results are visualized in a two-dimensional diagram, the required abstraction can be an obstacle for the learners. An example is the representation of the gravitational potential on the surface of a mountain using equipotential lines.
- Whether representations are static or dynamic: If the spatial relationship between the laboratory system and the coordinate system of the diagram changes during data acquisition, the translation process can be challenging. An example of such an experiment is the determination of the directional characteristic of a receiver dipole. The emitter is usually more massive and therefore more difficult to move than the receiver. Because of this, it is useful to turn the receiver instead of moving the emitter. Simultaneously, by rotating the receiver, the laboratory system is rotated. In contrast, the orientation of the coordinate system of the diagram stays the same.

These examples show that the challenges of documenting data in diagrams are diverse. Therefore, the supporting multimedia content must be adapted for the specific task.

The case investigated concerns documenting measured values of the field of a permanent magnet in a virtual laboratory. To construct multimedia tools to support this task, the exact area of application must be defined first.

Measuring the field of a permanent magnet necessitates a sensitive but unwieldy measurement apparatus. This is because the absolute value of the magnetic flux density decreases rapidly with increasing distance from the magnet. Minimal deviations in the alignment of the sensor cause large fluctuations in the measured value. When constructing such a laboratory, instead of moving the measurement device around the magnet, it is appropriate to restrict its movement to the radial direction. Using a linear drive enables highly accurate changes in the distance between the middle of the magnet and the sensor. This linear movement in combination with a rotation of the magnet allows the determination of the magnetic field around the magnet. However, by rotating the magnet, the associated laboratory system is rotated (see Fig. 1). As derived from the DeFT framework [7] above, the changing orientation between the laboratory system and the coordinate system of the diagram can pose challenges for the learners.

The focus of the investigation is therefore on finding multimedia techniques for computer-based experiments that support this transfer between the laboratory system of the experiment and the coordinate system in the diagram.

Beside physical considerations, since this process can also be seen as a spatial transformation between the laboratory system and the coordinate system in the diagram, success in this task may also depend on the learners' spatial thinking abilities (see Sec. I A 2). If so, this should be considered when choosing appropriate supportive multimedia tools. In the next step, possible approaches are derived from theory. Subsequently, their effectiveness is tested. Eye tracking data will help to interpret the results.

A. Theory

1. Experimental skills

In recent years, more and more attention has been paid to teaching experimental skills in physics courses (e.g., Refs. [8–12]). There are various categorizations in the literature to classify experimental skills. An overview is given by Emden [13]. A categorization of experimental skills in undergraduate labs is provided by the AAPT Committee on Laboratories [11]. Experimental skills can be roughly classified as skills required to prepare experiments, skills necessary to carry out experiments, or skills relevant to evaluate experimental data [14]. Recent studies have shown that lab courses focused on teaching experimentation do not necessarily convey less content knowledge than those concentrating on content reinforcement [15].

With the rapid development of digital applications, conducting experiments is no longer tied to working in a laboratory. Digital applications offer new possibilities to carry out experiments on the computer [8]. When experimenting with digital applications, however, it is questionable to what extent experimental skills can be acquired.

Many computer-based experiments document measured data automatically. Such automated documentation can free up cognitive resources for other learning activities [16]. However, by documenting measured values automatically it cannot be taken for granted that the students' ability to document measured values in diagrams is supported. Nevertheless, having a wide range of application, multimedia can conceivably help training the documentation of measurement results in diagrams.

2. Spatial rotation ability

There are different categorizations of spatial abilities in the literature [17–19]. The review by Cole *et al.* [20] gives an overview and summarizes research results on the role of spatial thinking in the teaching of astronomy. A component of spatial skills that is included in all of the categorizations is mental rotation ability. This ability concerns the mental rotation of two- and three-dimensional objects. A common test to measure two-dimensional rotation ability is the card rotation test developed by Ekstrom *et al.* [21] (see also Ref. [22]). A suitable test for measuring three-dimensional rotation ability is the cube comparison test [21] or the mental rotation test, originally developed by Vandenberg and Kuse [23] (for an overview, see Ref. [22]). There is also a revised version of the latter test [24].

Findings suggest that students with higher visual processing skills find it easier to use visualizations and other multimedia modules (for a summary, see Ref. [22]). Several research papers report that spatial thinking influences problem solving and interpreting graphs in science [25–29]. Especially the translation between frames of reference, a challenging task for learners [30], is influenced by the learners' visual spatial abilities [29].

In the case examined, the learners have to map positions from the laboratory system to the diagram. Since this task is similar to the problem of translating between reference systems, it should be considered whether this task also depends on the learners' spatial thinking ability.

3. Linking experiment and diagram

To support the transfer of measurement results of the magnetic flux density from the laboratory system to the correct position in the coordinate system of the diagram, so-called relational cues [31] could help. These can emphasize the connection between the experiment and the diagram. Multimedia offers various options for providing such cues in simulations and virtual or remote laboratories. Various approaches in the literature are promising.

Animated visualizations. Computer-based experiments that automatically document data display a static diagram containing the correctly plotted measurement values. The literature shows that compared to a static representation, an animation could be superior in supporting the transfer between the laboratory system and the diagram.

The results of a meta-analysis by Höffler and Leutner [32] suggest that animations seem to outperform static pictures if the motion depicted in the animation explicitly refers to the learning topic. Then animations can help to build mental models of the dynamic [33]. Berney and Bétrancourt [34], in their meta-analysis, are only partially able to replicate those results. They confirm that animations seem to be more effective compared to static pictures, but they cannot find significant differences in this effect between different types of knowledge. Instead of a comparison of static pictures and animations, the authors request an analysis of when, why, and for whom animations are beneficial for learning [34,35].

An important property of animations is that they can attract attention and help focus on relevant aspects. The onset of motion or appearance of objects can guide the focus of the learners [31,36,37]. However, results show that attention guidance does not necessarily lead to better learning outcomes [38–40].

When designing animations, the characteristics of the learning content should be taken into account. It was previously described that the translation between the laboratory system and the coordinate system of the diagram could be influenced by the students' visual spatial abilities. If the learners indeed use a spatial translation process for the documentation task according to Salomon's supplantation framework [41], an animation illustrating the mental translation task may foster the internal mental process. Therefore, animations could be particularly effective for supporting such tasks. Accordingly, Gilligan *et al.* [42] show that a continuous animation depicting the transformation of an object can have a positive effect on later spatial rotation tasks.

In computer-based experiments, a dynamic animation could be an appropriate aid for both binding visual attention and improving the transfer of measured values of the magnetic flux density from the laboratory system to the correct point in the coordinate system in the diagram. A possible implementation is the following: When a learner measures the field at a specific position in the laboratory system, an animation shows how to transfer this position to the corresponding point in the diagram. After the animation has ended, the measured value stays marked in the diagram.

Interactivity and feedback. Chi and Wylie [43] distinguish between four groups of learning activities. They point out that learning increases while "activities move from *passive* to *active* to *constructive* to *interactive*." Animations could leave the learners in a passive recipient role. Thus, a more active task could improve the transfer of measured values from the laboratory system to the diagram to a greater extent. The feature of interactivity distinguishes new media from "classic" media, which show the same information to every learner without taking the learners' activities into account. Interactivity is a key property for the successful use of multimedia [44]. Interactivity in multimedia applications is characterized by mutual actions between the learner and the learning environment [45].

A model for interaction in multimedia applications is described in Ref. [46]. In addition to the interaction between the learning environment and the learner, the authors also consider learner characteristics, motivational and emotional aspects, cognitive and metacognitive characteristics, and the learner's mental models. An optimal learning outcome requires finding the right level of interactivity in the learning environment. However, finding this level is difficult [47].

For cognitive tutors, interactive elements were identified that are contained in all such systems [47]. These elements are implicit yes-no feedback, specific feedback messages for commonly occurring errors, and next-step hints.

In the context of this application, feedback can be defined as "information communicated to the learner that is intended to modify his or her thinking or behavior to improve learning" [48].

Therefore, for the case examined, providing feedback helping the learners to rethink and reflect on their answers could induce deeper processing. Immediate feedback at task level, highlighting the correct position in the diagram, could cover these requirements. This is also justified by theory.

Hattie and Timperley [49] give an overview of research results showing that corrective feedback at task level can be effective, especially if it draws attention to faulty interpretations. Van der Kleij *et al.* [50] showed that students paid more attention to immediate feedback than to delayed feedback. Suddenly appearing feedback could additionally enforce this effect as research shows that the appearance of objects can capture attention [31,37,51]. In this way, giving immediate feedback clearly delineated from the rest of the application can be used to guide the learners' attention. It is worthy of note that, like for animations, binding attention does not necessarily result in better learning outcomes [38].

In summary, immediate feedback in computer-based experiments could be helpful for improving the transfer of measured vector quantities from the laboratory system to the coordinate system of the diagram. Such feedback could be integrated as follows: After learners have carried out a measurement in the computer-based experiment, they are asked to mark the point at which they would plot the measured value in the diagram. Simultaneously with the appearance of this marking, the laboratory visualizes and highlights the measured quantity at the correct position in the diagram.

4. Eye tracking

In recent years, there has been an increasing number of eye tracking studies in teaching and learning (for an overview, see the metastudies [52,53]). Recent investigations show how eye tracking data can provide insights into students' visual attention in physics education (for example, Refs. [54–56]).

Eye tracking systems allow the recording of learners' eye movement during a learning task. The eye-mind assumption [57] states that eye fixations indicate processing of the fixated information. Accordingly, eye tracking data help checking whether elements intended for learning were perceived in the expected way or if other salient elements attracted the learners' attention.

For the investigation described in this paper, eye tracking enables the analysis and comparison of eye movements during training. In this way, processes can be identified that are caused by the training and that are decisive for improving the task. For the comparability of the eye movements between the different training conditions, it is important that the conditions include the same components of the user interface and differ only in the given training stimulus.

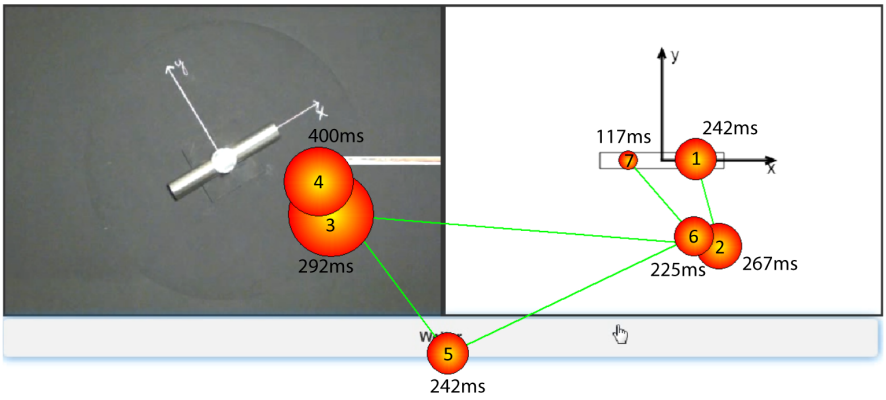

FIG. 2. Visualization of the scan path of a subject. The red circles represent fixations. The green lines show saccadic movements. The order in which the fixations occurred is described by the number in the circles. The duration of each fixation is indicated by the size of the corresponding circle and, in addition, is displayed next to the circle.

In the past, research concentrated mainly on the analysis of the total number of fixations, the duration of fixations, and the dwell time in AOIs. Also, the total number, direction, length, and duration of the movements between two fixations, the so-called saccades, were investigated (for an overview, see Refs. [52,58]). Figure 2 shows the scan path of a test person as an example. The circles correspond to fixations, with the size of the circles representing the duration of each fixation.

The number of saccadic movements was often examined to analyze integration processes between representations. While in some studies an increased number of saccades led to better performance (e.g., Refs. [56,59–61]), this could not be confirmed in other studies (e.g., Refs. [62–64]).

In the case investigated, the relationship between elements in the laboratory system and the diagram is of particular interest. Appropriate eye movements may be necessary to perceive these relationships. Therefore, a combined view of the directional and length distributions of saccades could reveal the perception of relations between the depicted elements. A suitable procedure was developed and is presented below.

B. Research questions

This paper examines four research questions (RQs).

In general, the laboratory system and the coordinate system in the diagram have no fixed spatial relationship to each other.

RQ1: If the laboratory system and the coordinate system in the diagram can be transformed into each other using rotation and translation and if the task is to transfer a measurement location of the laboratory system as accurately as possible into the coordinate system of the diagram, is there a correlation between spatial rotation ability and success in this assignment task?

In most computer-based experiments, as soon as measurements are recorded, they are immediately visualized at the correct positions in a diagram. In comparison, an animated visualization or an interactive assignment task with feedback could improve the subjects' accuracy when performing the assignment on their own.

RQ2: When the laboratory system is rotated relative to the coordinate system of the diagram, are there significant differences in the accuracy with which measurement locations are transferred from the laboratory system to the diagram after subjects attend one of the following three trainings:

(i) Training with animations illustrating the transfer between the laboratory system and the coordinate system of the diagram.
(ii) Training with interactive assignment tasks with feedback where subjects have to transfer positions from the laboratory system to the diagram.
(iii) Training with depictions of correctly solved assignment tasks.

As described above, the three training conditions have different characteristics. Subjects with different spatial abilities could therefore benefit from training in different forms. To select the most suitable training, it should thus be examined whether a relationship exists between the spatial abilities of the subjects and the performance gain in the individual training conditions.

RQ3: Is there a relationship between the spatial abilities of the subjects and the performance gain during the three training conditions?

Analyzing the fixations and saccadic movements provides information about the visual processing of the content of the training. It is crucial for the training that the information content is perceived and that it receives the visual attention of the learners. This can be determined by analyzing fixations. However, saccadic eye movements can also indicate cognitive processes. Such eye movements, for example, are necessary to perceive the position of the measuring location in relation to the laboratory system or to check the position of marked values in a diagram relative to

the corresponding coordinate system. Other characteristic saccadic movements should result from establishing a relationship between the laboratory system and the coordinate system of the diagram. The various training conditions could have different effects on those components of gaze behavior.

RQ4: Is there a difference between the training groups in the number of fixations and in the distribution of the saccades during training? Furthermore, can those measures indicate visual processes important for successful training?

II. METHODS

A. Participants

119 students (81 males, 38 females) from eight different 11th grade classes took part in the study. The curriculum of the 11th grade in a Bavarian high school includes describing magnetic fields as vector quantities. To guarantee that all students bring basic knowledge about magnetic fields with them, the examination was carried out shortly before the end of the school year.

B. Design

The design of the study is depicted in Fig. 3. At the beginning, the 119 subjects worked on the card rotation test (CRT). This test indicates the spatial rotation ability of the subjects and is a paper and pencil test. Subsequently, they took the pretest. After the pretest, the intervention took place. For the intervention, the subjects were divided randomly into three groups. The animation group consisted of $N_{\text{animation}} = 38$, the interactive group of $N_{\text{interactive}} = 41$, and the reference group of $N_{\text{reference}} = 40$ participants. Finally, all of the subjects took the post-test. The examination on the computer took only a short time. To ensure that the test subjects were not disturbed by the eye tracking and in order not to interrupt the sequence of pretest, training, and post-test, the eye tracking system was started when the subjects began to work with the computer application. For this reason, the eye tracker was calibrated only once at the beginning. Thus, eye movements were recorded while the subjects worked on the pretest, the intervention, and the post-test.

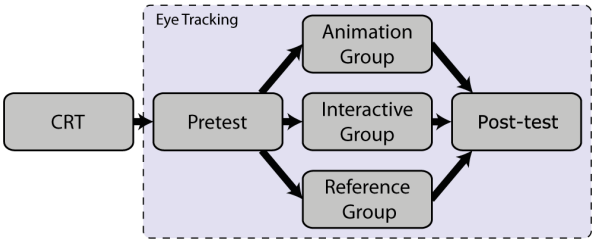

FIG. 3. Illustration of the design of the study. First the subjects worked on the card rotation test. Afterward, eye movements were recorded while the subjects participated in the pretest, the intervention, and the post-test.

C. Material

In this section, the technical equipment, the tests used, and the different training conditions are described. Also, the procedures of the investigation and statistical methods are presented.

1. Computer environment

The pretest, intervention, and post-test were carried out on a Windows 10 computer workstation. The resolution of the 24" monitor was 1920×1200 pixels. The application was written in HTML and JavaScript. It was opened in the Chrome browser. During the pretest and the post-test, click data were recorded using JavaScript and PHP. A high-quality mouse and mousepad were used to ensure best conditions. The mouse resolution was the same for all participants. An average setting was chosen based on results of a pilot test ($N = 8$).

2. Computer-based experiment

A modified version of a virtual laboratory with which the field of a permanent magnet can be measured [65] was used for the assessment. The application runs in all standard browsers. Figure 1 shows a screenshot. The virtual lab is based on a real experiment. To give a realistic impression, animated pictures of the real experiment are used in the virtual lab. For the measurement of the field of the magnet, the sensor can be moved radially while the magnet can be rotated. Thus, the apparatus allows a highly accurate positioning of the sensor at all positions within 100 mm around the magnet (further explanations can be found in the introduction).

3. Eye tracker

The eye tracking system used was an Eye-Follower from LC Technology. The system has an accuracy of less than 0.4° of visual angle over the whole range of head movement. Four cameras are used for tracking eye movements. Two of them record the motion of the head. In this way, the system can accurately track the movements of the eyes, even if the subjects move their head during computer operation. This makes it possible to work on the computer as naturally as possible, uninfluenced by the eye tracking system. The sampling rate of the system is 120 Hz. A nine-point calibration was carried out for every subject. The calibration process was successful if the accuracy was less than 0.63 cm (0.25 inches). To discriminate between fixations and saccades, the dispersion-based algorithm LC Fixation Detector was used (for more information about eye tracking algorithms, see Ref. [66]).

4. Tests

Card rotation test. To measure the spatial rotation ability of the subjects, the card rotation test of the kit of factor-referenced cognitive tests was chosen [21]. This test is a

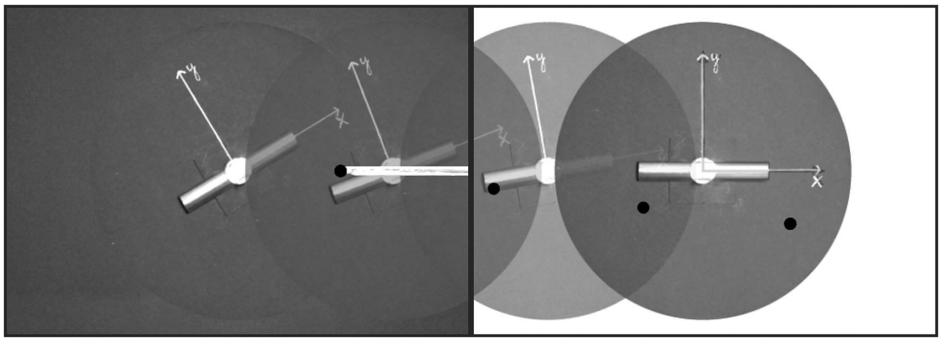

FIG. 4. Illustration of the animation as a series of still images captured at different points in time. The animation dynamically transforms the laboratory system to the coordinate system of the diagram.

paper-and-pencil test and measures the capacity to mentally rotate two-dimensional geometries. Its internal consistency is excellent ($\alpha = 0.96$) as Burton and Fogarty [67] reported. As the introductory material was available only in English, it was translated into German. In the test, one two-dimensional object is depicted on the left of a line and eight objects on the right of this line. The eight objects can be transformed into the object on the left using rotation and reflection. The task in this test is to compare if the objects on the right of the line either are the same but rotated or are mirrored in comparison to the referential object on the left of the line. The subjects received one point for each correctly answered subtask, and one point was deducted for each incorrectly answered subtask. The time on task was limited to 3 min.

Pretest. In the pretest, the subjects' task was to click as accurately as possible the position in the diagram corresponding to the location of the sensor in the laboratory system (see Fig. 1). Each subject worked on eight such assignment tasks. The eight measuring locations were evenly distributed over the four quadrants of the coordinate system. In the background, the system recorded the distance between the clicked location and the correct one. The smaller this distance was, the more points the participants received. A maximum of five points could be achieved with each task. Overall, this resulted in a maximum of 40 points for the pretest. The total pretest score was used for further analysis.

Post-test. In the post-test, the subjects had to fulfil the same task as in the pretest. The only difference was that eight new measuring locations were chosen. Like in the pretest, these locations were distributed over the coordinate system so that there were exactly two in each quadrant of the coordinate system. The system again recorded the distance between the clicked location and the correct one. A maximum of five points could be achieved with each task. Overall, this resulted in a maximum of 40 points for the post-test. For further analysis, the total post-test score was used.

5. Training conditions

Between the pretest and the post-test, the training took place. To avoid time pressure negatively impacting learning, participants could freely choose when to move on to the next training task. Each training provided information only on the correct assignment of the measurement location to the diagram. Subjects were distributed randomly to the following three conditions:

- In the animation group, a partially transparent dynamic animation visualized the transfer of the laboratory system including the measurement location to the diagram. From the time the transformed laboratory system matched the coordinate system of the diagram, the correct point in the diagram remained marked. Later the animation showed how to transfer the coordinate system of the diagram back to the laboratory system of the experiment. Figure 4 depicts video stills of an animation for a measurement location in the fourth quadrant. The training consisted of eight animations showing how to transfer the measurement locations of the pretest. By clicking the "Next" button, the subjects could move on to the next animation.
- In the interactive group, the subjects were asked to do the pretest tasks again. This time, when they clicked on the diagram, the point they clicked was marked in black. Simultaneously, the correct position was marked in red (see Fig. 5). The markers remained in their positions until the subjects clicked the "Next" button to proceed to the next training task.
- In the reference group, for each of the eight measurement locations of the pretest, the correct solutions were presented. The correct points were marked in the diagram. By clicking the "Next" button, the subjects could move through the eight solutions. Figure 6 shows the marking for the first of the eight measurement locations. The marking of the correct position corresponds to the behavior when in computer-based experiments, the measurement values are automatically documented at the correct position in the diagram. This group provides a reference level. In

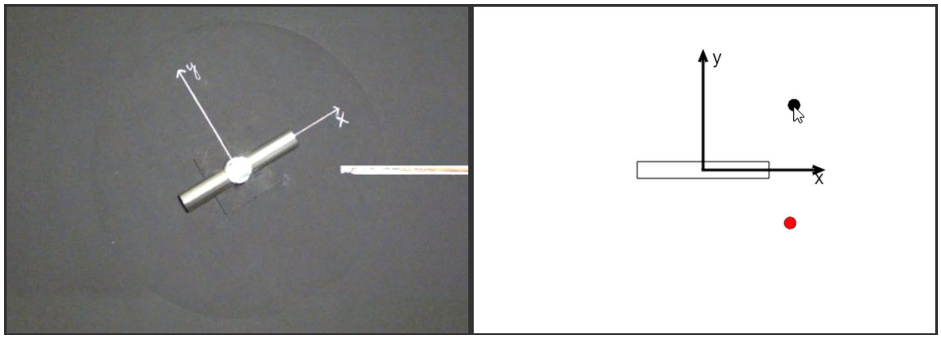

FIG. 5. Screenshot of the first training task in the interactive group. After clicking on the diagram, the clicked position is marked in black and the correct position appears in red.

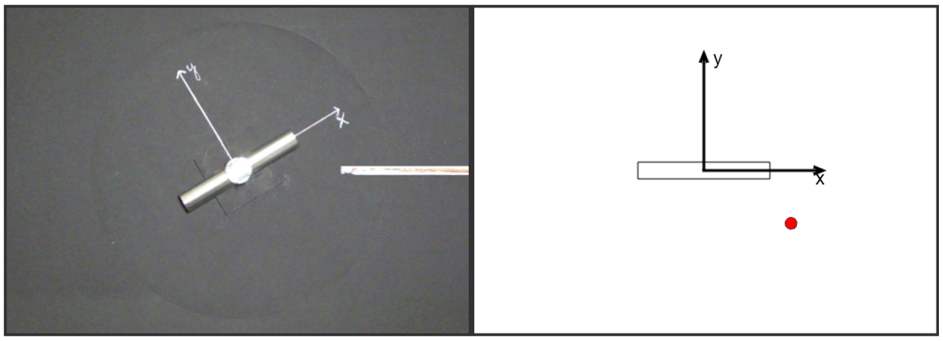

FIG. 6. Screenshot of the first part of the training in the reference group. The correct position is marked in red in the diagram.

relation to this level, the effects of the other two trainings were assessed.

D. Procedure

The investigation was part of class visits at the Chair of Physics Education at the LMU Munich. The classes visited the university on different days.

At the beginning of each visit, all students took the CRT. An automated slide show led through the manual of the test. After the subjects' questions were answered, the test was carried out. Testing time was limited to 3 min. An acoustical signal indicated the end. After this, the subjects worked in groups on physics experiments using Arduinos. Those experiments required no documentation of measurement results so that the results of the investigation were not falsified.

While the others continued to work on their experiments, a researcher led the subjects one by one into a separate room in which a computer workstation and the eye tracking system were set up. After calibration of the eye tracking system, the students started to work with the computer application. From this point onwards, the eye tracker recorded the eye movements.

The computer application was separated into four parts:

1. Part one introduced the user interface, which contains a view of the experimental setup including the axes of the laboratory system and the diagram to which the measurement locations should be transferred (see Fig. 1). On-screen text described the individual components. The introduction finished with two example tasks where students could practice assigning measurement locations to the diagram. The first part lasted an average of 185.1 sec.
2. Part two of the application consisted of the pretest. The pretest lasted an average of 98.0 sec.
3. In part three, the training was carried out. Participants were randomly assigned to one of three training conditions. The training lasted on average 108.5 sec.
4. In part four, the subjects worked on the post-test. This part took an average of 90.1 sec.

In total, each student worked with the application for an average of 481.7 sec.

E. Statistical procedures

1. Pearson's product-moment correlation coefficient

To assess the relationship between the CRT score and the pretest score as well as between the CRT score and the performance gain, Pearson's correlation coefficients were calculated. By using boxplots, the data was checked beforehand for outliers. Also, the assumptions of linearity and homoscedasticity were tested using a scatter plot. To prove the assumption of bivariate normality, Henze-Zirkler's test was calculated.

2. ANCOVA

Following the suggestions of Dimitrov and Rumrill [68], a one-way analysis of covariance (ANCOVA) was conducted to evaluate the results of the pretest and the post-test. The pretest score was used as a covariate and the post-test score as the dependent variable. In this way, the variance that existed in the pretest is corrected in the post-test score. Before calculating the ANCOVA, the underlying assumptions were checked. Data were searched for outliers using boxplots. The Shapiro-Wilk test showed whether data was normally distributed for each group. Furthermore, the homogeneity of variance was tested by calculating Levene's test. The independence of the covariate from the group membership was investigated by using an ANOVA. To test whether the assumption of homogeneity of regression slopes was met, a customized ANCOVA model was calculated that included the interaction between the covariate and the independent variable.

Contrasts for pairwise comparisons allowed the analysis of the results of the ANCOVA in more detail.

3. Kruskal-Wallis test

The time on task and the numbers of eye fixations and saccades were not normally distributed. To investigate differences between the groups in those variables, Kruskal-Wallis tests were calculated. To examine the differences in more detail, Mann-Whitney U tests were used for pairwise comparisons of the medians. In accordance with Divine *et al.* [69], the distributions of the dependent variable were checked for differences between the groups. Therefore, Kolmogorov-Smirnov tests were calculated. For reporting the effects of the pairwise comparisons, a Bonferroni correction was taken into account.

F. Preliminary analysis

To check whether the randomization of the 119 subjects into the animation group, interactive group, and reference group was successful, the pretest results were examined for differences between the training groups.

A one-way ANOVA was conducted to assess if there were differences in the pretest score. There were no outliers, according to inspection with a boxplot. Data was normally distributed for each group (Shapiro-Wilk test, $p > 0.05$) and the assumption of homogeneity of variance (Levene's test, $p > 0.05$) was met.

The mean score in the pretest decreased from the reference group ($M = 11.20$, $SD = 5.73$) via the interactive group ($M = 10.78$, $SD = 5.64$) to the animation group ($M = 10.39$, $SD = 4.73$).

There was no significant difference in the pretest score between training conditions, $F(2, 116) = 0.217$, $p > 0.05$, partial $\eta^2 = 0.004$. This indicates that the randomization was successful.

III. RESULTS

A. Spatial ability and task performance

To answer RQ1, Pearson's correlation coefficient was calculated to assess the relationship between the CRT scores and the scores in the pretest. It was suspected that there is a positive correlation between students' CRT scores and the pretest scores.

There were no outliers in the data; also, the assumptions of linearity, homoscedasticity and bivariate normality (Henze-Zirkler's test, $p > 0.05$) were fulfilled. A small positive correlation between the CRT score and the pretest score, $r = 0.234$, $N = 119$, p (one-tailed)< 0.01, with an $R^2 = 0.055$, was found. This shows that 5.5% of the variability in the CRT score is shared with the pretest score.

B. Training effect on task performance

To answer RQ2, it was tested if the type of training affected post-test performance. To avoid the influence of time pressure, the subjects were free to decide when to move on to the next training task. They had enough time to think about all the details that seemed relevant to them. The average duration of training differed between the groups. The animation group spent the most time on task ($M = 149.2$), followed by the interactive group ($M = 103.5$) and the reference group ($M = 75.0$). The time is given in seconds. The time on task was not normally distributed for each group; therefore, a Kruskal-Wallis test was calculated to check whether there are significant differences in the processing time of the training.

Indeed, the processing time was significantly affected by group membership, $H(2) = 78.68$, $p < 0.001$. To examine the differences between the groups in more detail, Mann-Whitney U tests were used. Considering a Bonferroni correction, all effects are reported at a 0.0167 level of significance.

A Mann-Whitney U test was calculated to determine if there were differences in the processing time of the training between the animation group and the reference group. The distributions differed between both groups (Kolmogorov-Smirnov, $p < 0.05$). There was a statistically significant difference in the processing time of the training between the animation group ($M_{\text{rank}} = 59.5$) and the reference group ($M_{\text{rank}} = 20.5$), $U = 0.0$, $z = -7.60$, $p < 0.001$, $r = -0.86$.

Also, a Mann-Whitney U test was calculated to determine if there were differences in the processing time of the training between the interactive group and the reference group. The distributions did not differ between both groups (Kolmogorov-Smirnov, $p > 0.05$). There was a statistically significant difference in the median of the processing time of the training between the interactive group ($Mdn = 106.1$) and the reference group ($Mdn = 72.26$), $U = 324.5$, $z = -4.68$, $p < 0.001$, $r = -0.52$.

A Mann-Whitney U test was calculated to determine if there were differences in the processing time of the training

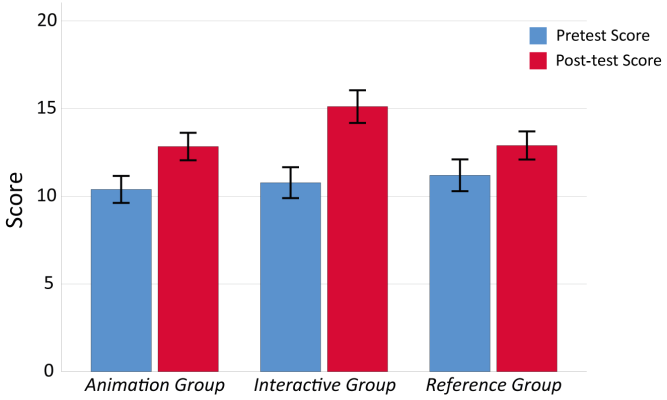

FIG. 7. Average scores of the pretest and the post-test for the three training groups. The error bars represent the standard error of the mean.

between the animation group and the interactive group. The distributions differed between both groups (Kolmogorov-Smirnov, $p < 0.05$). There was a statistically significant difference in the processing time of the training between the animation group ($M_{\text{rank}} = 57.5$) and the interactive group ($M_{\text{rank}} = 23.8$), $U = 113.0$, $z = -6.54$, $p < 0.001$, $r = -0.74$.

The mean score in the post-test decreased from the interactive group ($M = 15.12$, $SD = 5.97$) via the reference group ($M = 12.90$, $SD = 5.07$) to the animation group ($M = 12.84$, $SD = 4.84$). Figure 7 depicts the means of the pretest and post-test scores for each group and the corresponding standard errors.

A one-way ANCOVA was calculated to assess the effect of training condition on the post-test score while controlling for the pretest score. There were no outliers. Data was normally distributed for each group (Shapiro-Wilk test, $p > 0.05$) and the assumption of homogeneity of variance (Levene's test, $p > 0.05$) as well as homogeneity of regression slopes was met. Also, the preliminary analysis showed that the covariate did not differ between groups.

The covariate (pretest score) was significantly related to the post-test score, $F(1, 115) = 89.306$, $p < .001$, $r = .66$. There was also a significant effect of the training condition on the post-test score when controlling for the pretest score, $F(2, 115) = 4.42$, $p < 0.05$, partial $\eta^2 = 0.071$. According to Cohen [70], this corresponds to a medium effect.

Planned contrasts revealed that belonging to the interactive group significantly increased the post-test score in comparison to belonging to the reference group, $t(115) = -2,796$, $p < 0.01$, $r = 0.25$, $d = 0.521$. In contrast, belonging to the animation group did not significantly increase post-test performance in comparison to belonging to the reference group, $t(115) = 0.51$, $p > 0.05$.

C. Spatial ability and performance gain

For the examination of RQ3, Pearson's correlation coefficient was calculated to assess the relationship between the CRT score and the performance gain in the training groups. The three groups were analyzed separately.

In the animation group, there were no outliers in the data; also, the assumptions of linearity, homoscedasticity, and bivariate normality (Henze-Zirkler's test, $p > 0.05$) were fulfilled. There was no significant relationship found ($p > 0.05$).

In the interactive group, there were no outliers in the data; also, the assumptions of linearity, homoscedasticity and bivariate normality (Henze-Zirkler's test, $p > 0.05$) were fulfilled. Like in the animation group, no significant relationship was found ($p > 0.05$).

In the reference group, there were no outliers in the data; also, the assumptions of linearity, homoscedasticity and bivariate normality (Henze-Zirkler's test, $p > 0.05$) were fulfilled. Like for the other two groups, no significant relationship was found ($p > 0.05$).

D. Analysis of eye gaze pattern during training

To answer RQ4 for all three training conditions, the same area of interest (AOI) was defined. This AOI corresponds to the area within the red dashed line in Fig. 1. It contains the experimental setting and the diagram. The decision to choose this AOI was made for two reasons:

- While quasistatic stimuli are shown in the training of the interactive group and in the reference group, the training stimulus in the animation group moves dynamically between the laboratory system and the coordinate system. So that the eye movements of the three training conditions can be compared, the defined AOI must be suitable for all of them. The smallest suitable option is therefore the area marked in red in Fig. 1.
- The AOI should not be chosen larger as it should record only eye movements directly related to processing the training stimulus. Eye movements that enter or leave the area where the training stimulus is presented do not connect any causally related elements of the training and should therefore be filtered out by the choice of the AOI. For this reason, the AOI should be chosen as small as possible and should contain only the training stimulus.

As the eye tracking data was incomplete for six subjects, they were removed from the data set for further analysis. Those missing values were due to dropouts in the eye tracking system. The subjects were still almost evenly distributed across the groups ($N_{\text{animation}} = 38$, $N_{\text{interactive}} = 37$, $N_{\text{reference}} = 38$).

1. Analysis of the total number of fixations in the AOI

First the investigation focused on looking for differences in the total number of fixations in the AOI between the three groups. Data were not normally distributed for each group; therefore, a Kruskal-Wallis test was calculated. The total number of fixations was significantly affected by group

membership, $H(2) = 56.51$, $p < 0.001$. As before when analyzing differences in the processing time of the training, Mann-Whitney U tests were calculated to investigate the differences in the total number of fixations in the AOI in more detail. Again, a Bonferroni correction was taken into account.

A Mann-Whitney U test was calculated to determine if there were differences in the total number of fixations between the animation group and the reference group. The distributions did not differ between both groups (Kolmogorov-Smirnov, $p > 0.05$). There was a statistically significant difference in the median of the total number of fixations between the animation group ($Mdn = 201.5$) and the reference group ($Mdn = 88.0$), $U = 41.0$, $z = -7.08$, $p < 0.001$, $r = -0.81$. According to Cohen [70], this corresponds to a large effect.

Also, a Mann-Whitney U test was calculated to determine if there were differences in the total number of fixations between the interactive group and the reference group. The distributions did not differ between both groups (Kolmogorov-Smirnov, $p > 0.05$). There was a statistically significant difference in the median of the total number of fixations between the interactive group ($Mdn = 138.0$) and the reference group ($Mdn = 88.0$), $U = 282.5$, $z = -4.46$, $p < 0.001$, $r = -0.51$. According to Cohen [70], this corresponds to a large effect.

A Mann-Whitney U test was calculated to determine if there were differences in the total number of fixations between the animation group and the interactive group. The distributions did not differ between both groups (Kolmogorov-Smirnov, $p > 0.05$). There was a statistically significant difference in the median of the total number of fixations between the animation group ($Mdn = 201.5$) and the interactive group ($Mdn = 138.0$), $U = 341.0$, $z = -3.84$, $p < 0.001$, $r = -0.44$. According to Cohen [70], this corresponds to a medium effect.

2. Analysis of the saccadic distribution inside the AOI

As shown in the previous section, there are significant differences in the total number of fixations between the three training conditions. The three training conditions differ only in the training stimulus. The rest of the user interface is identical. Because of this common underlying visual structure, eye gaze patterns should show similarities in all of the three training conditions. However, the total number of fixations neither showed any of those structural similarities during training nor reflected the differences in test performance between training groups. To analyze similarities and differences of the eye pattern in more detail, the distribution of saccades, which led to the fixations, was investigated. Therefore, polar diagrams were created for each training task as follows.

The diagrams visualize saccades in dependence on their direction, length and number. According to their direction, saccades are classified in angular fields of 30° each. The saccadic length is encoded in the radial distance to the origin of the diagram. When visualizing saccades this way, at some positions in the diagram, the markers would overlap. To take this into account, colored markers were used to indicate the frequency of saccades which are of about the same length and direction. As shown above, the total number of fixations and, accordingly, the number of saccades that lead to these fixations differed between the training conditions. To improve the comparability of the diagrams, the number of saccades was normalized for each training condition. The color of the marked saccade corresponds to the value of a probability density function that describes the relative frequency of each saccade in a given training condition. These probability density functions of the saccades were estimated by using a kernel density estimation. In the following section, the diagrams of the first of the eight training tasks are discussed for each of the three training groups. All the other seven training tasks showed a similar structure.

Figure 8 presents the saccadic distributions during training task 1 for the three training conditions. It can be seen that in all training conditions, most saccades can be classified into the two angular fields of 345°–15° and 165°–195°. What is remarkable is that the saccadic distribution in these sections differs between the training conditions. While in the animation group the probability for having short saccadic amplitudes is higher than for having long saccadic amplitudes, in the interactive group and the reference group long saccadic amplitudes are more likely.

To analyze those differences in more detail, the saccadic distributions for the two angular fields (345°–15° and 165°–195°) were investigated. Those fields contained most of the saccades. Again, kernel density estimations were calculated for the saccades in each of the two angular fields for each of the three training groups and for all of the eight training tasks. In total 48 functions were calculated. For the first training task, the probability density functions in the angular field 345°–15° are shown in blue in Fig. 9. The figure additionally contains histograms showing, in gradations of one centimeter, the relative frequency of saccades in the corresponding length interval.

The probability density functions (Fig. 9) contain two local maxima at about the same saccadic lengths. Between those maxima, the probability density reaches a local minimum. This indicates that saccadic movements can be divided into two categories according to their length. A total of 42 of the 48 calculated functions showed this structure with a local minimum between the two local maxima. The remaining six functions, all belonging to the animation group, formed inflection points instead. This deviation was caused by the large number of short saccades compared to the number of long saccades in the animation group.

A threshold was defined to divide the saccades into two categories according to their length. For this purpose, the

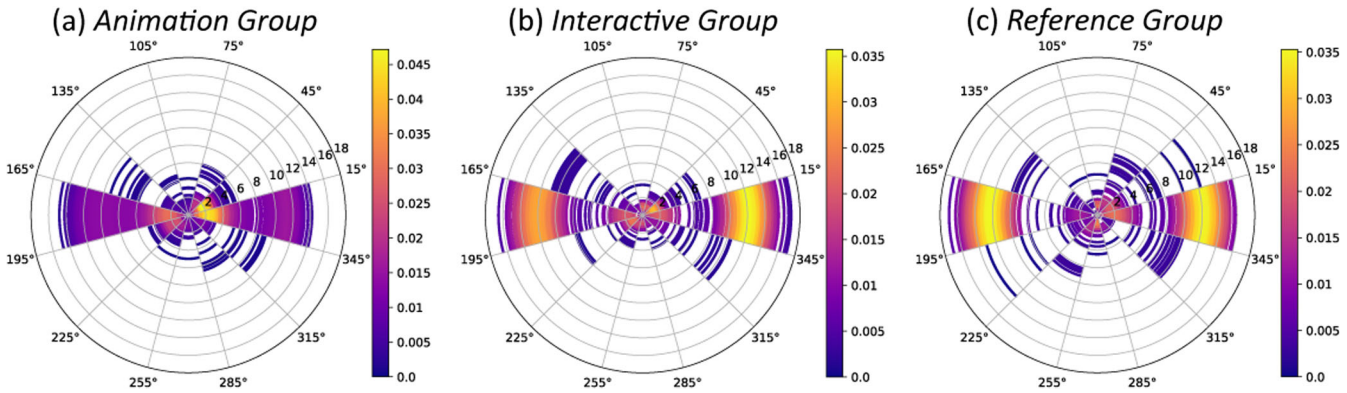

FIG. 8. The diagrams show the lengths and angular distributions of saccades during training task 1 in the animation group (a), interactive group (b), and reference group (c). Saccades were grouped into angular fields depending on their direction. The saccadic length in centimeters is encoded in the radial distance to the origin. The probability density of the saccades is color coded.

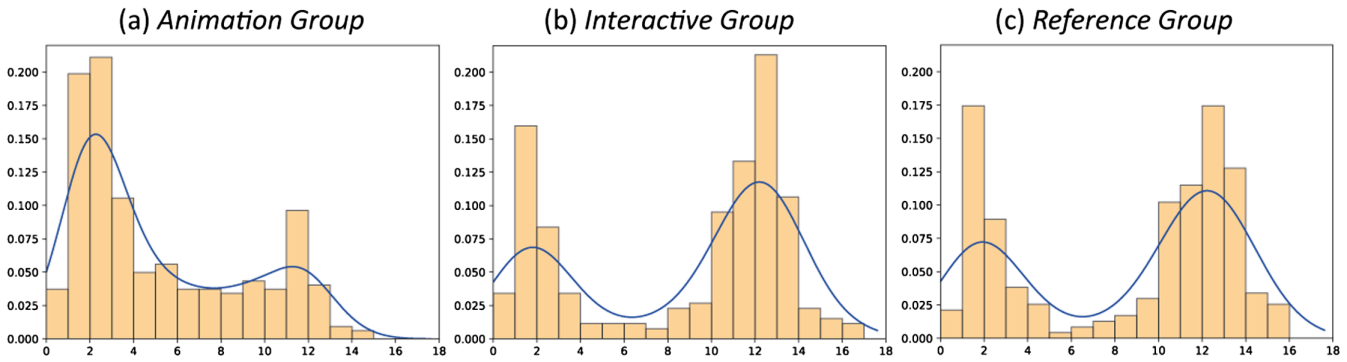

FIG. 9. The subfigures show the probability density function of saccades (blue line) in the angular field 345°–15° according to their saccadic length (in cm) during training task 1 in the animation group (a), interactive group (b), and reference group (c). While in the animation group the probability for shorter saccades is higher than for longer saccades, the interactive group and reference group show the opposite behavior.

mean was calculated from the positions of the 42 local minima ($M = 6.80$ cm, $SD = 1.71$ cm).

This division indicates further differences between the training groups. In the animation group, most of the saccades have a length less than 6.8 cm. In the interactive group and the reference group, saccadic lengths above this threshold are more likely. As shown before, the total number of fixations in the animation group was significantly higher than in the two other training groups. Consequently, most of those fixations in the animation group are the end points of saccades that have a length below 6.8 cm. Conversely, in the interactive group and the reference group, the probability is higher that the fixations are end points of saccades of length larger than 6.8 cm. As the interactive group performed best in the post-test, especially those saccades could indicate processes beneficial for improving the assignment task. This assumption would be reinforced if there were differences between the groups in the total number of saccades of length larger than 6.8 cm. As the diagrams were normalized, they do not give this information. For this reason, the next section deals with the question of whether significant differences exist between the training groups in the total number of saccades longer than 6.8 cm.

3. Analysis of the total number of saccades of length larger than 6.8 cm

The following examines the subjects' total number of saccadic movements with a length larger than 6.8 cm that occurred during the training tasks in the AOI. This total number was calculated by summing over all training tasks. As the total number of saccades inside the AOI that were longer than 6.8 cm was not normally distributed for each group, a Kruskal-Wallis test was calculated.

The total number of saccades longer than 6.8 cm was significantly affected by group membership, $H(2) = 47.482$, $p < 0.001$.

To examine the differences between the groups more closely, Mann-Whitney U tests with Bonferroni correction were carried out again (see above).

A Mann-Whitney U test was calculated to determine if there were differences in the total number of saccades longer than 6.8 cm between the animation group and the reference group. The distributions did not differ between both groups (Kolmogorov-Smirnov, $p > 0.05$). There was a statistically significant difference in the median of the total number of saccades longer than 6.8 cm between the animation group ($Mdn = 27.0$) and the reference group ($Mdn = 36.0$), $U = 398.5$, $z = -3.37$, $p < 0.01$, $r = -0.39$. According to Cohen [70], this corresponds to a medium effect.

Also, a Mann-Whitney U test was calculated to determine if there were differences in the total number of saccades longer than 6.8 cm between the interactive group and the reference group. The distributions did not differ between both groups (Kolmogorov-Smirnov, $p > 0.05$). There was a statistically significant difference in the median of the total number of saccades longer than 6.8 cm between the interactive group ($Mdn = 59.0$) and the reference group ($Mdn = 36.0$), $U = 245.0$, $z = -4.86$, $p < 0.001$, $r = -0.56$. According to Cohen [70], this corresponds to a large effect.

A Mann-Whitney U test was calculated to determine if there were differences in the total number of saccades longer than 6.8 cm between the animation group and the interactive group. The distributions did not differ between both groups (Kolmogorov-Smirnov, $p > 0.05$). There was a statistically significant difference in the median of the total number of saccades longer than 6.8 cm between the animation group ($Mdn = 27.0$) and the interactive group ($Mdn = 59.0$), $U = 124.5$, $z = -6.13$, $p < 0.001$, $r = -0.71$. According to Cohen [70], this corresponds to a large effect.

These differences therefore only partially reflect the assumption that the information acquired through long saccades is decisive for success in the training. The interactive group, whose training was the most successful, had the most long saccadic movements, as expected. The number of long saccadic movements in the animation group and the reference group differed, while the training was almost equally successful. A possible explanation for the deviation from the expectation is given in the discussion.

IV. DISCUSSION AND CONCLUSIONS

Results regarding RQ1 showed a significant positive correlation between the pretest score and the CRT score. This indicates that the pretest task is indeed related to the spatial thinking ability of the subjects. Especially students with weak spatial skills need more support when training in the documentation of measured values. However, the size of the correlation was small. A larger correlation coefficient was not to be expected as the task in the CRT was of lower complexity than the task in the pretest was: The CRT just measures the subjects' ability to transform two-dimensional figures into each other using rotation and reflection. Similarly, in the pretest, the laboratory system and the coordinate system must first be related to each other. In addition, the information obtained this way must be used to transfer the measuring position in the laboratory system to the correct position in the diagram. The results found indicate that spatial rotation processes are involved when solving the pretest task. According to Salomon [41], an external visualization could support the internal spatial translation process and contribute to a performance gain in the post-test. Therefore, as described in Sec. I A, considering animation as a possible supportive tool was justified.

The results regarding RQ2 show that the three training conditions differ in their influence on the post-test performance. Group membership affected the performance significantly with a medium effect size. Remarkably, the medium effect size resulted despite the short time on task during the intervention ($M = 108.5$ sec).

The animation group spent more time on tasks than the reference group did. Accordingly, this shows that the subjects took more time to complete the training. The processing depth, however, was not improved as the post-test performance of the animation group did not differ significantly from the results of the reference group. Therefore, the animation group did not profit significantly more from training with animations than the reference group did from inspecting the correctly marked positions. The intended better support from the animation could not be observed. Previous studies (see Sec. I A) reported similar observations. Animation often attracted attention but did not necessarily contribute to learning. Also, the transience of the animation limited the processing depth of the information. When discussing the results regarding RQ4, eye tracking data will help to interpret the inefficiency of the training in the animation group.

The interactive group spent more time on task than the reference group did. This indicates that subjects in the interactive group took more time to complete the training than subjects in the reference group did. In contrast to the animation group, the interactive group indeed outperformed the reference group significantly in the post-test. The presented information in the reference group and the interactive group concerning the correct point in the diagram did not differ in its content. The better performance of the interactive group therefore shows that the interactive task and the feedback helped to process the presented information more deeply.

The investigation of RQ3 showed that there is no relationship between the spatial rotation ability of the learners and the performance gain during the three trainings. This result suggests that, in the case under investigation, considering the spatial rotation abilities of the learners is not necessary when choosing the most suitable training.

The analysis of RQ4 showed similarities and differences in eye tracking measures during the three training conditions. Eye movements characteristic of the different trainings were identified. For the analysis, first the total number of fixations within the defined AOI was investigated. Results showed significantly more fixations during training in the animation group than in the interactive group and significantly more during training in the interactive group than in the reference group. This result is closely linked to the differences in processing time described above. However, processing time alone does not provide information about whether the content of the training was processed or whether the test subjects' gaze moved outside the user interface or even outside the screen. An animation that runs too slowly can, for example, due to boredom, lead the test subjects to observe objects that are not part of the training. Therefore, compared to the time on task alone, the number of fixations within the AOI gives more accurate insight. The high number of fixations in the animation group compared to the other two training conditions shows that the animation worked best to attract visual attention. In the interactive group, there were fewer fixations than in the animation group, however the number of fixations in this group was larger than in the reference group. Thus, compared to the reference group, also the training in the interactive group was successful in attracting visual attention.

Most of the saccadic movements during training can be assigned to the two angular fields from 345° to 15° and from 165° to 195° and correspond to eye movements in the horizontal direction. Thus, further analysis concentrated on the distribution of those saccadic movements. The structure of the calculated probability density functions (see Fig. 9 for the probability density functions of the first training task in the angular field 345° to 15°) indicated similarities and further differences in the eye movement between the three training conditions.

The number of saccadic movements was shown to consist of two components. For the division of the saccades into long saccadic movements and short saccadic movements, a saccadic length of 6.8 cm was derived from the dataset as a threshold. This length corresponds to approximately a quarter of the width of the AOI (half the distance between the origin of the laboratory system and the origin of the coordinate system in the diagram) (see Fig. 1).

Correspondingly, it is possible to differentiate between two characteristic saccadic movements:

- Short saccadic movements of length below 6.8 cm: These movements are more likely to occur within either the illustration of the experimental setup or the diagram. This can involve, for example, those eye movements that detect the position of the sensor in relation to the laboratory system. Such short saccadic movements are also required to follow the course of the animation. Because of their short length, these saccades affect matching processes between the experiment and diagram less.
- Long saccadic movements of length larger than 6.8 cm: These movements especially indicate matching processes between the experiment and the diagram. Such saccades are necessary, for example, to compare locations in the laboratory system of the experiment with positions in the coordinate system of the diagram.

The probability of long and short saccadic movements differed between the groups (see Fig. 9). The majority of the saccades in the animation group, in contrast to the interactive group and reference group, belong to the category of short saccadic movements. The recorded eye tracking data showed, that the reason for the increased occurrence of these movements is that the subjects in the animation group followed the course of the animation. In contrast, in the interactive group and the reference group, those saccades belonging to the group of long saccadic movements occurred more often than the shorter ones (see Fig. 9). Consequently, most of the fixations in the animation group were end points of short saccadic movements, while in the interactive group and the reference group, most of the fixations were end points of long saccadic movements. This shows that the three training conditions affected the visual attention of the test subjects differently. Since the training of the interactive group was the most successful and many of the long saccadic movements were observed in this group, it was assumed that especially the deeper processing of the information perceived through the longer saccadic movements is important for improving the assignment task during the training.

The statistical analysis of the number of saccadic movements of length above 6.8 cm showed that there are significantly more of them in the interactive group than there are in the reference group and that there are significantly more in the reference group than in the animation group. Thus, the interactive task in the interactive group led to the most eye movements comparing the experiment and the diagram. The immediate feedback encouraged further comparison between the experiment and diagram. The animation in the animation group, however, suppressed these eye movements in favor of the shorter saccadic movements with which the animation was followed. Interestingly, even despite the fact that the time on task in the animation group was longer than it was in the interactive group and the reference group, the number of saccadic movements of length of above 6.8 cm was smallest in the animation group.

It is therefore true for the interactive group that the highest number of long saccadic movements occurred in the most effective training. The subjects in this group had by far the most long saccadic movements ($Mdn = 59.0$).

On the other hand, when comparing the animation group and the reference group, the results are not so clear. The

reference group scored slightly better in the post-test than the animation group did, but this difference already existed in the pretest. Thus, no significant differences in training success were found between these two groups, while there were significantly more long saccadic movements in the reference group ($Mdn = 36.0$) than in the animation group ($Mdn = 27.0$). However, this difference in the number of long saccadic movements is far smaller than that between the interactive group and the reference group. That no significant difference in the effect of the training was found can therefore also be attributed to the number of long saccadic movements in the two groups not being sufficiently different.

In any case, the better performance of the interactive group shows that processing information obtained through the long saccadic movements seems to be particularly important for improving the assignment task. Long saccadic movements are necessary to compare positions in the laboratory system and positions in the coordinate system of the diagram. These results suggest that a higher number of saccadic movements leads to better integration of the information presented and thus to better performance. In previous research, this relationship was not always evident (see Sec. I A). As an explanation, it was assumed that the perceived elements are initially stored in memory. Subsequently, integration takes place without the need for comparative eye movements [62]. In the study presented in this paper, most long saccadic movements occurring in the most successful training could indicate that the two elements to be related to each other (laboratory system and diagram) were too complex to be stored in memory at the same time. This would necessitate saccadic eye movements to integrate the two elements.

However, it should be noted that it is not just about the number of saccadic movements but rather the active processing of the information obtained through them. This processing was facilitated by the immediate feedback. The subjects were immediately made aware of their errors, which prompted them to investigate the reason for their mistakes. A large number of saccadic movements without the corresponding cognitive activation of the test subjects would not necessarily have had the same effect.

For example, the passivity of the test subjects could be responsible for the poorer performance in the reference group. The test subjects are immediately shown the correct positions in the diagram. Passive test subjects may therefore not even be aware that they would have chosen an incorrect point on their own initiative.

During the training in the animation group, no indications were found that the animation might have helped to support the translation process between the laboratory system and diagram. However, this could be due to the animation not being similar enough to the internal cognitive translation process it should support (see Ref. [41]). In the analysis of the gaze data of the animation group, the longer saccadic movements necessary to link elements in the laboratory system and the coordinate system were rarely observed. Instead, many short saccadic movements took place to follow the dynamics of the animation. Attention was bound by the dynamics, which obviously did not lead to an increase in performance.

However, the results do not imply that animation is a poor teaching tool in general. Only for the described area of application, involving the most accurate documentation possible, were there no advantages over the other two training type. If only a visualization of the relationship between the laboratory system and the coordinate system in the diagram is intended, or if the gaze of the learners is to be guided, an animation can be the means of choice.

Overall, the training in the interactive group was best suited to support the documentation of measured vector quantities in diagrams when experimenting with computer-based experiments. An example showing a possible implementation in a virtual laboratory can be found in Ref. [71].

V. LIMITATIONS

For the investigation, the eye tracking data of the trainings in the animation group, the interactive group, and the reference group were compared. The trainings in the interactive group and the reference group show quasi-static training stimuli, while the training in the animation group contains a dynamic one. In general, it is difficult to compare eye tracking data from static and dynamic stimuli. In the present study, however, care was taken to ensure that the evaluated parameters are comparable. On the one hand, this applies to the total number of fixations as this quantity was used only as a measure of visual attention. On the other hand, when interpreting the saccadic movements in the animation group, the following was considered.

For saccadic movements in the animation group, three processes are responsible: saccadic movements within the animated element, saccadic movements between the dynamic and the static element of the display, and saccadic movements within the static element of the display.

When interpreting the results found, it was assumed that these three categories have the following influence on saccadic movements:

- Saccadic movements within the animated element: The animation dynamically transforms an entire part of the display, namely, the laboratory system of the experimental setup, into the part of the user interface which contains the diagram. Thus, in terms of saccadic length, eye movements that occur within the animated element are comparable to those that take place within the laboratory system. Eye movements that follow the animation should therefore belong to the category of short saccadic movements.
- Saccadic movements between the dynamic and the static elements of the display: The animation moves with uniform speed between the laboratory system and

the diagram. Therefore, saccadic movements that take place between the animated element and the laboratory system as well as between the animated element and the diagram were assumed to be distributed evenly over all lengths. Consequently, this effect increases the total number of observed saccadic movements but does not impair the distinction between the categories of long and short saccadic movements.

- Saccadic movements within the static element of the display: Because the animation is transparent, the underlying structure of the application can be seen at all times. Since this structure is the same as in the quasi-static trainings, such eye movements are comparable between all training groups.

The results show that the number of long saccadic movements was highest in the most successful training. The procedure that led to this statement was purely exploratory.

This is only a first hint that there is a relationship between the number of long saccadic movements and the increase in performance in the task under consideration. To verify this, further investigations are necessary. These should also be extended to other similar measurement documentation tasks (e.g., the documentation of measured values of the directional characteristic of a receiver dipole, of the directional characteristic of a speaker box, or of the spatial distribution of luminance for different LED lamps).

The results described were examined in a situation where the laboratory system and coordinate system could be converted into each other by using the geometric operations of rotation and translation. A generalization to more complex situations is conceivable but not a matter of course. Further research has to show whether the results found also apply if the axes are coded differently (for example, if a movement is visualized in a time-velocity diagram).

VI. FUTURE PROSPECTS

In the future, further research has to show if multimedia features in computer-based experiments can also support more general documentation processes. Additionally, the transferability of the results found to real experiments should be investigated. It is thinkable that augmented reality features in three-dimensional space could support the transfer between the laboratory system and the coordinate system of a diagram in a similar way as shown here for two-dimensional computer-based experiments.

[1] Committee on a Conceptual Framework for New K-12 Science Education Standards; National Research Council, *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (The National Academies Press, Washington, DC, 2012).

[2] N. D. Finkelstein, W. K. Adams, C. J. Keller, P. B. Kohl, K. K. Perkins, N. S. Podolefsky, S. Reid, and R. LeMaster, When learning about the real world is better done virtually: A study of substituting computer simulations for laboratory equipment, Phys. Rev. ST Phys. Educ. Res. **1**, 010103 (2005).

[3] G. Martínez, F. L. Naranjo, Á. L. Pérez, M. I. Suero, and P. J. Pardo, Comparative study of the effectiveness of three learning environments: Hyper-realistic virtual simulations, traditional schematic simulations and traditional laboratory, Phys. Rev. ST Phys. Educ. Res. **7**, 020111 (2011).

[4] V. Potkonjak, M. Gardner, V. Callaghan, P. Mattila, C. Guetl, V. M. Petrović, and K. Jovanović,Virtual laboratories for education in science, technology, and engineering: A review, Comput. Educ. **95**, 309 (2016).

[5] A. Hofstein and V. N. Lunetta, The laboratory in science education: Foundations for the twenty-first century, Sci. Educ. **88**, 28 (2004).

[6] S. Ainsworth, The functions of multiple representations, Comput. Educ. **33**, 131 (1999).

[7] S. Ainsworth, DeFT: A conceptual framework for considering learning with multiple representations, Learn. Instr. **16**, 183 (2006).

[8] R. Trumper, The Physics laboratory—a historical overview and future perspectives, Sci. Educ. **12**, 645 (2003).

[9] S. R. Singer, M. L. Hilton, and H. A. Schweingruber, *America's Lab Report* (National Academies Press, Washington, DC, 2006).

[10] S. Olson and D. G. Riordan, *Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics, Report to the President* (Executive Office of the President, Washington, DC, 2012).

[11] J. Kozminski, H. J. Lewandowski, N. Beverly, S. Lindaas, D. Deardorff, A. Reagan, R. Dietz, R. Tagg, M. Eblen-Zayas, J. Williams, R. Hobbs, and B. M. Zwickl, *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum* (AAPT, College Park, MD, 2014).

[12] N. G. Holmes and C. E. Wieman, Introductory physics labs: We can do better, Phys. Today **71**, No. 1, 38 (2018).

[13] M. Emden and E. Dumfleth, Assessing students' experimentation processes in guided inquiry, Int. J. Sci. Math. Educ. **14**, 29 (2016).

[14] H. Theyßen, M. Dickmann, K. Neumann, H. Schecker, and B. Eickhorst, Measuring experimental skills in large scale assessment: A simulation-based test instrument, in

Electronic Proceedings of ESERA 2015 Conference. Science Education Research: Engaging Learners for Sustainable Future, edited by J. Lavonen, K. Juuti, J. Lampiselkä, A. Uitto, and K. Hahl (co-ed. Dolin, Jens; Kind, Per) (University of Helsinki, Helsinki, 2016), pp. 598–606.

[15] E. M. Smith, M. M. Stein, C. Walsh, and N. G. Holmes, Direct Measurement of the Impact of Teaching Experimentation in Physics Labs, Phys. Rev. X **10**, 011029 (2020).

[16] S. Becker, P. Klein, A. Gößling, and J. Kuhn, Using mobile devices to enhance inquiry-based learning processes, Learn. Instr. **69**, 101350 (2020).

[17] M. C. Linn and A. C. Petersen, Emergence and characterization of sex differences in spatial ability: A meta-analysis, Child Development **56**, 1479 (1985).

[18] J. B. Carroll, *Human Cognitive Abilities* (Cambridge University Press, New York, NY, 1993).

[19] N. S. Newcombe and T. F. Shipley, Thinking about spatial thinking: New typology, new assessments, in *Studying Visual and Spatial Reasoning for Design Creativity*, edited by J. S. Gero (Springer Netherlands, Dordrecht, 2015), pp. 179–192.

[20] M. Cole, C. Cohen, J. Wilhelm, and R. Lindell, Spatial thinking in astronomy education research, Phys. Rev. Phys. Educ. Res. **14**, 010139 (2018).

[21] R. B. Ekstrom, J. W. French, H. H. Harmann, and D. Dermen, *Manual for Kit of Factor-Referenced Cognitive Tests* (Educational Testing Service, Princeton, NJ, 1976).

[22] J. C. Castro-Alonso, P. Ayres, M. Wong, and F. Paas, Visuospatial tests and multimedia learning, in *Advances in Cognitive Load Theory*, edited by S. Tindall-Ford, S. Agostinho, and J. Sweller (Routledge, New York, 2019), pp. 89–100.

[23] S. G. Vandenberg and A. R. Kuse, Mental rotation, a group test of three-dimensional spatial visualization, Percept. Mot. Skills **47**, 599 (1978).

[24] M. Peters, B. Laeng, K. Latham, M. Jackson, R. Zaiyouna, and C. Richardson, A redraw Vandenberg and Kuse mental rotations test: Different versions and factors that affect performance, Brain Cognit. **28**, 39 (1995).

[25] M. Hegarty and V. K. Sims, Individual differences in mental animation during mechanical reasoning, Mem. Cogn. **22**, 411 (1994).

[26] M. Kozhevnikov, M. Hegarty, and R. E. Mayer, Revising the visualizer-verbalizer dimension: Evidence for two types of visualizers, Cognit. Instr. **20**, 47 (2002).

[27] M. Kozhevnikov and R. Thornton, Real-time data display, spatial visualization ability, and learning force and motion concepts, J. Sci. Educ. Technol. **15**, 111 (2006).

[28] M. Stieff, Mental rotation and diagrammatic reasoning in science, Learn. Instr. **17**, 219 (2007).

[29] M. Kozhevnikov, M. A. Motes, and M. Hegarty, Spatial visualization in physics problem solving, Cogn. Sci. **31**, 549 (2007).

[30] S. Küchemann, P. Klein, H. Fouckhardt, S. Gröber, and J. Kuhn, Students' understanding of non-inertial frames of reference, Phys. Rev. Phys. Educ. Res. **16**, 010112 (2020).

[31] B. B. de Koning, H. K. Tabbers, R. M. J. P. Rikers, and F. Paas, Towards a framework for attention cueing in instructional animations: guidelines for research and design, Educ. Psychol. Rev. **21**, 113 (2009).

[32] T. N. Höffler and D. Leutner, Instructional animation versus static pictures: A meta-analysis, Learn. Instr. **17**, 722 (2007).

[33] M. Bétrancourt and B. Tversky, Effect of computer animation on users' performance: A review, Trav. Hum. **63**, 311 (2000).

[34] S. Berney and M. Bétrancourt, Does animation enhance learning? A meta-analysis, Comput. Educ. **101**, 150 (2016).

[35] R. K. Lowe and W. Schnotz, Animation Principles in Multimedia Learning, in *The Cambridge Handbook of Multimedia Learning*, 2nd ed., edited by R. Mayer (Cambridge University Press, New York, 2014), pp. 513–546.

[36] R. A. Abrams and S. E. Christ, Motion onset captures attention, Psychol. Sci. **14**, 427 (2003).

[37] A. P. Hillstrom and Y.-C. Chai, Factors that guide or disrupt attentive visual processing, Comput. Hum. Behav. **22**, 648 (2006).

[38] S. Kriz and M. Hegarty, Top-down and bottom-up influences on learning from animations, Int. J. Hum-Comput. St. **65**, 911 (2007).

[39] J. R. Kirby, Mental representations, cognitive strategies, and individual differences in learning with animations: Commentaries on sections one, and two, in *Learning with Animation*, edited by R. Lowe and W. Schnotz (Cambridge University Press, New York, 2008), pp. 165–182.

[40] B. B. de Koning, H. K. Tabbers, R. M. J. P. Rikers, and F. Paas, Attention guidance in learning from a complex animation: Seeing is understanding?, Learn. Instr. **20**, 111 (2010).

[41] G. Salomon, *Interaction of Media Cognition and Learning* (Erlbaum, Hillsdale, NJ, 1979).

[42] K. A. Gilligan, M. S. C. Thomas, and E. K. Farran, First demonstration of effective spatial training for near transfer to spatial performance and far transfer to a range of mathematics skills at 8 years, Dev. Sci. **23**, e12909 (2019).

[43] M. T. H. Chi and R. Wylie, The ICAP framework: Linking cognitive engagement to active learning outcomes, Educ. Psychol. **49**, 219 (2014).

[44] T. Feldmann, *Multimedia* (Chapman & Hall, London, 1994).

[45] R. E. Clark and D. F. Feldon, Ten common but questionable principles of multimedia learning, in *The Cambridge Handbook of Multimedia Learning*, edited by R. Mayer (Cambridge University Press, Cambridge, 2014), pp. 151–173.

[46] S. Domagk, R. N. Schwartz, and J. L. Plass, Interactivity in multimedia learning: An integrated model, Comput. Hum. Behav. **26**, 1024 (2010).

[47] K. R. Koedinger and V. Aleven, Exploring the assistance dilemma in experiments with cognitive tutors, Educ. Psychol. Rev. **19**, 239 (2007).

[48] V. J. Shute, Focus on formative feedback, Rev. Educ. Res. **78**, 153 (2008).

[49] J. Hattie and H. Timperley, The power of feedback, Rev. Educ. Res. **77**, 81 (2007).

[50] F. M. van der Kleij, T. J. H. M. Eggen, C. F. Timmers, and B. P. Veldkamp, Effects of feedback in a computer-based assessment for learning, Comput. Educ. **58**, 263 (2012).

[51] H. M. Oonk and R. A. Abrams, New perceptual objects that capture attention produce inhibition of return, Psychon. Bull. Rev. **5**, 510 (1998).

[52] M.-L. Lai, M.-J. Tsai, F.-Y. Yang, C.-Y. Hsu, T.-C. Liu, S. W.-Y. Lee, M.-H. Lee, G.-L. Chiou, J.-C. Liang, and C.-C. Tsai, A review of using eye-tracking technology in exploring learning from 2000 to 2012, Educ. Res. Rev. **10**, 90 (2013).

[53] E. Alemdag and K. Cagiltay, A systematic review of eye tracking research on multimedia learning, Comput. Educ. **125**, 413 (2018).

[54] P. Klein, J. Viiri, S. Mozaffari, A. Dengel, and J. Kuhn, Instruction-based clinical eye-tracking study on the visual interpretation of divergence: How do students look at vector field plots?, Phys. Rev. Phys. Educ. Res. **14**, 010116 (2018).

[55] P. Klein, S. Küchemann, S. Brückner, O. Zlatkin-Troitschanskaia, and J. Kuhn, Student understanding of graph slope and area under a curve: A replication study comparing first-year physics and economics students, Phys. Rev. Phys. Educ. Res. **15**, 020116 (2019).

[56] P. Klein, J. Viiri, and J. Kuhn, Visual cues improve students' understanding of divergence and curl: Evidence from eye movements during reading and problem solving, Phys. Rev. Phys. Educ. Res. **15**, 010126 (2019).

[57] M. A. Just and P. Carpenter, A theory of reading: from eye fixations to comprehension, Psychol. Rev. **87**, 329 (1980).

[58] K. Holmqvist and R. Andersson, *Eye Tracking: A Comprehensive Guide to Methods, Paradigms, and Measures* (Lund Eye-Tracking Research Institute, Lund, 2017).

[59] C. I. Johnson and R. E. Mayer, An eye movement analysis of the spatial contiguity effect in multimedia learning, J. Exp. Psychol. Appl. **18**, 178 (2012).

[60] L. Mason, M. C. Tornatora, and P. Pluchino, Do fourth graders integrate text and pictures in processing and learning from an illustrated science text? Evidence from eye-movement patterns, Comput. Educ. **60**, 95 (2013).

[61] P. A. O'Keefe, S. M. Letourneau, B. D. Homer, R. N. Schwartz, and J. L. Plass, Learning from multiple representations: An examination of fixation patterns in a science simulation, Comput. Hum. Behav. **35**, 234 (2014).

[62] K. Scheiter and A. Eitel, Signals foster multimedia learning by supporting integration of highlighted text and diagram elements, Learn. Instr. **36**, 11 (2015).

[63] A. Schüler, Investigating gaze behavior during processing of inconsistent text-picture information: Evidence for text-picture integration, Learn. Instr. **49**, 218 (2017).

[64] N. Ott, R. Brünken, M. Vogel, and S. Malone, Multiple symbolic representations: The combination of formula and text supports problem solving in the mathematical field of propositional logic, Learn. Instr. **58**, 88 (2018).

[65] C. Hoyer and R. Girwidz, A remote lab for measuring, visualizing and analysing the field of a cylindrical permanent magnet, Eur. J. Phys. **39**, 065808 (2018).

[66] D. D. Salvucci and J. H. Goldberg, Identifying fixations and saccades in eye-tracking protocols, in *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (Association for Computing Machinery, New York, 2000), pp. 71–78.

[67] L. J. Burton and G. J. Fogarty, The factor structure of visual imagery and spatial abilities, Science **31**, 289 (2003).

[68] D. M. Dimitrov and P. D. Rumrill, Jr., Pretest-posttest designs and measurement of change, Work **20**, 159 (2003), https://content.iospress.com/articles/work/wor00285.

[69] G. W. Divine, H. J. Norton, A. E. Barón, and E. Juarez-Colunga, The Wilcoxon–Mann–Whitney procedure fails as a test of medians, Am. Stat. **72**, 278 (2018).

[70] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Erlbaum, Hillsdale, NJ, 1988).

[71] https://www.didaktik.physik.uni-muenchen.de/sims/documentation/.