Graphical representations of data improve student understanding of measurement and uncertainty: An eye-tracking study

Ana Susac,[1,2,*] Andreja Bubic,[3] Petra Martinjak,[1] Maja Planinic,[1] and Marijan Palmovic[4]

[1]*Department of Physics, Faculty of Science, University of Zagreb, Bijenicka 32, 10000 Zagreb, Croatia*
[2]*Department of Applied Physics, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 10000 Zagreb, Croatia*
[3]*Chair for Psychology, Faculty of Humanities and Social Sciences, University of Split, Sinjska 2, 21000 Split, Croatia*
[4]*Laboratory for Psycholinguistic Research, Department of Speech and Language Pathology, University of Zagreb, Borongajska cesta 83h, 10000 Zagreb, Croatia*

(Received 10 January 2016; revised manuscript received 28 July 2017; published 31 October 2017)

Developing a better understanding of the measurement process and measurement uncertainty is one of the main goals of university physics laboratory courses. This study investigated the influence of graphical representation of data on student understanding and interpreting of measurement results. A sample of 101 undergraduate students (48 first year students and 53 third and fifth year students) from the Department of Physics, University of Zagreb were tested with a paper-and-pencil test consisting of eight multiple-choice test items about measurement uncertainties. One version of the test items included graphical representations of the measurement data. About half of the students solved that version of the test while the remaining students solved the same test without graphical representations. The results have shown that the students who had the graphical representation of data scored higher than their colleagues without graphical representation. In the second part of the study, measurements of eye movements were carried out on a sample of thirty undergraduate students from the Department of Physics, University of Zagreb while students were solving the same test on a computer screen. The results revealed that students who had the graphical representation of data spent considerably less time viewing the numerical data than the other group of students. These results indicate that graphical representation may be beneficial for data processing and data comparison. Graphical representation helps with visualization of data and therefore reduces the cognitive load on students while performing measurement data analysis, so students should be encouraged to use it.

DOI: 10.1103/PhysRevPhysEducRes.13.020125

I. INTRODUCTION

Measurement is the basis of the scientific method, and as such is fundamental for student understanding of experimental work. This work is very complex and consists of experimental design, data collection, data analysis, and interpretation of the obtained results. Understanding of measurement and measurement uncertainty is crucial for all phases of experimental work and, consequently, these concepts are introduced through physics laboratories and statistics courses. The professional association of physics teachers emphasizes that "students should learn enough about uncertainties to understand the inherent limitations of measurement processes" [1]. Several physics education research (PER) studies focused on student understanding of measurement uncertainty and their ability to process and compare experimental data [2–17].

*Corresponding author.
ana.susac@fer.hr

Published by the American Physical Society under the terms of the Creative Commons Attribution 3.0 License. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Previous studies have shown that students worldwide have similar difficulties with understanding of measurement and the associated uncertainty [2–17]. Some students, especially at a younger age, think there is no need for repeated measurements [2,5–9,16,17]. Many students believe that it is possible to make a perfect measurement and that such "true values" may be determined by making more measurements [2–5,7–9,17]. However, as a final result they usually report only the arithmetic mean and take into account only mean values when comparing data sets [3–5,7–13,15–17]. Overall, students rarely consider uncertainty of the measurement. They often do not distinguish between random and systematic errors, or between precision and accuracy of measurement [2,4,14,15,17].

Based on the results of their studies, researchers from the University of York, UK and the University of Cape Town, South Africa developed the Physics Measurement Questionnaire to probe students' reasoning about measurement [18]. It was a basis for many studies mentioned above. Other assessment tools have also been proposed, such as the Concise Data Processing Assessment [14] and the Measurement Uncertainty Quiz [19].

A probabilistic interpretation of measurement is recommended by the Joint Committee for Guides in Metrology [20].

This probabilistic approach has also been suggested for teaching measurement in the introductory physics laboratory [21]. The researchers argue that besides "highlighting the uncertain and tentative, yet quantifiable, nature of scientific knowledge," the introduction to probabilistic language may be useful for other areas of physics such as quantum mechanics and statistical mechanics [21]. The evaluations of the courses based on the probabilistic approach indicated that they are more effective than traditional courses for learning measurement and uncertainty in the introductory laboratory [22,23].

One of the skills that students are expected to develop in physics laboratories is to graphically represent the obtained data. Numerous PER studies have reported students' difficulties with the interpretation of graphs, e.g., Refs. [24–29]. When asked to graphically represent the measurement data, students are usually able to do so. However, when students were asked to compare two data sets in one study, none of the more than 200 students drew a graphical representation of the data with error bars [15]. A short survey on students' criteria for agreement between measured values with graphical representation of data was administered to two small groups of teaching assistants ($N = 11$) and students ($N = 12$) at the University of North Carolina at Chapel Hill. In addition, one question on comparison of two data sets with graphical representations was given to another group of students ($N = 44$) at the North Carolina State University. The authors concluded that the overlap method for the comparison of two data sets is intuitive for both undergraduate and graduate students. The overlap method refers to consideration of the overlapping confidence intervals. However, the authors admit that the question "How effective is the graphical error bar representation at getting students to use the uncertainty of their measurements to draw a valid conclusion about the agreement or difference between two values?" was raised but not fully answered in their study [15].

Although data graphing seems to be a useful method for comparing the values and the uncertainty intervals, this topic has not yet been systematically investigated in PER. In this study we wanted to explore if, and when, graphical representation of the measurement data can help in interpreting that data.

Representing data graphically might be useful in visualization, i.e., forming a mental image of data that could help in better understanding and comparing of data. Previous PER studies have shown that multiple representations could be beneficial for student understanding of physics and problem solving in general, but they also pose significant challenges to students, e.g., Refs. [30–32]. Based on previous findings [15], we hypothesized that additional graphical representation would be beneficial for student understanding of measurement. Cognitive load theory suggests that learning can effectively occur only if information is provided in such a way that it does not "overload" the mental capacity [33]. Graphical representation of data might reduce the cognitive load of measurement analysis compared to only numerical representation of data, thus leaving more cognitive resources available for processing different aspects of measured data. For example, if students are asked to compare two data sets, graphical representation would facilitate grasping of the data and enable its further analysis, i.e., evaluation of whether the intervals overlap.

For the purpose of this study, we have developed a short test covering basic measurement skills such as reporting measurement results, treating outliers, comparing two measurements, and differentiating between accuracy and precision. All test items were multiple-choice questions, and some of them were two tiered, i.e., in addition to answering the question regarding the abovementioned issues, students were asked why they had selected a particular response. Besides numerical values of data, their graphical representation on a number line was also given in one version of the test. In the second version of the test neither a graphical representation of data nor a number line were presented. Since many researchers have emphasized that measurements and uncertainty should be addressed as early as possible in physics teaching, e.g., Refs. [4,15], we did not specify the measurement uncertainty in the test as a standard deviation so that the test could be administered to high school students too [3,34]. As most of the previous research on student understanding of measurement and uncertainty was done on the interpretation of single variable data [2–17], we have also used only single variable data with and without number line representation in this study. Thus, the reader should bear in mind that throughout this article the term "graphical representation of data" refers to the number line representation of single variable data.

In addition to administering this test in a paper-and-pencil form, in the present study we also recorded students' eye movements while they were solving the test. In the last few decades, eye movements were often used as a probe of attention because they offer a much more direct and accurate measure of visual attention than traditional manual reaction time measures [35] or participants' reports [36]. Measurement of eye movements can be used for dissociating different stages in problem solving by analyzing temporal characteristics of eye-tracking data [37] or finding critical areas in problem display attended by successful problem solvers by analyzing spatial characteristics of eye-tracking data [38]. In addition, eye-tracking data can give insight on how long participants process certain information, thus providing an indirect measure of cognitive load.

Recently, eye-tracking technology started to be more widely used in educational studies in mathematics and science [36,39–48]. In a few PER studies, eye movements were measured to explore where students allocate visual attention during problem solving [41–43] in order to develop visual cues that might help them in the process [44]. Eye tracking was also used to gain insight into students' strategies in solving multiple-choice science

problems [49,50]. Analysis of eye-tracking data of the preservice science teachers with different levels of expertise in physics, chemistry, and biology revealed differences in eye-movement patterns across different science disciplines and similarities among participants with similar science backgrounds [49]. While solving multiple-choice science problems students paid more attention to chosen options than to rejected alternatives, and successful participants inspected longer the relevant factors in problems than the irrelevant ones [50]. Correspondingly, in the present study we used eye tracking to measure if, and how long, students pay attention to graphical representations of data.

II. RESEARCH QUESTIONS

In this paper we aim to answer the following research questions:

(i) Does the graphical representation of data help students in interpreting the measured data and the related uncertainty?

(ii) How does the graphical representation of data influence participants' eye movements while solving problems concerning the analysis of the measured data and the related uncertainty?

We performed two studies to address these research questions.

Based on the previously reported research results in this field [15], our hypothesis was that the graphical representation of data helps in interpreting the measured data and the related uncertainties. In the reported study, small groups of participants answered one or two questions on comparison of two data sets with graphical representation. In our study, we used more questions and more participants to experimentally answer our first research question. Additionally, we wanted to compare Croatian students' difficulties with measurements and the related uncertainty to research results from other countries. Our hypothesis was that Croatian students will show similar difficulties as students in other countries.

The second research question had a more exploratory character since no eye-tracking studies on understanding measurement and measurement uncertainty had been previously reported. However, based on the cognitive load theory, we hypothesized that students would pay attention to both numerical and graphical representation of data, and we expected that data shown in both formats would reduce cognitive load compared to data in numerical format only, which would have positive effect on data processing and task completion.

III. STUDY 1: PAPER-AND-PENCIL ASSESSMENT OF STUDENT UNDERSTANDING OF MEASUREMENT

A. Methods

1. Participants

The study included 101 undergraduate students from the Department of Physics, University of Zagreb. About half of the participants (48) were first-year students and they were tested during their first semester, so they did not have any prior experience with university physics laboratory. The rest of the participants were senior-year (mostly third and fifth year) students (53) who had previously completed at least two physics laboratory courses; some of them attended introductory statistics course as well. All students were prospective physics teachers.

Physics is taught as a compulsory subject in the last two grades of all elementary schools and throughout four years of most high schools in Croatia. Physical measurements are discussed in school physics classes (e.g., a need for multiple measurements) but only in relatively few schools do students actually perform measurements.

2. Materials

Eight multiple-choice test items were constructed by the authors. Some were new and some were modified questions from the previous studies [5,16,17] and the Physics Measurement Questionnaire (PMQ) [18]. The complete test is given in Appendix A.

The experiment with the dropping of a ball in the sand was chosen as the context for the first five questions for its simplicity [17]. The experimental situation is described and illustrated (see Appendix A). Students release balls without initial velocity from the same height and measure the diameter of the mark in the sand. Item 1 probes student understanding of the mean of measured values as the best representation of a set of measurements and their treatment of an outlier; it was modified from question 4 from the PMQ [18] and question 3 (treatment of outlier, p. 31) from Ref. [17]. As students discussed rough measurement errors (mistakes) during instruction, they were expected to recognize an outlier in the measurement data and omit it while calculating the mean. In item 2, students are asked to recognize that it is not possible to find "a true value" of the measured quantity. This item is to a certain degree similar to question 3 from the PMQ [18]. In item 3, students are asked about the quality of the measured data, i.e., they are supposed to compare the dispersion of the measured data. Item 3 was written on the basis of question 5 from the PMQ [18], question 2 from the Multiple-Choice Survey, p. 234 [16], and question 4 (same mean, different spread, p. 32) from Ref. [17]. In items 4 and 5, two data sets are compared—one where there is a significant overlap, and another where there is no overlap between the two sets. Item 4 was modified from question 6 from PMQ [18] and question 3 from the Multiple-Choice Survey, p. 235 [16], whereas item 5 was based on question 4 from the Multiple-Choice Survey, p. 236 [16].

The first five items are two tiered, i.e., after giving an answer to a multiple-choice question, students are asked to justify their choice. Within the multiple-choice questions the distractors are based on typical answers from the previous studies [5,16,17] but students can also choose the answer "other" and give their own explanation. We

modified questions from the previous studies so that the values of the measurements are small numbers that are easy to calculate, even without a calculator.

The rest of the items were newly constructed for this test. The context for the last three items is the measurement of the *free-fall acceleration g*. Items 6 and 7 probe student understanding of the terms accuracy and precision. We asked students to rate the accuracy and precision for two sets of measurement—precise but not accurate (item 6) and accurate but not precise (item 7). The distractors are four possible answers (accurate and precise, accurate and imprecise, precise and inaccurate, inaccurate imprecise). Item 8 deals with reporting the final result of the measurement and the treatment of significant digits, and it is also two tiered. The distractors are based on the typical students' answers from our teaching experience.

We prepared two versions of the test, with and without graphical representation of data. Within both versions, item 8 was used as a control item as it did not include a graphical representation of data in either version. By comparing two groups' scores on the test item 8 we could check the presence of potential differences between groups as such. For the first seven items, the graphical representation of the measurements on a number line was also provided in one version of the test.

3. Procedure

The test was administered to students during their regular classes in Fundamentals of Physics (first year students), Electrodynamics (mostly third year students), and Methods of Physics Teaching (mostly fifth year students). About half of the participants (49) were given the test with graphical representation of data, whereas the remaining participants were given the same test without graphical representation. There was no time limit to answer the 8 items (14 multiple-choice questions) but it took students usually 15–20 min to complete the test. Each question was presented on one page and participants were asked to work questions in the order of presentation in the booklet. Two authors carefully observed students taking the test, and they did not notice any students answering the questions in different order. One of the authors was an instructor for the course Methods of Physics Teaching, another author was a lab instructor (for fourth year), whereas the third author was a student of the fifth year collecting data for her diploma thesis. The remaining two authors were not associated with students in any manner.

4. Data analysis

The test was scored independently by two authors. On two-tiered items, if a correct answer was given with a correct explanation, the student was awarded 2 points. If students had chosen the correct answer and gave their own incomplete explanation, they were awarded 1 point. If a correct answer was given with a wrong explanation, the student was awarded 0 points. The maximum score was 14 points.

The agreement between the raters was very high (almost 100%) because students had selected their answers out of the multiple choices in most cases (96% of all answers). In the remaining 4% of all answers students gave their own explanation, most often on item 2b. In only seven out of 1414 student responses the raters did not initially agree. The differences in scoring have been discussed and consensually resolved. Here are examples of explanations given by students who have chosen the correct answer (d) on item 2 (see Appendix A) and the corresponding scores:

> *"The solution is somewhere in between due to the impossibility of exact determination." (correct explanation, 2 points)*
>
> *"I did not want to take the mean value. The ball has a diameter between 18 and 26 mm." (incomplete explanation, 1 point)*
>
> *"The value must be in the range of 18–26 mm because these are min and max values. More precisely, we can add up all the values and divide by the number of measurements and get 22 mm." (incorrect explanation, 0 points)*

We converted individual point scores on each test item to the percentage scores dividing given points by maximal number of points. To determine the effects of the year of study (first vs senior) and the graphical representation (with vs without graphical representation) on students' scores, a two-way analysis of variance (ANOVA) was conducted. The data that were analyzed using ANOVAs and reported in the manuscript satisfied the assumptions required for conducting ANOVAs. Distributions of data and residuals were normal, and the homogeneity of variance was tested using Levene's test that was not significant. The chi-square test was used for comparing the scores on individual test items between groups with and without graphical representation. A threshold of $p = 0.05$ was used for determining the level of effect significance within all conducted tests.

B. Results

1. Analysis of students' scores

The mean score (and standard deviation) on the paper-and-pencil test was $(49 \pm 20)\%$. The distribution of scores for the whole sample of students is shown in Fig. 1. The Shapiro-Wilk test showed that the distribution was normal ($W = 0.98$, $p = 0.08$). The largest number of students had test score between 41% and 60%. Only 3 students scored higher than 80%, whereas 10 students scored less than 20%. The results indicate that the test difficulty was adequate for the tested sample of students.

To compare the scores of students who solved tests with and without graphical representation of data across different years of study, we conducted a two-way ANOVA on average scores for test items 1–7, with factors being

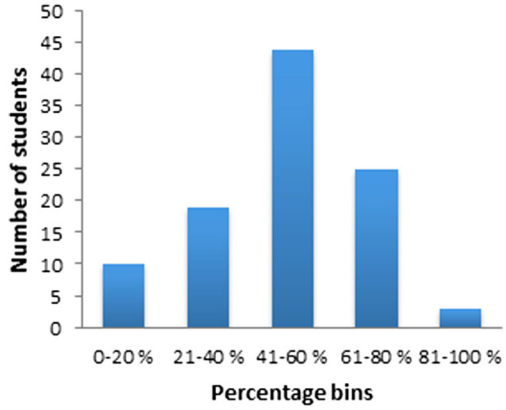


FIG. 1. Distribution of students' scores on paper-and-pencil test.

graphical representation (with vs without graphical representation) and year of study (first vs senior). The obtained results showed a statistically significant main effect of both factors, graphical representation [$F(1,97) = 10.21$, $p < 0.01$, $\eta_p{}^2 = 0.095$] and year of study [$F(1,97) = 8.09$, $p < 0.01$, $\eta_p{}^2 = 0.077$], whereas the interaction effect was not significant [$F(1,97) = 0.65$, $p > 0.05$, $\eta_p{}^2 = 0.077$]. Students who solved the test with graphical representation of data had higher scores than their peers who solved the test without graphical representation (Fig. 2). Senior-year students scored higher than first-year students.

Test item 8 did not have graphical representation of data in either group, so it was used as a control item. A corresponding two-way ANOVA on average scores for test item 8, with factors graphical representation (i.e., assignment to one of the two groups, with vs without graphical representation) and year of study was performed to test for group differences. A significant main effect of year of study [$F(1,97) = 15.35$, $p < 0.001$, $\eta_p{}^2 = 0.137$]

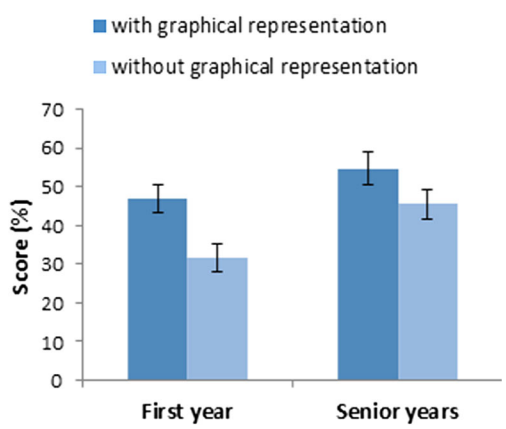


FIG. 2. Scores on paper-and-pencil test for first-year students and senior-year students divided into groups with graphical representation of data and without it. Average scores for test items 1–7 are shown. The error bars represent 1 SEM (standard error of mean).

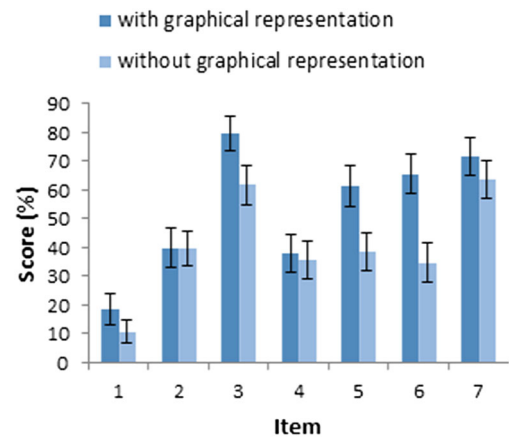


FIG. 3. Scores of all students on seven test items in study 1 divided into groups with graphical representation of data and without it. The error bars represent 1 SEM.

was found, whereas the effect of graphical representation [$F(1,97) = 1.27$, $p > 0.05$, $\eta_p{}^2 = 0.013$] and the interaction effect [$F(1,97) = 0.139$, $p > 0.05$, $\eta_p{}^2 = 0.001$] were not significant. Thus, random assignment to one of the two groups (with and without graphical representation) was successful, i.e., the groups did not differ *per se*.

Furthermore, we wanted to explore the effect of graphical representation on students' scores for each test item (Fig. 3). The test item 1 was the most difficult in the test. Students had the best scores on test items 3 and 7. The chi-square test revealed statistically significant differences in scores between groups with and without graphical representation for test item 5 [$\chi^2(2) = 7.267$, $p = 0.026$] and item 6 [$\chi^2(2) = 9.506$, $p = 0.002$], whereas no significant differences were revealed for other test items. When p values were adjusted for seven comparisons the difference between the two groups was statistically significant only for item 6 ($p = 0.014$).

2. *Analysis of students' responses*

Furthermore, we wanted to investigate in more detail students' difficulties with measurements and the related uncertainty. We analyzed responses of all participants ($N = 101$). The average scores of all students are given in the following paragraphs. The distributions of students' responses across distractors for all test items (questions and explanations) are shown in Appendix B (Fig. 11). We report here the main findings.

Item 1 was the most difficult item in the test (Fig. 3); students' average score on it was 14%. Students were asked which number was the best representation of a set of measurements. They were supposed to recognize an outlier and calculate the average of the remaining measurement values. Only 18% of the participants chose the correct answer "(a) 22 mm," whereas 26% selected the correct explanation "(b) This number is obtained if measurement

40 mm is ignored, then remaining measurements summed and divided by 5." More students chose the correct explanations than the correct answer itself probably because some students had not correctly calculated the average, whereas others had not recognized the outlier at first but after being prompted with multiple-choice explanations they realized that the outlier should be ignored when calculating the average. Most students chose the answer "(c) 25 mm" (47%) and explanation that this number is obtained if all measurements are summed and divided by 6 (49%). The second most popular answer was "(b) 23 mm" (30%). 11% of student selected the explanation "(c) This number appeared twice in the measurements, whereas the others appeared only once" and additional 9% chose "(d) This number is in the middle of the measurement results".

Item 2 addressed the student understanding of the measurement uncertainty. About half of the students (55%) chose the correct answer "(d) The measured quantity is somewhere between 18 and 26 mm" and 39% gave the correct explanation "(d) We can never know the true value of the measured quantity". The most popular incorrect answer was "(c) The measured quantity is somewhere between 18 and 23 mm" (20%), followed by "(a) The measured quantity is 22 mm" (15%). The explanation "(a) This number is obtained if all measurements are summed and divided by 5" was selected by 29% of students. On this test item the largest number of students gave explanation in their own words (23%), mostly repeating the statement that they have already chosen, i.e., that the measured quantity is somewhere in the interval between the smallest and the largest measured value.

Item 3 was the easiest item in the test (Fig. 3); the average score on it was 70%. The students were asked to compare the quality of the measurement results of the two groups with the same average but different spread. The correct answer "(a) The results of group A are better than the results of group B" was selected by 84% of students and 61% also selected the correct answer when comparing the intervals of the measured values ["(b) The results of group *A* are between 20 and 30 mm, and the results of group *B* are between 11 and 41 mm"]. Some students (11%) gave an explanation in their own words, mostly correctly indicating the different spread of data for the two groups.

Item 4 asked students if the results of the two sets of data agree. The averages of the two data sets were different but they had a significant overlap. About half of the students (44%) thought that the measurements agree, and 36% chose the explanation "(a) The intervals of the measured values mostly overlap." Equal proportion of students (44%) stated that the measurements do not agree, and 27% chose the explanation "(c) Average values of measurements of both groups are different." The explanation "(f) The measured values are too scattered" was selected by 12% of students.

Item 5 addressed the same concept as item 4 but in this item the two data sets did not overlap. A majority of students (91%) answered that the two measurement results do not agree. The most popular explanation was the correct statement "(a) The intervals of the measured values obtained by groups A and B do not overlap" (50%), followed by "(b) Average values of measurements of both groups are different" (24%) and "(d) The difference of 7 mm between the two averages is small compared to the measured value" (10%).

Items 6 and 7 regarded the distinction between the terms accuracy and precision. In item 6, where the measurement was precise, but not accurate, 49% of students selected the correct answer. Further 29% of students thought that the measurement was accurate, but not precise, whereas 14% chose the answer "(d) The measurement is neither accurate nor precise." In item 7 where the measurement was accurate, but not precise, 67% of students chose the correct answer while 26% selected the answer "(a) The measurement is accurate and precise".

Item 8 referred to correct reporting of the measurement result and correct treatment of significant digits. It was the second best solved item in the test; the average score on it was 70%. The correct answer "(d) $g = (9.80 \pm 0.02)\ \mathrm{m/s^2}$" was selected by 76% of students, followed by 14% who chose "(c) $g = (9.79945 \pm 0.02)\ \mathrm{m/s^2}$." The most popular explanation was the correct statement "(d) The number of digits in the result is determined by the error, which is an essential part of the results" (78%), followed by the explanation "(c) The average should be precisely reported (with more decimal places) along with the corresponding error" (16%).

At the end of the test, students who had the graphical representation of data were asked if it helped them to reply to questions. More than half of the students (59%) answered that the graphical representation of data was helpful. Typical explanations for this opinion included the following: "In some tasks it could be seen that certain values deviate from each other," "Graphical data is clearer than purely numerical data, it is easier to visualize".

Students who did not have the graphical representation of data did not make any attempt to show the data graphically.

IV. STUDY 2: EYE-TRACKING MEASUREMENT

A. Methods

1. Participants

The participants in this study were thirty undergraduate students from the Department of Physics, University of Zagreb. All participants were senior-year (mostly fourth and fifth year) prospective physics teachers who attended at least two physics laboratory courses; some of them attended basic statistics course as well. Each participant

gave an informed written consent before taking part in the experiment.

2. Materials

The test developed for study 1 was also used in study 2 (see Appendix A). We prepared two versions of the test, with and without graphical representation of data.

3. Apparatus

Eye-movement data were recorded using a stationary eye-tracking system with a temporal resolution of 500 Hz and a spatial resolution of 0.25°–0.50° (SMI iView Hi-Speed system, Senso Motoric Instruments G.m.b.H.). The distance between the eyes and the monitor was 50 cm. Prior to every recording, the gaze of each participant was calibrated with a 13-point calibration algorithm. The gaze direction was calculated as a vector between corneal reflection (which is stable, i.e., it depends only on head movements) and pupil position (i.e., the calculated center of the pupil). A fixation can be defined as the state when the eye remains still over a period of time, while a saccade is the rapid motion of the eye from one fixation to another. Smaller eye movements that occur during fixations, such as tremors, drifts, and flicks are called microsaccades. Microsaccades were automatically grouped in a fixation. The fixations were detected automatically using the "event detected method," which is built into the eye-tracking device. Blinks were corrected automatically.

4. Procedure

Before the measurement, participants were familiarized with the apparatus. They were instructed to respond by pressing the enter key on the keyboard and by choosing the answer using the mouse. The participants were asked to keep their head fixed during the measurements, so they could not use paper and pencil. After calibration, questions were presented to participant one by one. By choosing the answer, participant advanced to the next question. If a student had chosen "other" for explanation in two-tiered test items, they gave their oral explanation to the researcher after the eye-tracking measurement. There was no time limit to answer the 8 test items (14 multiple-choice questions). The whole procedure, including preparation, eye-movement calibration, recording, and eventual verbal explanation for two-tiered test items, lasted around 30 min. Half of the participants (15) solved the version of the test with graphical representation of data.

5. Data analysis

The participants' answers were scored in the same way as in the Study 1.

Recorded eye-movements data were analyzed using BeGaze software that calculated eye fixations and saccades. During the fixation the eyes remain relatively still, while during saccades the eyes rapidly change the point of fixation. The sequence of fixations and saccades is called a scan path. Figure 4 shows the scan path of one participant. The software allows the calculation of the viewing time (dwell time) and the number of fixations for any defined area of interest (AOI). We defined five areas of interest for each test item (eye-tracking data for explanations for two-tiered test items were not analyzed). AOIs were rectangles that included introduction text (*Introduction*), measured data (*Data*), multiple-choice question (*Question*), and the graphical representation of data on a number line (*Graphical representation*). AOI *All* includes all these AOIs. The viewing time (dwell time, time spent looking within the area of interest) and the number of fixations were evaluated for each defined AOI. These two measured values indicate the level of cognitive load during processing of particular AOI.

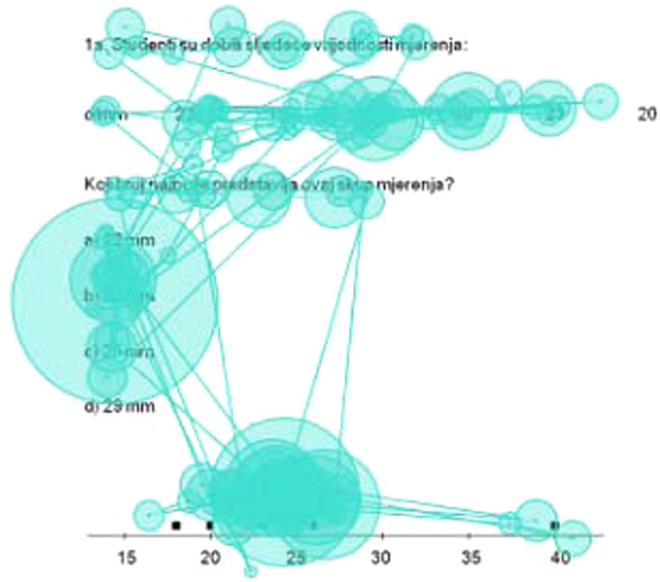


FIG. 4. Example of one participant's scan path. Centers of circles show position of fixations and lines show saccades. Radius of circle illustrates the duration of the fixation.

A two-way mixed-design ANOVA and Bonferroni corrected student's t tests were used to evaluate the difference in viewing patterns between two groups of participants (with and without graphical representation of data). As in study 1, the chi-square test was used for comparing the scores on different test items between groups with and without graphical representation. A threshold of $p = 0.05$ was used for determining the level of effect significance within all conducted tests.

B. Results

1. Analysis of students' scores

The mean score (and standard deviation) on the eye-tracking test was $(51 \pm 20)\%$. The Shapiro-Wilk test showed that the distribution was normal ($W = 0.95$, $p = 0.15$). The largest number of students (12) had test

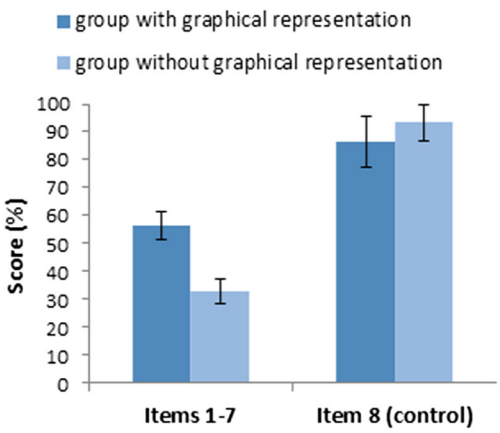


FIG. 5. Students' scores in the eye-tracking study for groups with and without graphical representation of data. Average scores for test items 1–7 are shown. Test item 8 did not have graphical representation of data and it was used to control for the differences between groups. The error bars represent 1 SEM.

score between 41% and 60%. Only three students scored more than 80%, whereas one student scored less than 20%; seven students scored 21%–40% and another seven students scored 61%–80%. The results corroborate previous results from the paper-and-pencil study. Mean scores in both studies were about 50% and distributions of students' scores were normal.

Students who solved the test with graphical representation of data had higher scores on the test items 1–7 than their colleagues who solved the test without graphical representation $[t(28) = 3.52,\ p = 0.001]$. However, the two groups' scores did not differ on test item 8 $[t(28) = 0.59,\ p > 0.05]$. Figure 5 shows students' scores for the two groups (with and without graphical representation) separately for test items 1–7 and for test item 8 (control test item, without graphical representation of data). The results indicate that graphical representation of data helped students in solving the test.

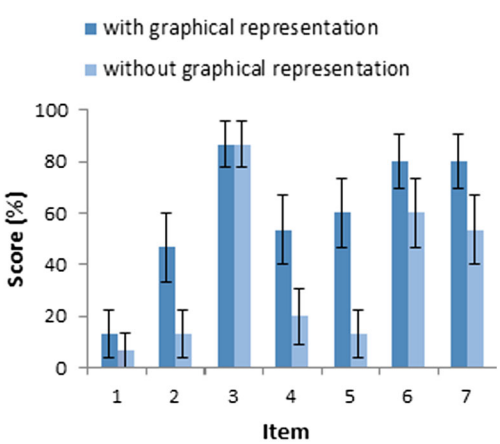


FIG. 6. Scores of all students on seven test items in Study 2 divided into groups with graphical representation of data and without it. The error bars represent 1 SEM.

As in study 1, we analyzed the effect of the graphical representation of data for each test item separately (Fig. 6). The chi-square test showed statistically significant differences in scores between groups with and without graphical representation for item 2 $[\chi^2(1) = 3.968,\ p = 0.046]$ and item 5 $[\chi^2(1) = 7.033,\ p = 0.008]$, whereas no significant differences were revealed with respect to other items. When p values were adjusted for seven comparisons neither difference was statistically significant.

2. Analysis of students' eye movements

To understand how the graphical representation of data helps students in understanding measurements, we analyzed their eye movements. The eye-tracking data gave us an insight to which part of the test items participants paid more attention. We were also able to compare eye-movement patterns of the participants with and without graphical representation of data. For example, Fig. 7 shows comparisons of the average heat maps of the two groups (with and without graphical representation) for the same test item. A very clear difference is observed in the area where the data are shown—students who did not have graphical representation of data spent more time looking at data. Corresponding comparison for the control test item did not reveal any apparent difference between two groups (Fig. 8). In both groups, participants had the longest fixation time at the average value of the measurement (9.799 45), followed by the measurement error (0.02 m/s^2) and offered answers (in particular the correct answer was viewed for a longer time than other multiple choices). The viewing patterns for the groups were rather similar, indicating that the groups were not different as such.

To quantify the observed differences for the test items 1–7, we defined four areas of interest comprising introduction text, measured data, multiple-choice question, and graphical representation of data. Figure 9(b) illustrates defined AOIs for one test item. An overall AOI including all four mentioned AOIs was also defined. We compared the viewing times and the number of fixations for the two groups (with and without graphical representation). These two variables are indirect measures of cognitive load. Longer viewing time and a larger number of fixations indicate higher cognitive load [51–54]. As the results are analogous and yield to the same conclusions, we report here only results for the participants' viewing times. Corresponding data on the total number of fixations for different AOIs are given in Appendix C (Fig. 12).

Figure 9(a) shows that the total viewing time (i.e., viewing time at the AOI All category, calculated as the sum of viewing times in items 1a, 2a, 3a, 4a, 5a, 6, and 7) did not differ for the groups with and without graphical representation $[t(28) = 0.47,\ p > 0.05]$. However, further analysis across smaller AOIs revealed the differences between the groups [Fig. 9(c)]. Bonferroni adjusted t

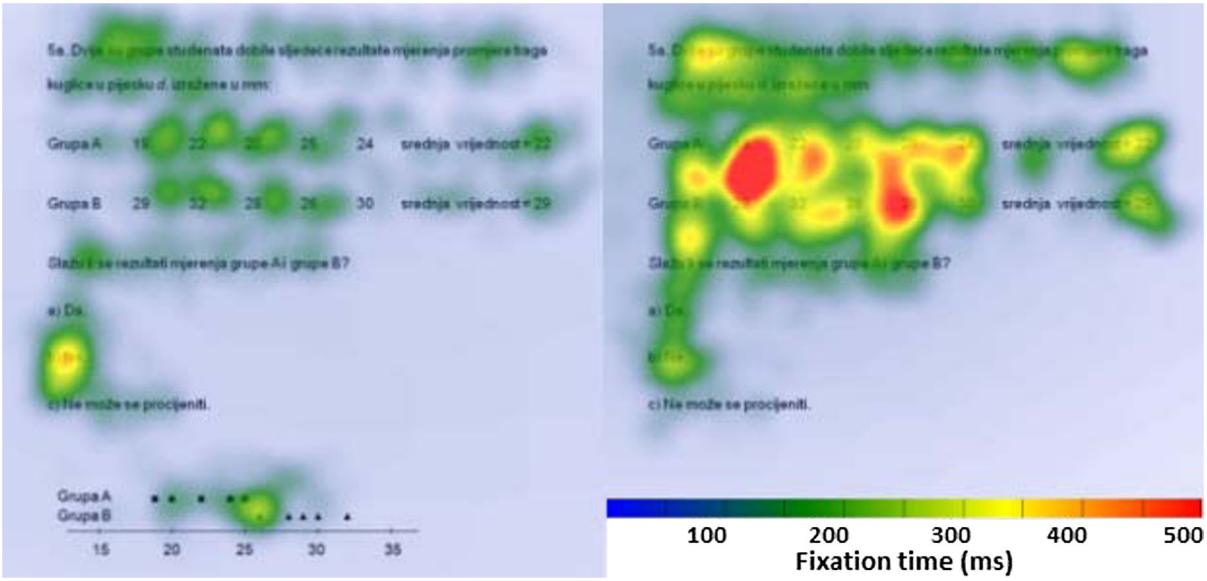


FIG. 7. Heat maps of students who solved the same test item with and without graphical representation of data. Heat maps show how long students looked at different parts of the test item. Red indicates the area of the highest fixation time.

tests showed that students who had graphical representation of data looked less at the AOI Data than their peers without graphical representation [$t(28) = 2.57$, $p = 0.06$]. It is due to the small number of participants and Bonferroni adjustment for four comparisons that this difference was only marginally significant. Both groups looked for the same time at the Introduction text [$t(28) = 0.04$, $p > 0.05$]. Although it may appear that students without graphical representation of data fixated on AOI Question (multiple-choice question) longer, the difference was not statistically significant [$t(28) = 1.34$, $p > 0.05$]. As expected, students who had graphical representations of data looked more at the AOI Graphical representation than their colleagues who only had blank area bellow the Question AOI [$t(28) = 8.27$, $p < 0.001$].

3. Comparisons between paper-and-pencil and eye-tracking results

Our students are not familiar with eye-tracking measurements, and most of the participants in study 2 took part in such a measurement for the first time. We wanted to explore if the new testing environment influenced their outcomes. To assure the comparability of the two samples, we took data only from a subset of participants from study 1 (senior-year students) and all data from study 2 (they were all senior-year students). A two-way mixed-design ANOVA was conducted

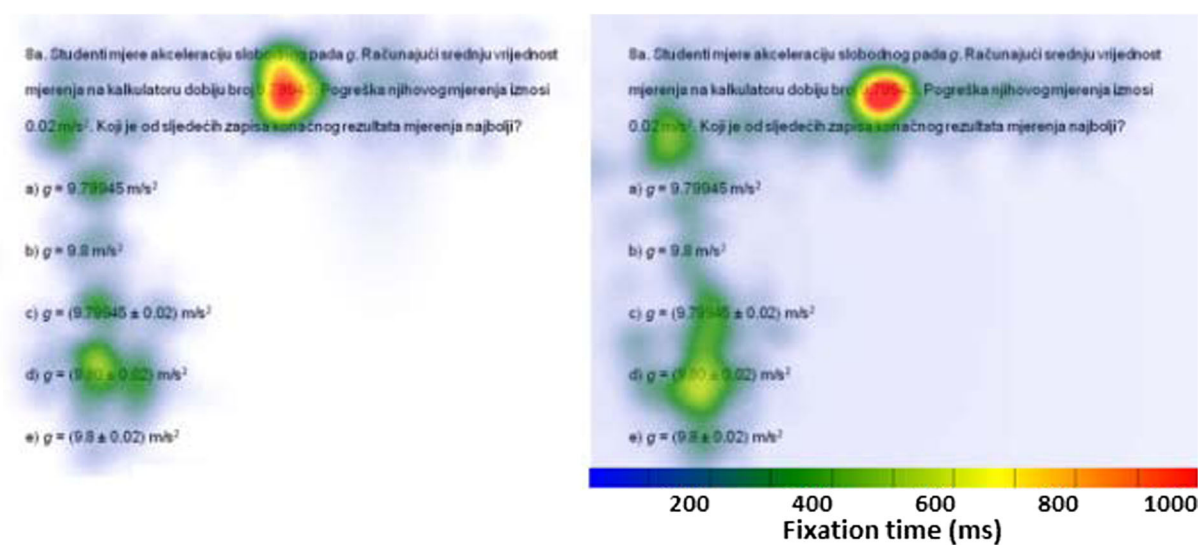


FIG. 8. Heat maps of students from two groups (with and without graphical representation of data) for the control test item 8 without graphical representation. Red indicates the area of the highest fixation time.

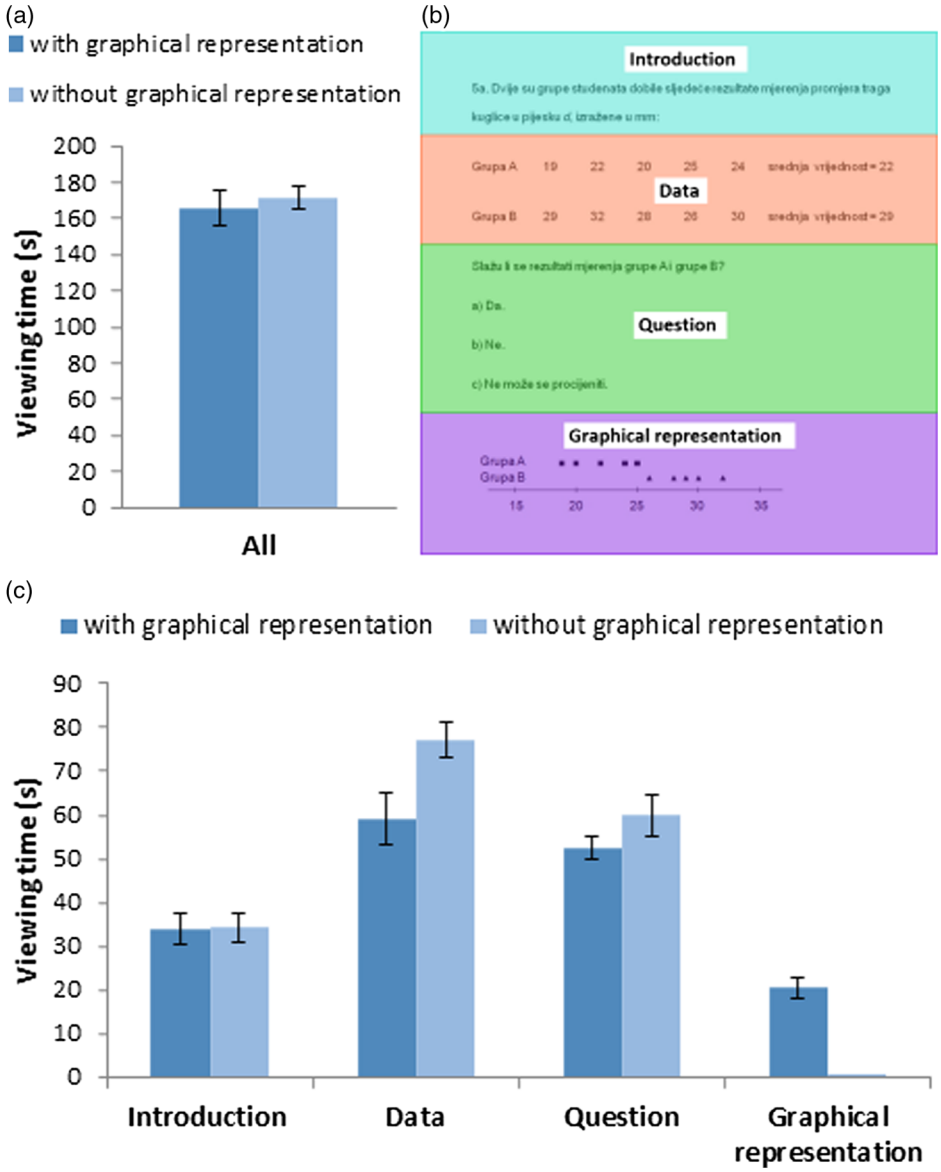


FIG. 9. (a) Average total viewing time at AOI *All* for groups with and without graphical representation of data. (b) Example of definition of AOIs for one test item. (c) Average total viewing times at defined AOIs for groups with and without graphical representation of data. Average total viewing time is calculated for test items 1–7. The error bars represent 1 SEM.

on students' scores with factors graphical representation (with vs without graphical representation) and study (1 vs 2). The obtained results showed a significant main effect of graphical representation [$F(1,79)=10.82$, $p<0.01$, $\eta_p{}^2=0.120$], while the main effect of study [$F(1,79)=1.07$, $p>0.05$, $\eta_p{}^2=0.013$] and interaction effect [$F(1,79)=1.58$, $p>0.05$, $\eta_p{}^2=0.020$] were not significant. Accordingly, we can conclude that the novel testing situation in study 2 did not affect students' test scores (Fig. 10).

V. DISCUSSION

The influence of graphical representation of data on understanding and interpreting measurement results was investigated within two studies among university students. For the purpose of these studies, a novel test was developed that measures student understanding of the measured data and related uncertainty. This test was constructed mostly based on the previous research, and proved to be adequate for our sample. However, it was disappointing that the average score was 49% and that only 3% of students scored more than 80% on the test. The obtained results corroborate previous studies that have shown students' rather poor understanding of physical measurements and the related uncertainty [2–17].

Our main aim was to explore if the graphical representation of data helps students in the interpretation of measured data and the related uncertainty. In both studies, students who

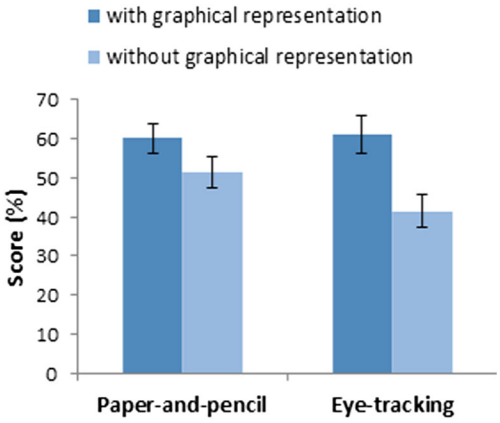


FIG. 10. Students' scores on paper-and-pencil test and eye-tracking test divided into groups with graphical representation of data and without it. Average scores for test items 1–8 (total test scores) are shown. The error bars represent 1 SEM.

had graphical representation of data scored higher than their colleagues without graphical representation. Our experimental groups were rather small (in particular, in study 2), so we used the control item 8 without graphical representation to check the existence of potential differences between groups as such, and the result was negative. The participants' success in solving item 8 in study 2 indicated a ceiling effect, therefore, it was not best suited to control for differences between groups. However, ANOVA results clearly show a beneficial role of graphical representation of data in understanding and interpreting the measured data. Figures 3 and 6 show that the effect of graphical representation of data was not statistically significant for each item 1–7 but the trend was always the same. This suggests that including a larger number of participants in a study would probably yield a statistically significant effect for most test items.

Our studies confirm and extend previous indications that graphical representation might be useful for comparing two data sets [15]. Previous results of a short survey have shown that the majority of a small group of students (8/11) responded that the graphical representation of the overlapping confidence intervals was most helpful when deciding whether their experimental results did or did not agree with a theoretical prediction. However, none of more than 200 students in a larger study ever drew such a graph to help them evaluate whether the two values overlapped [15]. The results have an important implication for teaching—students should be encouraged to graphically show the data. A previous study has reported that some students, when taught to use diagrams such as histograms or linear scatter plots, find it useful and adopt diagrams when reporting data [11].

Furthermore, the obtained results indicated that, as expected, senior-year students scored higher than their first-year colleagues. The improvement is probably due not only to attending physics laboratories but also to taking an introductory statistics course and other physics courses (and possibly also influenced by the dropout of weaker students). However, it was disappointing that the difference between first-year and senior-year students was not larger than was observed in study 1, despite the encouraging results indicating rather good results on test items concerning the quality of the measured data and correct reporting of the measurement result. Nevertheless, more attention should be given to the treatment of outliers, comparisons between measurements, and understanding of the concepts of accuracy and precision.

Our second research question on the influence of the graphical representation of data on participants' eye movements was addressed in study 2 in which we measured eye movements of the students who solved the test constructed for the first study. Students' scores on the test confirmed the results obtained in study 1—students who had graphical representations of data were more successful than their colleagues without graphical representations. Analysis of eye-tracking data showed that the total viewing time, i.e., time spent on solving questions with and without graphical representation, was the same for both groups of participants. Further analysis of smaller AOIs revealed that participants who had graphical representation of data spent less time looking at the AOI Data. This indicates that graphical representation helped them to better understand data, so they did not have to attend this AOI as much as their peers without graphical representation.

Our results can be interpreted in the framework of cognitive load theory [33]. It seems that the graphical representation of data helps visualization by presenting data in a systematic and focused way that reduces working memory load. Consequently, more cognitive resources remain available for further processing of data. Furthermore, graphical representation of data might probably direct attention to the important features of measured data (such as their spread) that are crucial for understanding and comparing different data sets. When data are presented in a numerical form only, a participant needs more cognitive resources to discern and visualize important data features, such as average values and uncertainty intervals. For data presented in a graphical form, it takes less time and effort to see important data features. It seems that the graphical form of data is advantageous not only for its efficiency in using cognitive resources but it can also have an important role as an indicator of important characteristics of data that leads to further data processing. For example, clearly seeing an outlier in a graphical representation of data might remind students to exclude it when calculating the mean. The students themselves concluded that graphical representation of data is useful because of data visualization.

It is important to note here that the obtained results further endorse usefulness of eye-tracking technology in educational studies [36,39–48]. First, the comparison of the students' scores on the test in studies 1 and 2 did not reveal any difference in the students' scores on

paper-and-pencil assessment and eye-tracking measurement. Although students probably felt more uncomfortable in the eye-tracking study because they had to sit still, it did not influence their results. Next, the eye-tracking recording provided an objective measure of visual attention during data processing, giving us an additional insight into the beneficial role of graphical representation of data. Students' reports on the helpfulness of graphical representation also corroborate the eye-tracking findings. A number of students noted that graphical representation of data helped them in visualizing the data. Overall, both studies suggest that graphical representation of data is beneficial.

In addition to the main research questions, within the present study we also compared the distributions of students' responses to previous research results. Analysis of the students' responses revealed that Croatian students have similar difficulties with understanding measurement and related uncertainty as students in other European counties, South Africa and the United States [2–17]. This is in agreement with the PER findings in other fields such as mechanics, electromagnetism etc., indicating that students express similar difficulties worldwide, e.g., Refs. [55,56].

With regard to specific test items, students in our study had most difficulties with item 1 where they were asked which number represents a set of measurements the best. A majority of the students realized that they should calculate the average of the measured values, but only a small number recognized the outlier value and discarded it. Our results agree with the results of one study on U.S. students where only less than 10% of participants omitted the outlier and averaged other measurements [15]. However, 35% of Japanese students in the same study omitted one outlier before calculating the mean and an additional 50% of participants omitted two data points because they were taught the "trimmed mean" procedure in statistics class (omitting the highest and the lowest value). About half of the students in South African study [5] and 60%–90% of the U.S. students in another study [17] excluded the outlier before calculating the mean. One of the reasons for such a discrepancy in the results probably lies in the form of the questions. In the latter studies [5,17], the students were explicitly confronted with the issue of excluding the outlier, while in our study [15] the students were supposed to recognize an outlier and then decide to exclude it. Furthermore, these results confirm that the instruction on the nature of measurement and error analysis has a crucial role in developing student understanding of measurements. Overall, students should be explicitly trained to recognize the outlier values in their measurements, to think about possible reasons for such deviations, and to decide whether they should be excluded from the data analysis.

In item 2 the majority of students identified measurement uncertainty. However, it was more difficult to evaluate the uncertainty and understand the cause of uncertainty [13]. About three quarters of students knew that it can be concluded that the measured quantity is somewhere in the interval of measured values, but only half of them agreed with the statement that we can never know the true value of the measured quantity.

In three test items, students were asked to compare the results of two data sets. Most students correctly solved item 3, where they were supposed to decide on the quality of the measurement results, based on the data spread. It was more difficult to decide if the results of the two data sets agree by looking at the intervals overlap (items 4 and 5). Again, the results corroborate the previous reports on students' difficulties on data comparison [5,7–11,15–17]. Croatian students are not often (if at all) confronted with the need to compare data sets in the physics laboratory, and more attention should be paid to such activities. As the previous study has found, it takes time and effort to develop basic skills related to data manipulation and comparison [13].

The distinction between the terms accuracy and precision represents a more technical skill but it makes a part of general understanding of measurements, so it was tested by items 6 and 7. About half of the students in our sample were able to differentiate between accuracy and precision, whereas the rest were confounding the two concepts. Our results are comparable with those from a previous study [15]. In a typical target shooting question students separately reported about accuracy and precision (in our study combined answers about accuracy and precision were offered). For example, on the question equivalent to our item 7 (accurate, but not precise measurement), 21% students answered that the accuracy was good and majority (95%) agreed that the precision was poor. We believe that even short interventions including discussion with examples such as target shooting [15] would substantially improve student understanding of the terms accuracy and precision.

Reporting the measurement result and treatment of significant digits (item 8) were among the easier skills for our tested sample. This is probably the result of considerable emphasis on these issues in the physics laboratory at the Department of Physics, University of Zagreb. Croatian students seemed to be more successful in reporting the measurement result than the American students in the previous studies [15,17]. More than 85% of students in one study [15] reported too many significant figures, while in another study [17], the results were considerably better with 70% of students reporting the correct number of significant figures. However, the format of the questions in the previous U.S. study and our study was different; hence, the results are not really comparable. The U.S. students solved open-ended test questions where they were asked to report the average and uncertainty estimate of a set of measurement data. Despite correct reporting of measurement results, it has been shown that students often apply rules of significant figures without a firm conceptual understanding of the reasons behind it [15].

To our knowledge, this study is the first systematic PER study on graphical representation of measurement data performed on a larger number of students. Our results indicate that graphical representation of data may help students to better understand data and measurement uncertainty. Also, for the first time, we have used the eye-tracking measurements to get a more detailed picture of the allocation of students' attention when solving questions related to measurements and uncertainty. The results from the eye-tracking study confirmed the positive effect of graphical representation on data processing found in the paper-and-pencil study. The eye-tracking results suggest that students pay attention to graphical representation of data and thus spend less time processing numerical data. This finding is in accordance with cognitive load theory; presentation of data in graphical form reduces the cognitive load and increases the amount of cognitive resources available to process the data. Overall, the results suggest that more emphasis should be placed on graphical representation of measurements data in physics laboratory courses.

Finally, it is also important to consider several factors that may limit the generalization of the obtained results. First, in the present study we had a rather small number of participants, especially in study 2, so the effect of the graphical representation of data on individual test items should be further explored on a larger number of participants in future studies. Besides, a larger number of participants in the eye-tracking study would allow the investigation of differences in eye movements of correct and incorrect problem solvers. In future studies, additional eye-tracking measures, such as fixation duration, might be used to further assess cognitive load related to processing data in a numerical and graphical format. The present experimental design cannot give a full picture of when or why graphical representations help in student understanding of measurement, but the item analyses can help make predictions. In future studies, it would also be interesting to compare scores and eye-movement patterns of students presented with graphs, but with no numerical data and students presented with both graphs and numerical data. The role of the position of the graphical representation of data within the stimuli (before or after the numerical data) should also be examined in future studies. It would also be useful to add more control items in future studies because students had high scores on the control item used in the present studies, thus indicating a possible ceiling effect that should be avoided.

VI. CONCLUSION

Previous PER studies suggested that graphical representation of the data might be helpful in understanding measurement and data processing, but this issue had not been systematically studied. We used paper-and-pencil assessment and measurement of eye movements to investigate the role of graphical representation in understanding and interpreting data. The results showed that graphical representation helped students to better understand and interpret data. The students who had graphical representation spent less time on the area of interest including data, and scored higher than their colleagues without graphical representation. The results suggest that students should be taught to graphically represent measurement data.

Croatian students showed similar difficulties with understanding measurement and related uncertainty as students from other countries reported in the previous studies. The lowest scores were found in the test items concerning the treatment of an outlier and comparison of the results of two data sets. Study 1 showed that senior-year students scored higher than first-year students. This indicates that the current teaching practice appears to have some positive effect on student understanding of measurements. However, we hope that the results of this and other PER studies will call for further efforts to improve university students' skills related to measurements and data processing.

ACKNOWLEDGMENTS

This research was funded by the University of Split, Croatia.

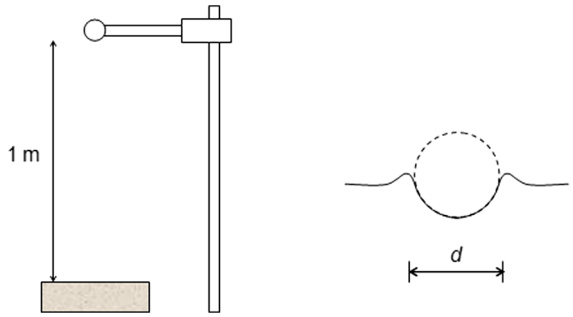
APPENDIX A: TEST ITEMS[1]

Name:_____

Instruction: Please answer the questions in order and do not return to those that you have already answered. In some test items an explanation will be required after you choose an answer in the multiple-choice question. Please choose the answer that best corresponds to the reason why you have selected the answer on the previous question. If among the offered answers you cannot find one that fits your reasoning, please write down the explanation in your own words.

Students carefully release the ball without the initial velocity from a height of 1 m and measure the diameter d which the ball leaves in the sand.

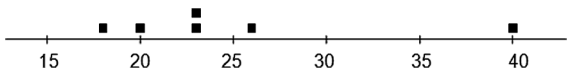
[1]Each question was presented on one page in the booklet.



1a. Students obtained the following measurement results:

d/mm	23	18	26	40	23	20

Which number best represents this set of measurements?
a) 22 mm
b) 23 mm
c) 25 mm
d) 29 mm

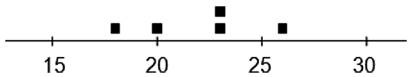


1b. Explanation:
a) This number is obtained if all measurements are summed and divided by 6.
b) This number is obtained if measurement 40 mm is ignored, then remaining measurements summed and divided by 5.
c) This number appeared twice in the measurements, whereas the others appeared only once.
d) This number is in the middle of the measurement results.
e) Other explanation: _____
_____

2a. Students obtained the following measurement results for the diameter d which the ball left in the sand:

d/mm	23	18	26	23	20

What can students conclude about the value of the measured quantity d?
a) The measured quantity is 22 mm.
b) The measured quantity is 23 mm.
c) The measured quantity is somewhere between 18 and 23 mm.
d) The measured quantity is somewhere between 18 and 26 mm.

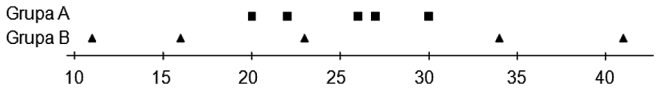


2b. Explanation:
a) This number is obtained if all measurements are summed and divided by 5.
b) Measurement 26 mm deviate from the mean value, so it should be ignored.
c) This number appeared twice in the measurements, whereas the others appeared only once.
d) We can never know the true value of the measured quantity.
e) Other explanation:_____
_____

3a. Two groups of students obtained the following measurement results for the diameter of the ball trace in the sand d, expressed in mm:

Group A	26	30	22	27	20	average = 25
Group B	23	11	41	16	34	average = 25

Compare the obtained results.
a) The results of group A are better than the results of group B.
b) The results of group B are better than the results of group A.
c) The results of both groups are equally good.
d) It cannot be estimated which results are better.



3b Explanation:
a) Both groups obtained the same average value.
b) The results of group A are between 20 mm and 30 mm, and the results of group B are between 11 mm and 41 mm.
c) The results of groups A and B do not agree—it is possible that the same average values are obtained by chance.
d) Groups A and B used different methods, but that does not matter if the final result is the same.
e) If multiple measurements are taken, the same average value is obtained, regardless of the error in the measurement.
f) Other explanation: _____
_____

4a Two groups of students obtained the following measurement results for the diameter of the ball trace in the sand d, expressed in mm:

Group A	18	25	20	28	24	average = 23
Group B	20	28	22	24	31	average = 25

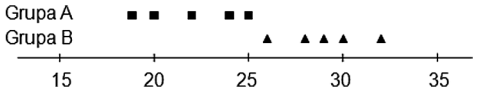
Do the measurement results of group A and group B agree?
a) Yes.
b) No.
c) It cannot be estimated.

4b Explanation:
a) The intervals of the measured values mostly overlap.
b) Both groups obtained measurement values 20, 24 and 28 mm.
c) Average values of measurements of both groups are different.
d) The interval of the measured values was 10 mm wide for group A, and 11 mm wide for group B.
e) The difference of 2 mm between the two averages is large compared to the measured value.
f) The measured values are too scattered.
g) Other explanation:_____
_____

5a Two groups of students obtained the following measurement results for the diameter of the ball trace in the sand d, expressed in mm:

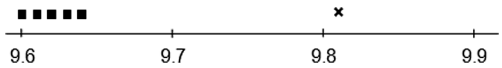
Group A	19	22	20	25	24	average = 22
Group B	29	32	28	26	30	average = 29

Do the measurement results of group A and group B agree?
a) Yes.
b) No.
c) It cannot be estimated.

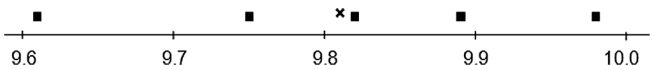


5b Explanation:
a) The intervals of the measured values obtained by groups A and B do not overlap.
b) Average values of measurements of both groups are different.
c) The interval of the measured values was 6 mm wide for both groups.
d) The difference of 7 mm between the two averages is small compared to the measured value.
e) The measured values are too scattered.
f) Other explanation:_____
_____

6. Students measured the free-fall acceleration g (in Zagreb g equals 9.81 m/s^2). They obtained the following measurement results for g, expressed in m/s^2:
9.63 9.64 9.62 9.60 9.61 average = 9.62
How would you describe the accuracy and precision of the measurement?
a) The measurement is accurate and precise.
b) The measurement is accurate, but not precise.
c) The measurement is precise, but not accurate.
d) The measurement is neither accurate nor precise.



7. Students measured the free-fall acceleration g (in Zagreb g equals 9.81 m/s^2). They obtained the following measurement results for g, expressed in m/s^2:
9.61 9.98 9.82 9.75 9.89 average = 9.81
How would you describe the accuracy and precision of the measurement?
a) The measurement is accurate and precise.
b) The measurement is accurate, but not precise.
c) The measurement is precise, but not accurate.
d) The measurement is neither accurate nor precise.



8a. Students measure the free-fall acceleration g. Calculation of the average on a calculator gives the number 9.79945. Error of the measurement is 0.02 m/s^2. Which of the following reports of the final measurement result is the best?
a) $g = 9.79945\ m/s^2$
b) $g = 9.8\ m/s^2$
c) $g = (9.79945 \pm 0.02)\ m/s^2$
d) $g = (9.80 \pm 0.02)\ m/s^2$
e) $g = (9.8 \pm 0.02)\ m/s^2$

8b. Explanation:
a) As many digits as possible should be written in the result to be more precise. The error depends on the number of measurements, so there is no sense in reporting it.
b) The result must always be rounded to one decimal place. The error depends on the number of measurements, so there is no sense in reporting it.
c) The average should be precisely reported (to more decimal places) along with the corresponding error.
d) The number of digits in the result is determined by the error, which is an essential part of the results.
e) Other explanation:_____
_____

APPENDIX B: DISTRIBUTION OF STUDENTS' RESPONSES

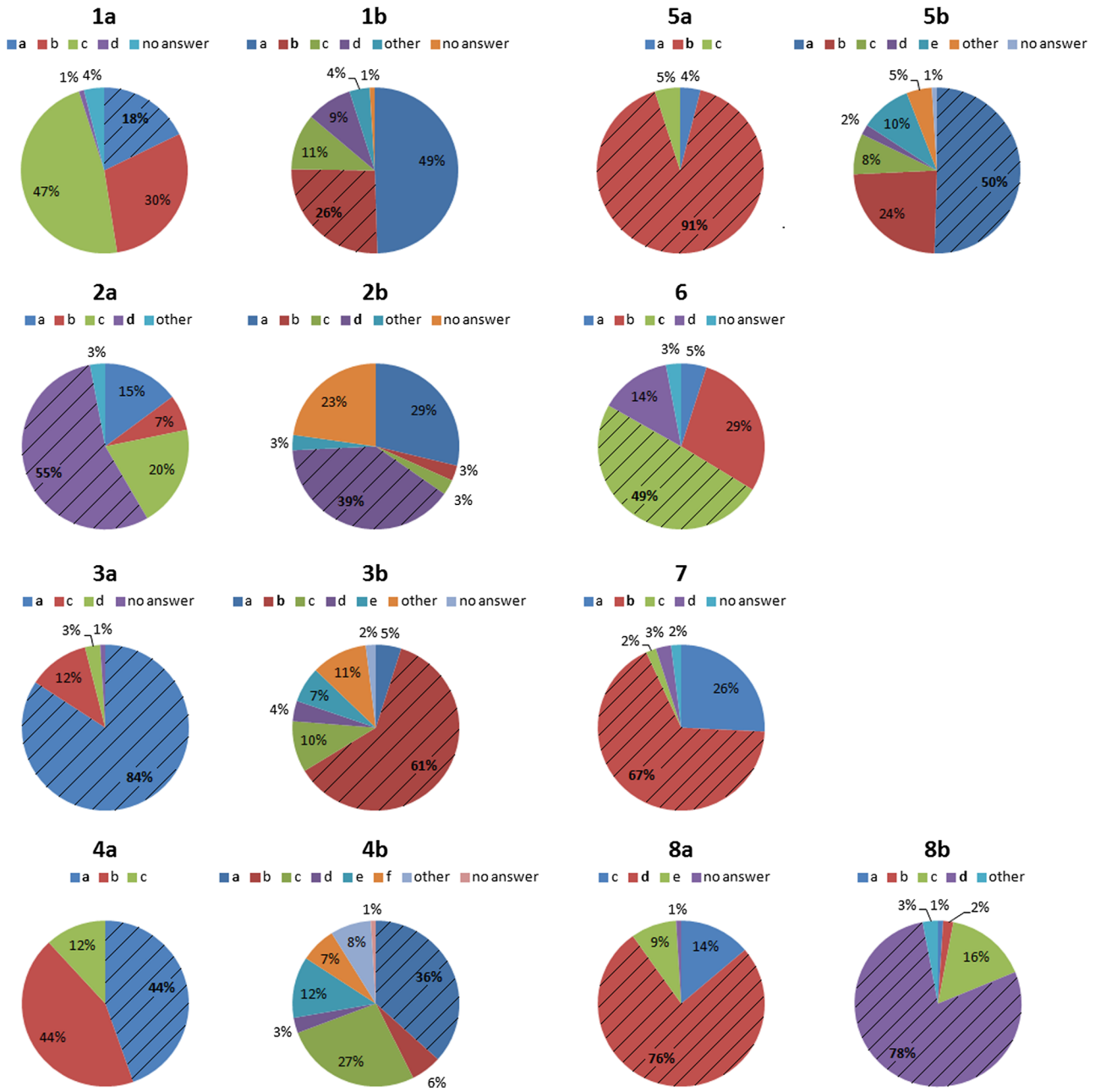


FIG. 11. Pie diagrams of distribution of students' responses for all questions and all participants in study 1. The correct answers are given in bold and denoted by stripes in pie diagrams.

APPENDIX C: TOTAL NUMBER OF FIXATIONS FOR DIFFERENT AOIs

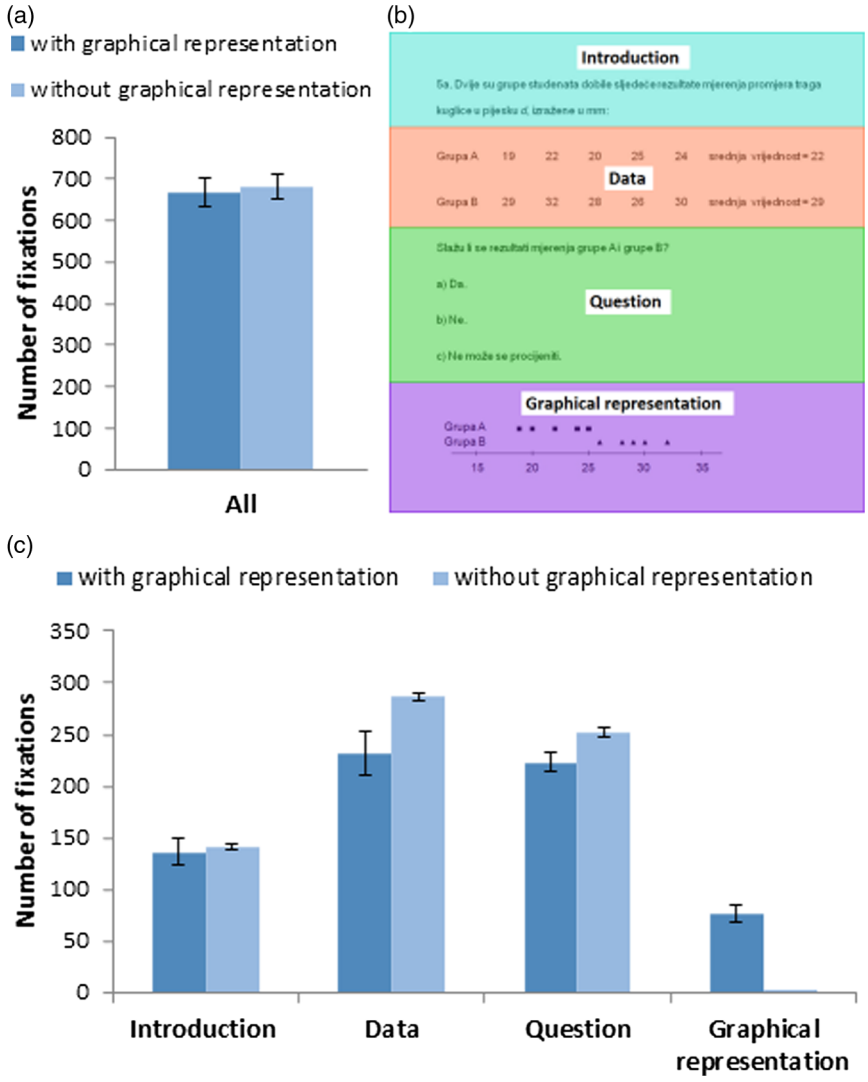


FIG. 12. (a) Average total number of fixations (calculated as the sum of number of fixations for items 1a, 2a, 3a, 4a, 5a, 6, and 7) at AOI *All* for groups with and without graphical representation of data. (b) Example of definition of AOIs for one test item. (c) Average total number of fixations at defined AOIs for groups with and without graphical representation of data. The error bars represent 1 SEM.

[1] American Association of Physics Teachers, Goals of the introductory physics laboratory, Am. J. Phys. **66**, 483 (1998).

[2] M.-G. Séré, R. Journaux, and C. Larcher, Learning the statistical analysis of measurement errors, Int. J. Sci. Educ. **15**, 427 (1993).

[3] M.-G. Séré, F. Fernandez-Gonzalez, J. Gallegos, F. Gonzales-Garcia, E. De Manuel, J. Perales, and J. Leach, Images of science linked to labwork: A survey of secondary school and university students, Res. Sci. Educ. **31**, 499 (2001).

[4] J. Leach, R. Millar, J. Ryder, M.-G. Séré, D. Hammelev, H. Niedderer, and V. Tselfes, Students' images of science as they relate to labwork learning. Labwork in Science Education, Working Paper 4, *Targeted Socio-Economic Research Programme Project PL 95–2005* (1998).

[5] S. Allie, A. Buffler, L. Kaunda, B. Campbell, and F. Lubben, First-year physics students' perceptions of the quality of experimental measurements, Int. J. Sci. Educ. **20**, 447 (1998).
[6] B. Campbell, L. Kaunda, S. Allie, A. Buffler, and F. Lubben, The communication of laboratory investigations by university entrants, J. Res. Sci. Teach. **37**, 839 (2000).
[7] F. Lubben, A. Buffler, S. Allie, and B. Campbell, Point and set reasoning in practical science measurement by entering university freshmen, Sci. Educ. **85**, 311 (2001).
[8] A. Buffler, S. Allie, F. Lubben, and B. Campbell, The development of first year physics students' ideas about measurement in terms of point and set paradigms, Int. J. Sci. Educ. **23**, 1137 (2001).
[9] T. S. Volkwyn, S. Allie, A. Buffler, and F. Lubben, Impact of a conventional introductory laboratory course on the understanding of measurement, Phys. Rev. ST Phys. Educ. Res. **4**, 010108 (2008).
[10] R. L. Kung, Teaching the concepts of measurement: An example of a concept-based laboratory course, Am. J. Phys. **73**, 771 (2005).
[11] R. Lippmann Kung and C. Linder, University students' ideas about data processing and data comparison in a physics laboratory course, NorDiNa **2**, 40 (2006).
[12] E. Etkina, S. Murthy, and X. Zou, Using introductory labs to engage students in experimental design, Am. J. Phys. **74**, 979 (2006).
[13] E. Etkina, A. Karelina, and M. Ruibal-Villasenor, How long does it take? A study of student acquisition of scientific abilities, Phys. Rev. ST Phys. Educ. Res. **4**, 020108 (2008).
[14] J. Day and D. Bonn, Development of the Concise Data Processing Assessment, Phys. Rev. ST Phys. Educ. Res. **7**, 010114 (2011).
[15] D. L. Deardorff, Ph.D. thesis, North Carolina State University, 2001.
[16] R. F. Lippmann, Ph.D. thesis, University of Maryland, 2003.
[17] D. S. Abbott, Ph.D. thesis, North Carolina State University, 2003.
[18] http://journals.aps.org/prper/supplemental/10.1103/PhysRevSTPER.4.010108/e010108_app.pdf.
[19] http://user.physics.unc.edu/~deardorf/uncertainty/quiz.html.
[20] Joint Committee for Guides in Metrology, Evaluation of measurement data—Guide to the expression of uncertainty in measurement, GUM 1995 with minor corrections (2008).
[21] S. Allie, A. Buffler, B. Campbell, F. Lubben, D. Evangelinos, D. Psillos, and O. Valassiades, Teaching measurement in the introductory physics laboratory, Phys. Teach. **41**, 394 (2003).
[22] A. Buffler, S. Allie, and F. Lubben, Teaching measurement and uncertainty the GUM way, Phys. Teach. **46**, 539 (2008).
[23] S. Pillay, A. Buffler, F. Lubben, and S. Allie, Effectiveness of a GUM-Compliant Course for Teaching Measurement in the Introductory Physics Laboratory, Eur. J. Phys. **29**, 647 (2008).
[24] L. C. McDermott, M. L. Rosenquist, and E. H. van Zee, Student difficulties in connecting graphs and physics: Examples from kinematics, Am. J. Phys. **55**, 503 (1987).
[25] R. J. Beichner, Testing student interpretations of kinematics graphs, Am. J. Phys. **62**, 750 (1994).
[26] W. M. Christensen and J. R. Thompson, Investigating graphical representations of slope and derivative without a physics context, Phys. Rev. ST Phys. Educ. Res. **8**, 023101 (2012).
[27] M. Planinic, Z. Milin-Sipus, H. Katic, A. Susac, and L. Ivanjek, Comparison of student understanding of line graph slope in physics and mathematics, Int. J. Sci. Math. Educ. **10**, 1393 (2012).
[28] M. Planinic, L. Ivanjek, A. Susac, and Zeljka Milin-Sipus, Comparison of university students' understanding of graphs in different contexts, Phys. Rev. ST Phys. Educ. Res. **9**, 020103 (2013).
[29] L. Ivanjek, A. Susac, M. Planinic, A. Andrasevic, and Zeljka Milin-Sipus, Student reasoning about graphs in different contexts, Phys. Rev. Phys. Educ. Res. **12**, 010106 (2016).
[30] D. Meltzer, Relation between students' problem-solving performance and representational mode, Am. J. Phys. **73**, 463 (2005).
[31] P. B. Kohl and N. D. Finkelstein, Student representational competence and self-assessment when solving physics problems, Phys. Rev. ST Phys. Educ. Res. **1**, 010104 (2005).
[32] M. De Cock, Representation use, and strategy choice in physics problem solving, Phys. Rev. ST Phys. Educ. Res. **8**, 020117 (2012).
[33] P. Chandler and J. Sweller, Cognitive load theory and the format of instruction, Cognit. Instr. **8**, 293 (1991).
[34] S. M. Coelho and M.-G. Séré, Pupils' reasoning in practice during hands-on activities in the measurement phase, Res. Sci. Technol. Educ. **16**, 79 (1998).
[35] A. H. Duc, P. Bays, and M. Husain, Eye movements as a probe of attention, Prog. Brain Res. **171**, 403 (2008).
[36] A. Susac, A. Bubic, J. Kaponja, M. Planinic, and M. Palmovic, Eye movements reveal students' strategies in simple equation solving, Int. J. Sci. Math. Educ. **12**, 555 (2014).
[37] K. Nitschke, N. Ruh, S. Kappler, C. Stahl, and C. P. Kaller, Dissociable stages of problem solving (I): temporal characteristics revealed by eye-movement analyses, Brain Cogn. **80**, 160 (2012).
[38] E. R. Grant and M. J. Spivey, Eye movements and problem solving: Guiding attention guides thought, Psychol. Sci. **14**, 462 (2003).
[39] R. Merkley and D. Ansari, Using eye tracking to study numerical cognition: The case of the ratio effect, Exp. Brain Res. **206**, 455 (2010).
[40] E. Schneider, M. Maruyama, S. Dehaene, and M. Sigman, Eye gaze reveals a fast, parallel extraction of the syntax of arithmetic formulas, Cognition **125**, 475 (2012).
[41] M. Kozhevnikov, M. A. Motes, and M. Hegarty, Spatial visualization in physics problem solving, Cogn. Sci. **31**, 549 (2007).
[42] A. D. Smith, J. P. Mestre, and B. H. Ross, Eye gaze patterns as students study worked out examples in mechanics, Phys. Rev. ST Phys. Educ. Res. **6**, 020118 (2010).
[43] A. M. Madsen, A. M. Larson, L. C. Loschky, and N. S. Rebello, Differences in visual attention between those who

correctly and incorrectly answer physics problems, Phys. Rev. ST Phys. Educ. Res. **8**, 010122 (2012).

[44] A. Madsen, A. Rouinfar, A. M. Larson, L. C. Loschky, and N. S. Rebello, Can short duration visual cues influence students' reasoning and eye movements in physics problems?, Phys. Rev. ST Phys. Educ. Res. **9**, 020104 (2013).

[45] A. Rouinfar, E. Agra, A. M. Larson, N. S. Rebello, and L. C. Loschky, Linking attentional processes and conceptual problem solving: Visual cues facilitate the automaticity of extracting relevant information from diagrams, Front. Psychol. **5**, 1094 (2014).

[46] H. Tang and N. Pienta, Eye-tracking study of complexity in gas law problems, J. Chem. Educ. **89**, 988 (2012).

[47] M. Cook, E. N. Wiebe, and G. Carter, The influence of prior knowledge on viewing and interpreting graphics with macroscopic and molecular representations, Sci. Educ. **92**, 848 (2008).

[48] M. Patrick, G. Carter, and E. Wiebe, Visual representations of DNA replication: Middle grades students' perceptions and interpretations, J. Sci. Educ. Technol. **14**, 353 (2005).

[49] R. H. Tai, F. J. Loehr, and F. J. Brigham, An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments, Int. J. Res. Meth. Educ. **29**, 185 (2006).

[50] M. J. Tsai, H. T. Hou, M. L. Lai, W. Y. Liu, and F. Y. Yang, Visual attention for solving multiple-choice science problem: An eye-tracking analysis, Comput. Educ. **58**, 375 (2012).

[51] Q. Wang, S. Yang, M. Liu, Z. Cao, and Q. Ma, An eye tracking study of website complexity from cognitive load perspective, Decision Support Systems **62**, 1 (2014).

[52] K. F. Van Orden, W. Limbert, S. Makeig, and T. P. Jung, Eye activity correlates of workload during a visuospatial memory task, Human Factors **43**, 111 (2001).

[53] M. De Rivecourt, M. N. Kuperus, W. J. Post, and L. J. M. Mulder, Cardiovascular, and eye activity measures as indices for momentary changes in mental effort during simulated flight, Ergonomics **51**, 1295 (2008).

[54] K. Rayner, K. H. Chase, T. J. Slattery, and J. Ashby, Eye movements as reflections of comprehension processes in reading, Sci. Stud. Read. **10**, 241 (2006).

[55] L. C. McDermott and E. F. Redish, Resource Letter: PER-1: Physics Education Research, Am. J. Phys. **67**, 755 (1999).

[56] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020119 (2014).