Gaze patterns enhance response prediction: More than correct or incorrect

Sebastian Becker[*]

University of Cologne, Faculty of Mathematics and Natural Sciences—Digital Education Research, 50931 Cologne, Germany

Stefan Küchemann[†]

University of Kaiserslautern, Department of Physics - Physics Education Research, 67663 Kaiserslautern, Germany

Pascal Klein

University of Göttingen, Faculty of Physics, Physics Education Research, 37077 Göttingen, Germany

Andreas Lichtenberger

ETH Zürich, Department of Physics - Solid-State Dynamics and Education, 8093 Zürich, Switzerland

Jochen Kuhn

LMU Munich, Faculty of Physics, Physics Education, 80333 München, Germany

(Received 11 February 2022; accepted 16 June 2022; published 26 July 2022)

Eye tracking enables the reconstruction of eye movements and thus the analysis of visual information selection and integration processes during problem solving. In this way, learner-specific difficulties can be identified and problem-solving process can be adapted accordingly. For such an adaptation, the prediction of response behavior plays a crucial role. To predict whether a problem is solved correctly or incorrectly, the segmentation of the visual stimulus into specific areas of interest (AOIs) is particularly crucial for the quality of a prediction based on eye-tracking data. In the study presented here, the gaze data of $N = 115$ students were analyzed while solving the Test of Understanding Graphs in Kinematics (TUG-K), a validated test instrument whose items include graphs of position, velocity, and acceleration versus time. For selected items, response accuracy was predicted based on visual attention using multiple logistic regression analysis, examining the influence of AOI segmentation. The prediction quality could be significantly improved when the diagram was not considered as contiguous AOI, but when it was divided into solution-relevant and solution-irrelevant areas. To verify that the AOIs selected by the regression algorithm are indeed relevant to the solution process, an expert rating was performed, which showed moderate to good agreement between the AOIs rated by the experts as relevant to the correct solution and the AOIs selected by the algorithm. There are also pairs of items in the TUG-K that require the same mathematical solution procedure but differ in the physical context. This opened the possibility to investigate a new approach. Based on response accuracy and allocation of visual attention to one item, the response accuracy of the other item of the pair was predicted. It could be shown that the prediction quality based on visual attention was significantly higher than the prediction based on response accuracy. This demonstrates the added value of collecting process-based data versus product-based data for prediction and thus for learner-specific adaptation. The results of this study indicate, first, that only certain areas are crucial for a correct solution when extracting information from diagrams and, second, that the application of mathematical procedures plays a crucial role in interpreting graphs of different physics quantities. These findings thus provide insight into the visual strategies involved in interpreting kinematic diagrams and can also serve as a basis for eye tracking-based adaptation of problem-solving processes, in which adaptation can occur even before an incorrect answer is given.

DOI: 10.1103/PhysRevPhysEducRes.18.020107

[*] Corresponding author.
sbeckerg@uni-koeln.de

[†] These authors contributed equally to this work.

I. INTRODUCTION

Capturing eye movement via eye tracking has become increasingly important for physics education research [1,2]. The recording of visual attention during learning or problem solving allows insights into cognitive processes that remain hidden in strictly product-based data collection methods [3]. This is particularly valuable when the processes can provide information about problem-solving strategies or student difficulties [4–7], or if the meaning of individual visual stimuli is the special focus of research interest [8,9]. The relationship between visual attention and performance is a frequently studied issue in eye tracking in physics education research [10]. If students eye-tracking data are used to predict performance, the visual information extraction processes can be linked to students' understanding [11], and as an extension, the prediction can be used to provide scaffolding support for students in an adaptive learning environment [12].

The analysis of eye-tracking data is especially beneficial when students learn or solve problems with multiple external representations [9,13] or synthesis problems [4], as the conceptual understanding is closely linked to the visual information extraction processes. Among external representations, graphs are a central visualization for learning to simplify complex concepts and to display measurement data across the science, technology, engineering, and mathematics (STEM) fields [14,15].

In the present study, we investigate how eye-tracking data can be used to predict response behavior regarding selected items of the Test of Understanding Graphs in Kinematics (TUG-K) [16], which has become a widely used test instrument to evaluate students' understanding of graphs. The items were created based on extensive research on students' difficulties and address several tasks like determining velocity, acceleration, displacement, and change in velocity from graphs of position, velocity, and acceleration versus time.

For a prediction based on eye-tracking data, the test items must be segmented into specific areas [so-called areas of interest; (AOIs)], for which the eye-tracking metrics are calculated. This segmentation can be done at different levels of granularity, e.g., globally only text and diagram or blockwise for each step of thought [17]. However, no study has yet systematically investigated the influence of granularity on prediction performance in this context, which is surprising since high predictive quality is essential for adapting test and learning formats. To fill this research gap, this study investigates which granularity of segmentation can achieve the highest accuracy in predicting students' response behavior.

To solve items of the TUG-K, students have to evaluate and interpret the slope of a curve or the area under a curve in graphs. Previous studies have shown that these mathematical procedures, rather than physics concepts, dominate when students solve problems on kinematic graphs [18,19]. For example, when students determine the acceleration from a time-velocity diagram, they apply the same mathematical procedure as when they calculate the velocity from a time-position diagram, which is assumed to be reflected in similar gaze behavior.

We used such item pairs of the TUG-K to investigate whether the same underlying mathematical procedure allows a prediction of the response behavior for one item by solely analyzing the gaze behavior for the other item. Such analyses might open up the possibilities to detect inaccurate solving strategies of students and to offer support based on these difficulties even before the other item of the pair was viewed. In an adaptive learning environment, demotivating failure in solving such items could be avoided in this way, and a learning process induced instead.

From a methodological perspective, we can draw conclusions about the relationships between eye-tracking data and performance that are relevant for adaptive gaze-based learning systems, for example. From an educational perspective, we learn how students extract relevant information from kinematic diagrams and identify the underlying mathematical solution strategies in different physical contexts.

II. STATE OF RESEARCH

A. Eye tracking in physics education research

The basic question of whether something as mundane as the low-level perceptual functions involved in visual attention can also provide information about performance on physical tasks was already posed by Rouinfar *et al.* [10], since physics problems "are among the most intellectually and cognitively demanding [processes] that human beings are capable of engaging in". The relationship between visual attention and performance in solving physics tasks is of particular research interest (see Ref. [2] for an overview, particularly Table IV). It was repeatedly shown that highly aggregated measures of visual attention, such as dwell time on questions, diagrams, or alternatives (in multiple choice assessment scenarios), were not significantly related to performance [20–23].

However, if one chooses a finer analysis, for example, by looking at attention to the correct response option (as opposed to aggregating time across all alternatives), one finds correlations with response accuracy [21,23,24], which is the basis for prediction. This is not surprising in the context of multiple-choice tasks, since the selected option is considered longer than unselected options.

On the other hand, if we look at relevant and irrelevant areas, for example, in diagrams, it has been shown that successful learners spend more time on relevant areas than on irrelevant areas, and for unsuccessful learners, it is the opposite [5,6,10,25–27].

This can be explained by the information reduction hypothesis of Gegenfurtner *et al.* [28], which states that experts are more efficient in identifying relevant areas in a representation and paying more visual attention to them. Consequently, it can be assumed that if a task is solved correctly, areas of the visual stimulus that are relevant (not relevant) to the solution will receive higher (lower) visual attention. In studies by Madsen *et al.* [25] and Rouinfar *et al.* [10], these relevant areas were identified in preliminary work through interviews with students.

The issue of granularity in the analysis of eye-tracking data has been raised in several physics education studies. Chien *et al.* [29] speak of global look zones when they divided their learning material into two sections, and Smith *et al.* [17] decompose their worked-out examples into text and equations at different levels of granularity. In the studies by Klein *et al.* [9] and Küchemann *et al.* [30], it was found that visual attention to concept-specific AOIs within graphs distinguishes correctly from incorrectly responding students as well as confident and unconfident students.

In the context of graphs, Klein *et al.* [9] observed that the cumulative visual attention to the entire graph area does not discriminate between high- and low-performing students which suggests that the cumulative visual attention on this AOI is not predictive for students' performance. Additionally, the authors found in an exploratory analysis that students who solve slope problems correctly attribute more visual attention on areas along the curve in a graph [9]. Specifically, for the TUG-K, it was found that the dwell time on the options is correlated to the students' self-confidence ratings and that the underlying test structure of the TUG-K is reflected in the students' gaze transitions between the question, the graph, and the options [23,31]. Thus, to infer performance from gaze data, a tailored selection of AOIs as well as the associated metrics is required.

For various representations, the predictability of student comprehension based on eye movements has already been investigated based on different approaches to defining AOIs [11,32]. Rebello *et al.* [32] found that the dwell time on patterns of 192 squarelike tiles that cover the entire area of representation-based items allows for a high prediction of student performance. In comparison, Küchemann *et al.* [11] found that the dwell time on relevant and irrelevant areas of a line graph reached a lower prediction accuracy of students' performance. Here, the authors also observed that the dwell time along the line graph is predictive for students' performance when determining the slope and that the dwell time on the area underneath contains information about students' understanding when determining the integral of a graph.

Based on these works, it remains an important open question to what extent the selection of AOIs in which the eye movements are evaluated affects the predictability.

B. Line graphs in STEM education

Line graphs are typically introduced in middle school and used throughout education, transferring to a variety of STEM contexts [33]. Independent of the context, the same mathematical procedures, such as determining the slope or area enclosed by a graph, are applied to evaluate graphs in different contexts and to communicate scientific concepts [34]. Both differentiating and integrating follow strict procedures in which specific regions in the graph must be evaluated. The mathematics concepts of slope and integral have been in the scope of research over the past years [e.g., [35–39]], and they were picked up by other assessment instruments such as the Kinematics Concept Test (KCT) [18], The Force and Motion Conceptual Evaluation (FMCE) [40], and the Kinematics Representational Competence Inventory (KiRC) [41]. Research has shown that mathematical procedures are crucial for the understanding of graphs in several contexts.

The studies from Lichtenberger *et al.* [18] as well as Bektasli and White [19] provided evidence that the mathematics concepts of slope and area are central in solving tasks about graphs in the field of kinematics. Conducting a structural analysis of student answers to kinematic graphs, Lichtenberger *et al.* [18] found that the graphical determination of slopes and areas were the main separable dimensions. Only in a finer-grained analysis was it possible to distinguish between acceleration and velocity as a rate and between displacement as an area under the curve in a v, t graph and change of velocity as area under the curve in an a, t graph. In other words, students who are able to correctly answer questions about velocity are also likely to correctly answer mathematically similar questions about acceleration. Given this result, the authors concluded that the understanding of the mathematical procedures is crucial for the interpretation of kinematics graphs. Evaluating a TUG-K dataset, Bektasli and White [19] also found two factors explaining the correlations of student answers, one for determining the slope of a curve and the other one for finding or interpreting the area under a curve. Physics concepts like velocity and acceleration were embedded in this structure and did not separate.

Such underlying mathematical structures should also be reflected in gaze behavior, which elicits the question of whether corresponding eye-tracking metrics can be used to predict response behavior when solving corresponding items of the TUG-K.

III. RESEARCH QUESTIONS

Based on the current state of research, the way in which the visual stimulus is segmented into AOIs is crucial for predicting response behavior based on gaze data. Since there are different approaches for such segmentation, we wanted to investigate the influence of the segmentation procedure on the prediction quality.

We focused on a comparison between a macro-level AOI segmentation (question text, diagram, response options) and a microlevel segmentation in which the diagram is systematically divided into smaller segments according to the information reduction hypothesis of Gegenfurtner *et al.* [28]. A finer segmentation into solution-relevant and solution-irrelevant segments should contribute to a better discrimination of visual attention between correct and incorrect responders and thus to a better prediction of response accuracy. Thus, the first research question relates to the prediction of response accuracy for individual selected items based on visual attention allocation with respect to different AOI segmentations.

RQ 1: When using eye-tracking data to predict the response accuracy, which AOI segmentation of the test items results in the best prediction quality?

The TUG-K has pairs of items that require the same mathematical procedure to solve but differ in the physical context. For example, in two items the negative slope must be determined quantitatively, once as velocity and once as acceleration. The studies presented above have shown that students who are able to apply the correct mathematical procedure are often successful in both contexts. In turn, the application of the mathematical procedure requires the selection of certain visual information from the given diagrams, which should be reflected in the eye-tracking data. We therefore assume that students who have successfully applied the procedure in one context will also do so in another context and thus show similar eye-tracking behavior. The question arises as to whether the inclusion of eye-tracking measures for one item of a pair can be used to improve the quality of the prediction of the answer accuracy of the other item. This approach is new and has not yet been investigated.

RQ 2: For a pair of items involving different physical contexts but the same mathematical solution procedure, can the prediction of the response accuracy of an item be improved based on the eye-tracking data of the corresponding item?

IV. METHODOLOGY

A. Sample

The data were collected from $N = 115$ German and Swiss gymnasium students (58 female, 57 male; all with normal or correct-to-normal vision). A gymnasium is a public school that provides higher education and constitutes the highest level of the educational system in Germany and Switzerland. Gymnasium students are comparable to U.S. high school students attending college preparatory classes.

B. Materials

The 26 single-choice items of the latest version of the TUG-K (version 4.0[1]) were translated into German and presented to the students in the original order of the test in two sets of 13 items, with a short break in between (an example item is shown in Fig. 1). The TUG-K contains the following three item pairs, which require the same mathematical procedure for the quantitative determination of the physical quantity and to which the analysis in this article is therefore limited. The corresponding items can be found in the Supplemental Material [42].

1. Determination of the positive slope
item 5 Velocity from position-time graph
item 7 Acceleration from velocity-time graph

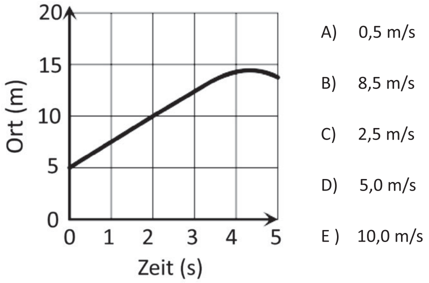

FIG. 1. Example item from the TUG-K. The task has been translated verbatim into German and reads as follows in the English original: *The following figure shows the position versus time graph of an object. The velocity of the object at* $t = 2$ s *is*:. The axis labels are also written in German and mean *time* (“Zeit”) and *position* (“Ort”).

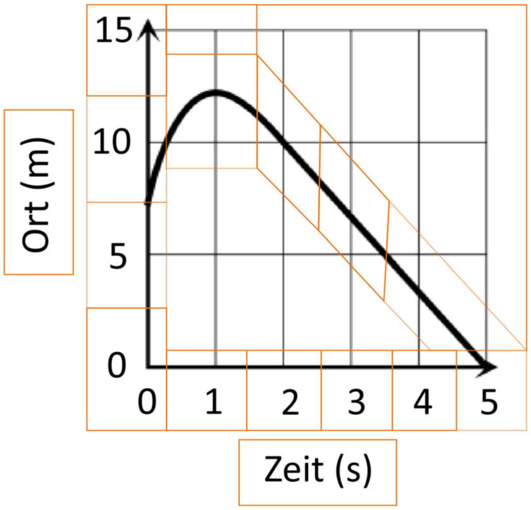

FIG. 2. Segmentation of the diagram of an exemplary item into different micro-AOIs. The axis labels are written in German and mean *time* (Zeit) and *position* (Ort).

2. Determination of the negative slope
item 18 Velocity from position-time graph
item 6 Acceleration from velocity-time graph
3. Determination of the area content
item 4 Displacement from velocity-time graph
item 16 Change in velocity from acceleration-time graph

C. Procedure and measures

1. Test procedure

Four identical eye-tracking systems were set up in school libraries and the students participated voluntarily in data collection, either in free periods or regular classes (with permission of their teachers). At the time of the testing procedure, the subject area kinematics were completed in all courses. The participants received no credit or gift for participating. First, the students were introduced to the eye tracker, and a nine-point calibration process was used for a fully customized and accurate gaze point calculation prior

[1] physport.org/assessments.

to each set of 13 questions. Subsequently, the items were presented on a 22-inch computer screen (1920 × 1080; refresh rate 75 Hz) equipped with a Tobii X3–120 remote eye-tracking system. The students then solved the items without interruption from the researcher and had as much time as needed to solve. After students were ready to give their answers, they pressed a key to move to the next page where they entered their answer. They did not receive feedback after completing an item and could not return to previous tasks.

2. AOI segmentation procedure

Based on the raw gaze points, which have been recorded at 120 Hz, we determined fixations based on an I-VT algorithm. The I-VT algorithm determines the speed between each gaze point, and all consecutive gaze points that are below a threshold of 30°/s belong to the same fixation. Once this threshold is exceeded, the fixation ends, and a new fixation starts when the speed is again below this threshold. The connection between two consecutive fixations is termed *saccade*, which is an rapid eye movement that does not exceed a duration of 100 ms. To extract the eye-tracking metric "total visit duration" [(TVD), cumulated times between first fixation in and first fixation outside an AOI; AOIs not visited are counted as 0] as a measure for visual attention allocation, the items presented to the students were segmented into AOIs on two different levels. At the macroscopic level, the question text, the diagram and the options were assigned an AOI (macro-AOIs).

At the microscopic level, the diagram AOIs were systematically positioned according to the information reduction hypothesis of Gegenfurtner *et al.* [28] described in the state of research. This means that two physics education experts identified the information in the graphs that was relevant and irrelevant to solving the tasks. In the graph tasks here, there is typically a large irrelevant area, such as a nonlinear part of the graph and a relevant area of the graph, in which the slope needs to be determined. In the first step, we discriminated between these two areas. Similarly, there are irrelevant areas on the axes (e.g., in Fig. 2 from 0 to 1 on the x axis and the value 15 on the y axis), which were also considered separate micro-AOIs.

Furthermore, we also considered the work by Klein *et al.* [9], who found that students who correctly determine the slope of the graph spent more time on regions located along the graph. Accordingly, we isolated the graph from the remaining area. Eventually, we isolated the value that is mentioned in the question text and the associated point along the graph (in Fig. 2: the value 3 on the x axis and the associated point on the graph). This value is a surface feature that is likely to attract the attention of both students who solve an item correctly and students who solve it incorrectly. We also isolated the values on the x and y axes required to solve the item correctly.

To determine the slope, it is necessary to divide the y-axis interval with the x-axis interval, and to determine the area underneath the graph, the students are required to multiply the x-axis interval with the y-axis interval and calculate the half of this area. This area determination is the simplified calculation of the integral underneath the graph for the case where the graph runs linearly through the origin. Therefore, we considered the end points of each interval as separate micro-AOIs (in Fig. 2 the values 2 and 5 on the x axis and the values 0 and 10 on the y axis).

3. Prediction of response accuracy

Multiple regression analysis was performed for different datasets to predict response accuracy based on the eye-tracking metric TVD. In the resulting models, the standardized TVD on the specific AOIs and their interactions were included as possible predictors of the dependent variable response accuracy. For each dataset, a stepwise regression based on Akaike's information criterion (AIC) was carried out using the step function from the R package "stats" (v.3.6.2) to find the best-fitting model by iteratively removing the least contributive predictor from the full model. Because response accuracy is a binary coded variable (0 indicates an incorrect answer, 1 indicates a correct answer), logistic regression was chosen (see, e.g., Ref. [43]). The resulting logistic models were compared using the function compareGLM from the R package "rcompanion" (v. 2.3.26). In this way, for each AOI segmentation, one obtains those AOIs that are relevant for predicting a correct or incorrect solution. The models can also be compared to find the one best suited for the prediction.

To compare the prediction models based on the different datasets, McFadden's pseudo R-squared (R^2_{McFadden}) was obtained as a measure of prediction quality and tested whether the corresponding prediction models differed significantly from each other in the explained proportion of variability. If the models differed significantly, the amount of deviance reduced (denoted as "Red. Dev.") by the better-fitting model was calculated.

V. RESULTS

First, with reference to RQ 1, the results on the influence of the two different levels of AOI segmentation (macroscopic and microscopic) on prediction quality for individual items are presented below. We call this intra-item prediction, because here the data regarding one item are used to predict the correctness of the same item (Sec. A). Moreover, in the process, we also tested whether adding the correct response option as an AOI leads to further improvement in prediction quality.

We then report the results for the inter-item prediction. Here, we use the data with respect to one item to predict the correctness of another item (Sec. B). In addition to the gaze data, we can also use the correctness of the base item here.

A. Intra-item prediction

1. Comparison of the datasets

The three datasets used to estimate response accuracy for each item are defined as follows:

TABLE I. Results of the intra-item prediction: McFadden's pseudo R squared (R^2_{McFadden}), reduced deviance compared with the previous model (Red. Dev.) and p value (p).

Dataset	R^2_{McFadden}	Red. Dev.	p
Item 4			
1	0.066	⋯	⋯
2	0.256	28.8	2.4×10^{-06}
3	0.396	21.3	2.8×10^{-04}
Item 5			
1	0.069	⋯	⋯
2	0.255	29.4	6.6×10^{-06}
3	0.440	29.3	2.1×10^{-05}
Item 6			
1	0.000	⋯	⋯
2	0.322	49.0	6.4×10^{-08}
3	0.679	54.3	6.5×10^{-10}
Item 7			
1	0.224	⋯	⋯
2	0.260	5.7	0.678
3	0.329	10.9	9.8×10^{-04}
Item 16			
1	0.044	⋯	⋯
2	0.129	12.8	3.5×10^{-04}
3	0.234	15.8	1.3×10^{-03}
Item 18			
1	0.000	⋯	⋯
2	0.179	28.1	3.4×10^{-05}
3	0.441	41.2	1.1×10^{-09}

- Dataset 1: We use TVD on the macro-AOIs (i.e., question text, diagram, and options). The macro-AOIs were defined *a priori* due to the surface features of the items.
- Dataset 2: We use TVD on the micro-AOIs resulting from a finer segmentation of the diagram.
- Dataset 3: We use TVD on the micro-AOIs of the diagram (as in dataset 2) and on the correct option.

By definition, the information level increases from dataset 1 to 3, and the degree of aggregation decreases. The results of the model comparison for the three datasets are shown in Table I for each of the examined items. For a clear representation, the pseudo-R^2 values are shown in Fig. 3.

In the graphical comparison of the pseudo R-square measure, it can be observed that for each item, the R-square measure increases from dataset 1 to 3. Thus, the increase in prediction quality with AOI segmentation at the microlevel (datasets 2 and 3) compared to segmentation at the macro level (dataset 1) can be identified pretty clearly. The limit for a sufficiently good fit of the model to the data, which is usually set at a value for McFadden's pseudo R^2 of 0.2, is reached for segmentation on the macro level (dataset 1) only for one item and otherwise remains far below. A prediction based on a finer segmentation of the diagram (dataset 2), on the other hand, leads to a good prediction quality being achieved for all items except item 7. If the TVD regarding the correct answer option is included in the model (dataset 3), the goodness of fit is increased for all items and thus the limit for a sufficiently good prediction quality is exceeded for each item.

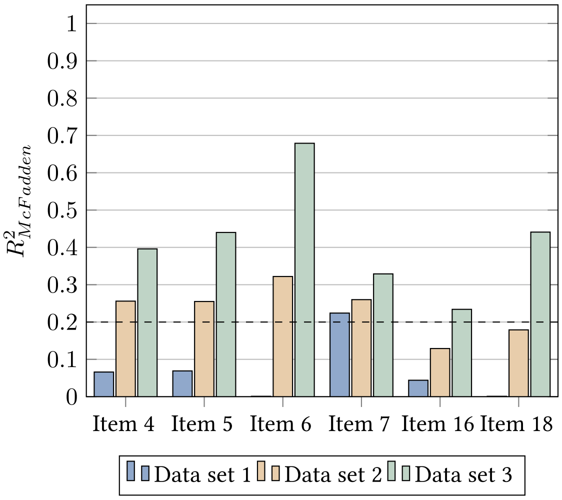

FIG. 3. Comparison of McFadden's pseudo R^2 for intra-item prediction based on the different datasets. The limit from which good prediction quality can be assumed is shown as a dashed line.

2. Relevant and irrelevant micro-AOIs for prediction

Based on the micro-AOIs selected by the algorithm (dataset 2), we examined whether the gaze data in each micro-AOI is helpful to discriminate between visual strategies that are associated with correct and incorrect solutions (i.e., we studied which of these task-relevant and irrelevant

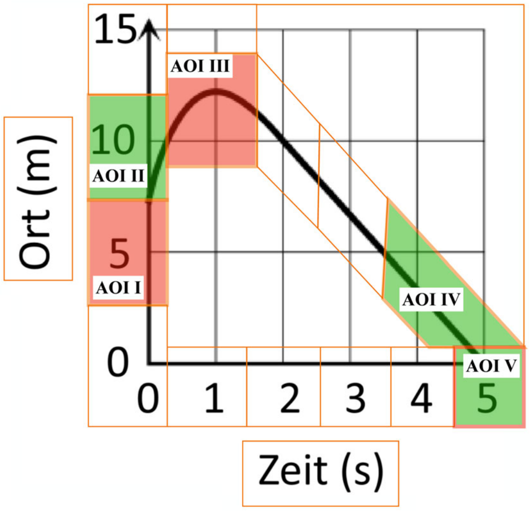

FIG. 4. Predictive AOIs of item 18 (red: longer TVD with incorrect answer, green: longer TVD with correct answer). The axis labels are written in German and mean *time* (Zeit) and *position* (Ort).

micro-AOIs are prediction relevant). The results for item 18 are shown in Fig. 4; the results for the other items are attached to the Supplemental Material [42]. The AOIs highlighted in red were viewed longer if the item was solved incorrectly, while the AOIs marked in green were viewed longer if the item was solved correctly. The remaining AOIs (with no highlights) were not relevant for distinguishing gaze behavior in correct and incorrect responses.

Since the regression procedure selects AOIs relevant for prediction based on eye-tracking data alone, an expert rating with eight experts from physics education research was conducted to verify whether the AOIs selected by the algorithm are indeed relevant for the solution process.

3. Expert rating

We present the results for item 18 in Table II, and the results for all other items can be found in the Supplemental Material [42]. The AOI numbers (column 1) refer to the labels in Fig. 4, and in the second column the direction of the correlation between longer TVD and answer accuracy is given. In the last columns, the number of experts who judge the AOI as relevant or not relevant is shown. For example, AOI III (corresponding to the area around the highest point in the graph) was judged as irrelevant by all experts, and a longer TVD on this AOI was associated with incorrect answers. Thus, there is a consistency between the data and the expert rating in this case. In the following section, we quantify this consistency by determining the agreement coefficient κ between the expert rating and the algorithmic classification of the AOIs via

$$\kappa = \frac{1}{N}\sum_{i=1}^{N}\frac{p_{0,i} - p_c}{1 - p_c}. \tag{1}$$

In Eq. (1), $N = 8$ is the number of raters, $p_{0,i} = (N_{i,\mathrm{rel}} + N_{i,\mathrm{irrel}})/(N_{\mathrm{tot,rel}} + N_{i,\mathrm{irrel}})$, where $N_{i,\mathrm{rel}}$ is the number of AOIs that have been rated as task relevant by the ith rater and that have received a significantly higher TVD by students who solved the item correctly, $N_{i,\mathrm{irrel}}$ is the number of AOIs that have been rated as task irrelevant by the ith rater and that have received a significantly higher TVD by students who solved the item incorrectly, $N_{\mathrm{tot,rel}}$ is the total number of AOIs that have received a significantly higher TVD by students who solved the item correctly, $N_{\mathrm{tot,irrel}}$ is the total number of AOIs that have received a significantly higher TVD by students who solved the item incorrectly, and p_c is the probability of having an agreement between TVD and expert rating by chance.

TABLE II. Expert rating regarding the solution relevance for AOIs of item 18.

AOI	Response for longer TVD	Relevant	Irrelevant
I	incorrect	3	5
II	correct	8	0
III	incorrect	0	8
IV	correct	7	1
V	correct	7	1

TABLE III. Agreement coefficient κ between the expert rating and the algorithmic classification of the AOIs.

Item 4	Item 5	Item 6	Item 7	Item 16	Item 18	Mean
0.31	0.01	0.77	0.50	0.88	0.73	0.53

Table III shows the agreement coefficient for all items. It is noticeable that the expert ratings and the classification shows a total moderate agreement of 0.53, ranging from 0.01 (item 5), which means that there is a slight agreement, to 0.88 (item 16), which implies almost perfect agreement (according to Landis and Koch [44]).

B. Inter-item prediction

In the following, the results of the prediction for item pairs of the TUG-K are presented, which require the same mathematical solution procedure, but associated with a different physical context. Predicted is the response accuracy for one item of the pair based on data (response accuracy, eye-tracking data) of the other item of the pair (inter-item prediction). By comparing different datasets, we address RQ2 (i.e., whether the prediction quality can be improved by adding eye-tracking data for such a prediction). Finally, we added the TVD with respect to the correct response option as AOI to the model to supposedly increase the prediction quality.

The quality of inter-item prediction was compared for the following datasets.

- Dataset 1: We use only accuracy data (correct or incorrect) of the base item.
- Dataset 2: We use TVD based on the micro-AOIs.
- Dataset 3: We combine the information from above (i.e., we use accuracy and TVD based on the micro-AOIs).
- Dataset 4: We additionally use TVD on the correct option. Thus, we use accuracy and TVD based on the micro-AOIs and TVD on the correct option.

While datasets 1 and 2 are clearly separated (product data vs process data), datasets 3 and 4 contain information from the previous datasets and thus are richer in information. The characteristic parameters of the corresponding models are shown in Table IV and graphically illustrated in Fig. 5.

For the first pair of items (4 and 16), the prediction quality increases steadily from dataset 1 to dataset 4 (i.e., the AIC value decreases and McFadden's pseudo R^2 increases [cf. Table IV]). By adding the TVD, the deviance is reduced, and the prediction model becomes significantly better. The highest prediction quality is achieved when the accuracy, the TVD regarding the micro-AOIs, and the

TABLE IV. Results from inter-item prediction: McFadden's pseudo R squared (R^2_{McFadden}), reduced deviance compared with the previous model (Red. Dev.) and p value (p).

Dataset	R^2_{McFadden}	Red. Dev.	p
Prediction from item 4 to item 16			
1	0.158	⋯	⋯
2	0.246	14.9	5.0×10^{-03}
3	0.416	24.3	$< 2.2 \times 10^{-16}$
4	0.522	16.1	3.1×10^{-04}
Prediction from item 5 to item 7			
1	0.156	⋯	⋯
2	0.213	9.0	2.5×10^{-01}
3	0.321	17.0	3.9×10^{-05}
4	0.321	⋯	⋯
Prediction from item 18 to item 6			
1	0.056	⋯	⋯
2	0.343	39.4	2.2×10^{-05}
3	0.354	5.9	$< 2.2 \times 10^{-16}$
4	0.354	⋯	⋯

correct option are used for the prediction. Including TVD on micro-AOIs also improves the prediction performance for the second (items 5 and 7) and third (items 18 and 6) pairs of items considered in our analysis. However, in these cases, the additional inclusion of the TVD regarding the correct option does not lead to a significantly better prediction model.

It is striking that the limit for a sufficiently good prediction quality (McFadden's pseudo $R^2 > 0.2$) is not reached if only the answer correctness (i.e., the product-based measure) is used as a data basis, which clearly demonstrates the added value of including process-based eye-tracking data to predict response behavior for intra-item prediction.

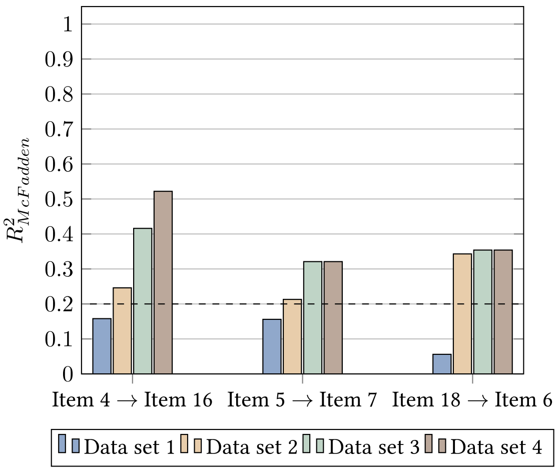

FIG. 5. Comparison of McFadden's pseudo R^2 for inter-item prediction based on the different datasets.

VI. DISCUSSION

A. Intra-item prediction

The aim of our analysis was to identify which AOI segmentation is best suited for predicting the correctness of items with graphs. In all items, it is noticeable that the prediction improves from the macro level to the micro level. The results also show that for only one of six items did the prediction of the correct answer using the TVD on macro-level AOIs reach a value of $R^2 > 0.2$, which corresponds to a good prediction. For a segmentation on the micro level, on the contrary, the prediction of the correct answer reaches a value $R^2 > 0.2$ in four out of six items.

It should be noted here that we only use a finer segmentation of the diagram area as AOIs that is guided and confirmed by experts; the amount of eye-tracking data is identical. Therefore, it is reasonable to assume that this finer division allows for a better distinction between the strategies of students with correct and incorrect answers. This implies that the graph exhibits an essential role in the problem-solving process and that the inclusion of eye-tracking data does not lead *per se* to the best prediction of the answer correctness.

A further increase in prediction quality can be achieved if the correct response option is included as an additional AOI. Inclusion of the TVD with respect to this AOI in the prediction model leads to an increase in predictive validity for all items. This is in agreement with previous studies that have already demonstrated the predictivity of the correct answer option in multiple choice tests for answer correctness [21,23].

To provide relevance and meaning to the AOIs, we included expert ratings of the relevance of the AOIs for the solution process. We found moderate agreement which implies a relation between the statistical difference in the predictability to the inherent structure of the tasks. For example, only slight agreement was found in item 5. In this case, the AOI at an x value of $t = 2$ s is highly relevant for the solution process (eight of eight experts consider this AOI as relevant) to determine the slope, but it received a higher TVD by students with an incorrect answer. The reason for this might be that the value of $t = 2$ was already mentioned in the question text. This means that students do not need to understand the concept of the slope to focus on this AOI, but rather follow the cue in the question text. Apart from this low value, there is one item that shows an almost perfect agreement (item 16) and two items with a substantial agreement (item 6 and item 18). These results imply that the areas in a graph that are relevant for the solution receive more attention from students with a correct answer.

B. Inter-item prediction

The basis for inter-item prediction is item pairs of TUG-K, which require the same mathematical procedure to solve, but differ in terms of the physical quantity that must be determined. The intention of inter-item prediction for these item pairs is to predict the response patterns of one item using data with respect to the other item. By comparing the underlying logistic regression models, it was demonstrated for all three item pairs that the model based on eye-tracking data is significantly better at predicting response patterns than the model based only on product-based response accuracy.

The pseudo *R*-square measure for the prediction based on the eye-tracking dataset is above the threshold of 0.2 for all three pairs of items, in contrast to the prediction based on response accuracy, which indicates a particularly good fit of the model in. Although the number of possible predictors is larger with the eye-tracking dataset compared to the single predictor variable response accuracy, the model comparison nevertheless proves the added value of the process-based eye-tracking data in predicting response accuracy.

Thus, the gaze behavior during the solution process of an item is of greater importance for the response behavior of the other item of an item pair than the knowledge about the correct or incorrect answer after completion of the solution process alone. This indicates that problem-solving strategies are manifested in the gaze behavior, which is used again in the same or similar way for an item that requires the same mathematical procedure. For example, if the velocity or acceleration is to be determined from a given linear graph, this task can be solved in both cases by applying a slope triangle, which is reflected in the gaze data. This supports the study results of Bektasli and White [19] and also Lichtenberger *et al.* [18] who both conclude that understanding the mathematical procedures is essential in interpreting kinematic graphs.

The prediction quality can be further improved by combining eye-tracking data and answer correctness. For all three items, the model based on the combined dataset is also significantly better. However, adding the TVD for the correct response option leads to a significantly better-fitting model only for item pair 4 and 5. For the other two item pairs, the TVD for the correct response option is not a significant predictor.

VII. CONCLUSION AND OUTLOOK

Our paper focuses on the investigation of the prediction of response accuracy based on visual attention for selected items from the TUG-K, which is a well-established test instrument for understanding kinematic diagrams in educational research. We were able to show that segmenting the diagram AOIs on the micro level leads to an increase in prediction quality, which could even be increased by adding the correct answer option as an additional AOI. The segmentation was based on a partitioning of the diagram into areas that are relevant or not relevant for the solution process and which, according to Gegenfurtner's information reduction hypothesis, should receive more or less visual attention if the solution is correct.

This indicates that a certain visual strategy underlies a correct interpretation of kinematic diagrams. Successful students pay more attention to relevant areas of the diagram in order to extract information for the problem-solving process. We infer that deficits occur at the beginning of the problem-solving process, in that students do not correctly identify the relevant information from the diagram. In physics instruction, this should be taken into account by teaching the students in which areas of a diagram they can find which information and how they can extract it from the diagram, which can be generalized to diagrams in different contexts.

Considering the data analysis process, the results clearly show the dependence of the prediction quality on the specification of the AOIs before the actual regression analysis. In order to achieve a good prediction quality based on eye-tracking data, a segmentation of the diagram is essential for the items investigated here, which divides the diagram area into relevant and nonrelevant areas for the solution. It should be noted that this requires knowledge about solution processes of such diagrams (i.e., expertise in subject didactics), that should be considered, especially with regard to an automated analysis of eye-tracking data.

In this context, a promising research perspective arises from the possibility of having the AOI segmentation performed by an objective system rather than on the basis of a subjective expert assessment. In this case, an algorithm could (automatically) identify AOIs on the basis of the collected gaze data, which could contribute to an increase in the quality of prediction.

The TUG-K has item pairs whose items require the same mathematical solution procedure but have different physical target quantities. This allowed us to investigate the prediction of answer correctness of one item based on data on the other item of the pair. Compared to product-based answer correctness, which can only be determined after the solution process, a higher prediction quality could be achieved using eye-tracking data, which is already determined during the solution process.

Such a process-based prediction of response behavior opens up the possibility of automatically supporting the problem-solving process even before the task is solved incorrectly, for example, by providing additional explanations or visual cues, which would avoid a demotivating failure in problem solving for the learner. In addition, gaze behavior could be used to identify erroneous strategies and offer (automated) support measures tailored to the task. This could be implemented in intelligent tutoring systems to use student gaze data to continuously adapt learning environments and thus personalize the learning process. However, this would require real-time processing of eye movement data that includes fixation filtering and determination of eye movement metrics in predefined AOIs.

Future research must therefore address the question of whether and at what point in the problem-solving process eye-tracking data allow evidence-based predictions about the probability of a correct or incorrect solution even before the answer is given. This opens a promising research perspective from our point of view and we encourage other research groups in this field to investigate time-resolved prediction based on gaze data in follow-up studies based on the results of this study.

In summary, the results demonstrate that collecting eye-tracking data during the problem-solving process is useful for gaining insight into the visual strategies used in interpreting kinematic diagrams and predicting the correctness of responses based on this data. These findings can be used to sensitize teachers to the problems students face in extracting information from diagrams and distinguishing them into relevant and nonrelevant information, as well as to give more attention to this issue in teacher education. Furthermore, these are important results in terms of adapting learning environments in general and problem-solving processes in particular, since high predictive quality is essential for successful adaptation.

However, the validity is limited to selected item pairs of a specific test instrument, which are based on the same mathematical solution procedure and further research is needed to enable such an automatized support. For example, it would have to be investigated at what point in the solution process sufficient eye-tracking data are available to enable a prediction of the answer behavior with sufficient quality. Machine learning methods could also be used, which could offer advantages over the classical regression analyses used in this contribution, both in terms of the quality of the prediction and the amount of data required. In the future, intelligent tutoring systems could use such data analysis methods to identify hurdles in the problem-solving process in real time and automatically initiate precisely tailored assistance actions without any time delay.

[1] K. Wright, Eye tracking gets complex, Physics **14**, 59 (2021).

[2] L. Hahn and P. Klein, Eye tracking in physics education research: A systematic literature review, Phys. Rev. Phys. Educ. Res. **18**, 013102 (2022).

[3] A. Susac, A. Bubic, P. Martinjak, M. Planinic, and M. Palmovic, Graphical representations of data improve student understanding of measurement and uncertainty: An eye-tracking study, Phys. Rev. Phys. Educ. Res. **13**, 020125 (2017).

[4] B. Ibrahim and L. Ding, Sequential and simultaneous synthesis problem solving: A comparison of students' gaze transitions, Phys. Rev. Phys. Educ. Res. **17**, 010126 (2021).

[5] M. Kekule and J. Viiri, Students' approaches to solving R-FCI tasks observed by eye-tracking method, Sci. Educ. **9**, 117 (2018), http://ojs.pedf.cuni.cz/index.php/scied/article/viewFile/1010/551.

[6] 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).

[7] M. Kekule, Students' approaches when dealing with kinematics graphs explored by eye-tracking research method, in *Proceedings of Frontiers in Mathematics and Science Education Research Conference, Famagusta, North Cyprus*, FISER (2014) 108–117.

[8] C.-J. Wu and C.-Y. Liu, Eye-movement study of high- and low-prior-knowledge students' scientific argumentations with multiple representations, Phys. Rev. Phys. Educ. Res. **17**, 010125 (2021).

[9] 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).

[10] 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).

[11] S. Küchemann, P. Klein, S. Becker, N. Kumari, and J. Kuhn, Classification of students' conceptual understanding in stem education using their visual attention distributions: A comparison of three machine-learning approaches, in *Proceedings of CSEDU* (SCITEPRESS, Prague, Czech Republic, 2020).

[12] K. Scheiter, C. Schubert, A. Schüler, H. Schmidt, G. Zimmermann, B. Wassermann, M.-C. Krebs, and T. Eder, Adaptive multimedia: Using gaze-contingent instructional guidance to provide personalized processing support, Comput. Educ. **139**, 31 (2019).

[13] P. Klein, J. Kuhn, and A. Müller, Förderung von Repräsentationskompetenz und Experimentbezug in den vorlesungsbegleitenden Übungen zur Experimentalphysik [Promotion of representation competence and experiment relevance in the lecture-accompanying exercises on experimental physics], Z. Naturwissenschaften **24**, 17 (2018).

[14] G. M. Bowen, W.-M. Roth, and M. K. McGinn, Interpretations of graphs by university biology students and practicing scientists: Toward a social practice view of scientific representation practices, J. Res. Sci. Teach. **36**, 1020 (1999).

[15] B. Strobel, M. A. Lindner, S. Saß, and O. Köller, Task-irrelevant data impair processing of graph reading tasks: An eye tracking study, Learn. Instr. **55**, 139 (2018).

[16] R. J. Beichner, Testing student interpretation of kinematics graphs, Am. J. Phys. **62**, 750 (1994).

[17] 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).

[18] A. Lichtenberger, C. Wagner, S. I. Hofer, E. Stern, and A. Vaterlaus, Validation and structural analysis of the kinematics concept test, Phys. Rev. Phys. Educ. Res. **13**, 010115 (2017).

[19] B. Bektasli and A. White, The relationships between logical thinking, gender, and kinematics graph interpretation skills, Egitim Arastirmalari—Eurasian J. Educ. Res. **12**, 1 (2012), https://files.eric.ed.gov/fulltext/EJ1057377.pdf.

[20] Y. Bi and T. Reid, Evaluating students' understanding of statics concepts using eye gaze data, Int. J. Engin. Educ. **33**, 225 (2017).

[21] J. Han, L. Chen, Z. Fu, J. Fritchman, and L. Bao, Eye-tracking of visual attention in web-based assessment using the force concept inventory, Eur. J. Phys. **38**, 045702 (2017).

[22] A. Susac, M. Planinic, A. Bubic, K. Jelicic, L. Ivanjek, K. Matejak Cvenic, and M Palmovic, Effect of students' investigative experiments on students' recognition of interference and diffraction patterns: An eye-tracking study, Phys. Rev. Phys. Educ. Res. **17**, 010110 (2021).

[23] P. Klein, A. Lichtenberger, S. Küchemann, S. Becker, M. Kekule, J. Viiri, C. Baadte, A. Vaterlaus, and J. Kuhn, Visual attention while solving the test of understanding graphs in kinematics: An eye-tracking analysis, Eur. J. Phys. **41**, 025701 (2020).

[24] A. Susac, M. Planinic, A. Bubic, L. Ivanjek, and M. Palmovic, Student recognition of interference and diffraction patterns: An eye-tracking study, Phys. Rev. Phys. Educ. Res. **16**, 020133 (2020).

[25] 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).

[26] 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).

[27] M. Kekule, Students' approaches when dealing with kinematics graphs explored by eye-tracking research method, Eur. J. Sci. Math. Educ. **2**, 108 (2014).

[28] A. Gegenfurtner, E. Lehtinen, and R. Säljö, Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains, Educ. Psychol. Rev. **23**, 523 (2011).

[29] K.-P. Chien, C.-Y. Tsai, H.-L. Chen, W.-H. Chang, and S. Chen, Learning differences and eye fixation patterns in virtual and physical science laboratories, Comput. Educ. **82**, 191 (2015).

[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] P. Klein, S. Becker, S. Küchemann, and J. Kuhn, Test of understanding graphs in kinematics: Item objectives confirmed by clustering eye movement transitions, Phys. Rev. Phys. Educ. Res. **17**, 013102 (2021).

[32] N. S. Rebello, M. H. Nguyen, Y. Wang, T. Zu, J. Hutson, and L. Loschky, Machine learning predicts responses to conceptual tasks using eye movements, in *Physics Education Research Conference 2018, PER Conference, Washington, DC*, 2018, 10.1119/perc.2018.pr.Rebello.

[33] G. Leinhardt, O. Zaslavsky, and M. K. Stein, Functions, graphs, and graphing: Tasks, learning, and teaching, Rev. Educ. Res. **60**, 1 (1990).

[34] N. Glazer, Challenges with graph interpretation: A review of the literature, Stud. Sci. Educ. **47**, 183 (2011).

[35] 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).

[36] 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).

[37] C. Nagle, D. Moore-Russo, J. Viglietti, and K. Martin, Calculus students' and instructors' conceptualizations of slope: A comparison across academic levels, Int. J. Sci. Math. Educ. **11**, 1491 (2013).

[38] A. Susac, A. Bubic, E. Kazotti, M. Planinic, and M. Palmovic, Student understanding of graph slope and area under a graph: A comparison of physics and nonphysics students, Phys. Rev. Phys. Educ. Res. **14**, 020109 (2018).

[39] S. Ceuppens, L. Bollen, J. Deprez, W. Dehaene, and M. De Cock, 9th grade students' understanding and strategies when solving $x(t)$ problems in 1D kinematics and $y(x)$ problems in mathematics, Phys. Rev. Phys. Educ. Res. **15**, 010101 (2019).

[40] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, Comparing the force and motion conceptual evaluation and the force concept inventory, Phys. Rev. ST Phys. Educ. Res. **5**, 010105 (2009).

[41] P. Klein, A. Müller, and J. Kuhn, Assessment of representational competence in kinematics, Phys. Rev. Phys. Educ. Res. **13**, 010132 (2017).

[42] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.18.020107 for the individual items as well as the AOIs including the corresponding TVD and the expert rating.

[43] A. Agresti, *An Introduction to Categorical Data Analysis*, 2nd ed. (Wiley, New York, NY, 2007).

[44] J. R. Landis and G. G. Koch, An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers, Biometrics **33**, 363 (1977).