

Gaze-based prediction of test performance using machine learning for adaptive learning systemsYavuz Dinc,^{1,*} Sarah Malone², Verena Ruf¹, Steffen Steinert¹,
Stefan Küchemann¹ and Jochen Kuhn¹¹*Ludwig-Maximilians-University, München, Germany*²*Saarland University, Department of Education, Saarbrücken, Germany*

(Received 15 December 2024; accepted 9 October 2025; published 26 November 2025)

Prior research has demonstrated that students' performance on physics test items can be accurately predicted using machine learning algorithms based on their gaze behavior. These gaze data are typically recorded during item completion and capture students' visual attention to both the verbal item stem and accompanying visual representations, such as graphs. This predictive capability highlights the potential of machine learning as a tool for the development of AI-based adaptive learning systems, such as intelligent textbooks, for improving personalized learning. With this study, we aim to replicate the findings on the predictive power of gaze data recorded during test completion in the field of kinematics and extend them by examining whether gaze data collected during a preceding learning phase can similarly predict performance on the test items. The study involved 88 high school students who first engaged with instructional material on kinematics and then completed test items from the test of understanding graphs in kinematics, a test aimed at assessing students' conceptual understanding of kinematic graphs. Eye-tracking data were recorded both while students engaged with the learning materials and while they worked on the test items presented after each slide. To analyze these data, we used a support vector machine to predict whether students would answer the test items correctly. We compared prediction accuracies based on gaze data recorded during the learning phase with those based on gaze data recorded during the test phase. The findings indicate that eye-tracking data gathered during the completion of test items were more accurate in predicting students' performance than data collected during the learning phases. Notably, paying close attention to graphs was one of the key behaviors that distinguished students who answered correctly from those who did not. Students' gaze behavior on the graph was also most predictive of students' performance. These findings highlight the potential of gaze-based adaptive systems and the importance of including graphical representations in personalized learning environments. They also suggest that future research should explore how intelligent textbooks can combine visual representations, such as graphs, with short conceptual questions to optimize real-time adaptive support.

DOI: [10.1103/PhysRevPhysEducRes.21.020148](https://doi.org/10.1103/PhysRevPhysEducRes.21.020148)**I. INTRODUCTION**

Adaptive learning systems are promising instruments to improve the quality of teaching and learning by considering heterogeneities in the target group [1]. Intelligent textbooks are designed with the objective of optimizing the reading experience and facilitating learning gains. They capture and interpret learners' interactions with the book [2]. Such interactions with the book may be obtained from comprehensive data sources, including log files. In instances where direct engagement with the digital textbook (e.g., turning pages and pressing response buttons) is limited, alternative

indicators are necessary to infer the learner's cognitive state. To obtain comprehensive insights into learners' engagement and cognitive processes, nonintrusive physiological measures, such as metrics of gaze behavior, can also be employed [3].

These insights can not only inform the future design of intelligent textbooks but also deepen our understanding of learning processes in complex domains. Moreover, the widespread adoption of browser-based eye-tracking frameworks such as WebGazer.js [4] and AI-driven methods like iTracker [5] has dramatically lowered the barrier to capturing precise gaze data with widely available hardware. Running directly in standard web browsers, these tools can deliver real-time gaze metrics without specialized eye-tracking setups. This democratization of gaze analytics enables seamless integration into digital learning environments, delivers feedback, and personalizes support in real time [6].

While eye-tracking shows great promise for adaptive learning systems [7], its value also lies in how research

*Contact author: Dinc.Yavuz@physik.uni-muenchen.de

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

indicates that specific gaze data effectively depict cognitive learning processes such as the selection, organization, and integration of multimedia (i.e., combinations of text and images) information [8]. However, an initial series of studies conducted by Scheiter *et al.* [9] revealed that solely relying on learners' observed gaze behavior to provide adaptive learning support does not necessarily lead to improved learning outcomes. One possible issue described by the authors was that the captured gaze metrics were rather coarse and, therefore, not easily interpretable, even though they were generally suitable for distinguishing between different groups of learners. Instead of solely examining gaze patterns on text and images as in previous studies, recent research in physics education has taken a more nuanced approach by analyzing eye movements within specific, meaningful regions of graphs [10,11]. Graphs play a central and ubiquitous role in physics (education)—as well as in other science, technology, engineering, and mathematics (STEM) disciplines. In this regard, research has, for example, identified expert-novice differences in gaze patterns during learning, problem solving, and performing test items with graphs [12]. Moreover, the predictive nature of gaze data in relation to performance has been demonstrated in various studies focusing on physics tasks involving graphs [10,11,13]. Klein *et al.* [14] demonstrated that specific gaze data can be used to distinguish solution accuracy in physics problems. Successful strategies for problem solving with graphs were consistently reflected by a higher focus on relevant graph areas and more transitions across the graph [11].

Based on these promising results from eye-tracking studies on solving physics tasks that include graphs, the present study aims to investigate whether gaze data collected during learning can predict conceptual understanding on graph tasks in the domain of kinematics. Rather than relying solely on eye-tracking data gathered during test item completion, our study also captures eye movements as learners interact with instructional materials that include graphs. This approach aims to determine whether such real-time metrics can serve as early indicators of a learner's mastery of the material.

We structure the manuscript as follows: First, we outline the relevant theoretical foundations and define key terms to establish the research framework. Next, we detail the methodology, including the design of the learning materials, the implementation of the eye-tracking environment, and the phases of data collection. We also describe the collected data and the analytical methods applied. In the subsequent section, we present the results, followed by a discussion that contextualizes the findings and explores their implications for educational research. Finally, we address the study's limitations, propose directions for future research, and conclude with a concise summary.

II. THEORETICAL BACKGROUND

A. Learning and testing with multiple external representations

Multiple external representations (MERs) are essential in teaching STEM concepts, especially in physics education. Physics as a discipline uses mathematical models to describe everyday phenomena and often supplements these with different forms of representation, such as formulas, images, or tables [15]. This diversity makes it possible to view complex content from different perspectives. The ability to switch between representations makes it easier for learners to find individually suitable approaches or to focus on the representation that best meets the specific requirements of a task [16]. Another advantage of MERs lies in their ability to promote learning by presenting information in both pictorial and textual forms [17]. This is assumed to be more effective than learning with text only. The simultaneous provision of multiple representations can encourage learners to develop a more flexible and diverse understanding, which can support the development of a deeper understanding of the underlying concepts [18].

Studies show that providing more than one representation can be particularly helpful when the representations present information that complements each other or limits certain possible interpretations so that misunderstandings are avoided [19]. By considering different representations, learners can develop a broader understanding that they can also transfer to new contexts and problem-solving tasks [20]. At the same time, MERs can help learners to make abstract concepts more accessible and link them to more concrete examples.

Moreover, a meta-analysis by Rexigel *et al.* [21] found that using even more than two external representations in STEM education positively impacts performance and reduces cognitive load, provided that appropriate support measures are in place.

MERs can also be relevant for educational assessment, testing, and problem solving [22,23].

The multimedia effect postulated for learning has also been found for testing, which means that performance in educational tests is enhanced when test items not only include text or numbers but also graphics and pictures providing essential information [24].

In summary, empirical evidence indicates that MERs can enhance performance, both in learning contexts and in subsequent assessments. This effect can be theoretically explained by referring to the specific cognitive processes assumed to be engaged when working with MERs. Regarding learning, Mayer's cognitive theory of multimedia learning (CTML) identifies three central cognitive processes: the selection of relevant information from the available representations, the organization of this information in mental structures, and the integration of these

structures with each other and with the prior knowledge stored in long-term memory [25].

Mayer originally developed CTML to explain learning from combinations of text and images, as these are among the most common multimedia formats. However, the principles are not limited to this specific combination. The three cognitive processes are also applicable to other combinations of representations, such as diagrams and equations. While Mayer's theory primarily relates to learning with MERs, there is also evidence that similar cognitive processes play a crucial role in solving test items that include MER. The problem representation phase, the initial stage of solving a test item, is considered to closely parallel the processes underlying learning with MERs [26]. During this phase, test takers need to construct a mental model of the problem that integrates information from the MERs. This process involves selecting relevant information from all available representations, mentally organizing and structuring it, and integrating it both across representations and with prior knowledge. Only after this foundational step do processes specific to test item solving, such as decision making, come into play.

Malone *et al.* [27] demonstrated that heterogeneous MERs, such as combinations of text and equations, improve test performance by supporting the integration of different forms of information. Similarly, Lindner *et al.* [28] found that multimedia effects originally postulated for learning also apply to testing and problem-solving tasks. Lindner [26], however, emphasized that a deeper investigation of the cognitive processes involved in solving test items that include MERs requires the use of process measures, such as eye tracking.

In light of this theoretical and empirical foundation, the present study focuses on tasks that include MERs, specifically combinations of verbal information and graphs, which are prevalent in physics education. Since both the learning materials and the test items used in our study are graph-based, they require learners to engage in the cognitive processes of selection, organization, and integration as described by CTML. These processes are assumed to be reflected in gaze behavior. Thus, analyzing gaze data in graph-based learning and testing contexts offers a promising avenue for gaining insights into students' cognitive

processing and for predicting their performance. The present study builds on this rationale by investigating whether gaze behavior during learning and testing with MER can serve as a valid predictor of conceptual understanding in kinematics.

B. Eye-tracking in physics education research

Cognitive processes during learning and solving test items with MERs—such as selecting, organizing, and integrating information—are visible in students' eye movements [29]. Therefore, researchers often investigate these processes by looking at participants' gaze data. In educational research on learning and testing with MERs, a broad approach is typically taken, focusing on how often and for how long individual representations are fixated, as well as how frequently gaze shifts occur between them [6,18]. In contrast to general educational research, physics education research commonly uses eye tracking as a methodology and places greater emphasis on detailed analyses of gaze behavior within individual representations—for example, how learners interact with essential parts of graphs [30]. For example, many studies use eye tracking to investigate specific cognitive processes involved in solving graph-based problems [12]. Klein *et al.* [14] investigated the gaze data of 41 physics majors on vector field tasks. They found that saccades and fixations can be indicative of success in test items. Other gaze metrics reflecting cognitive processes have also shown relevance during solving test items with graphs; for example, successful solvers tend to allocate more attention to task-relevant areas of the graph [12]. This aligns with the information-reduction hypothesis [31] and suggests increased organizing of the information [8]. This is consistent with findings that older students with more conceptual knowledge focus more on relevant parts of physics problems [32]. Successful item solvers also switch more between parts of the graph [11,12], which could indicate that more integrative processes are necessary for solving test items successfully [8]. Moreover, gaze data have been used in multiple studies to reliably predict success in solving test items [11,13,33]. An overview of relevant eye-tracking metrics is provided in Table I.

TABLE I. Definitions and interpretations of the eye-tracking metrics used in this study.

Metric	Definition	Interpretation
Dwell time	Sum of all fixation and saccade durations within a given AOI	Total amount of visual attention allocated to that region
Total fixation duration	Sum of all fixation durations (within a given AOI)	How long did participants fixate that AOI
Number of fixations	Count of individual fixation events (within a given AOI)	Frequency of attention shifts within that AOI
Transitions	Number of gaze shifts from one AOI to another	Degree of integration or switching between different elements of the stimulus

C. Machine learning with eye-tracking data

Previous research could successfully predict students' performance based on their eye movements using machine learning (ML)-based methods. For example, Küchemann *et al.* [11] used the gaze data of 115 high-school students who solved a line-graph item of the test of understanding graphs in kinematics (TUG-K) [34] to predict their success. They employed a support vector machine (SVM) to make these predictions. An SVM is a large-margin classifier that constructs a kernel-based, multidimensional decision boundary to distinguish between classes, such as correct and incorrect solvers, and seeks to maximize its margin to the training instances [35]. The input features included students' visit durations on areas of interest (AOIs) and gaze transitions between those AOIs [11]. They observed that the accuracy metrics increase with increasing training size. Other researchers compared various ML algorithms for classifying performance, including an SVM, k -nearest neighbor, and random forest, and found that the SVM performed best [13]. A high accuracy was achieved, for example, for classifying students' problem-solving strategies on vector field problems using SVMs trained on gaze data [36].

SVMs are trained on labeled data, such as problem-solving success, to predict the outcome for unknown data based on the input, such as gaze data [35], and can be used in real time [37,38]. This is a necessity for increasingly common adaptive learning systems [39] that should provide immediate feedback to learners, for example, in the context of intelligent textbooks. Adaptive learning systems often use ML techniques because they perform better than statistical methods [40]. Due to its successful application in predicting performance on physics tasks [11,13,33], we chose this algorithm for our analyses.

D. Research questions

Previous studies have analyzed and demonstrated the predictive nature of gaze data predominantly in the context of solving test items [10,11]. Eye movements within a test item could thus be successfully utilized to predict whether an individual would solve that specific problem correctly. While solving test items that include MERs involves cognitive processes similar to those described in the context of learning with MERs, there are also important differences in the depth and purpose of processing. As Hahn and Klein [30] note, gaze patterns and cognitive demands may differ between learning and assessment scenarios. Nonetheless, it is an open question whether gaze data collected during learning phases contain sufficient information to predict subsequent performance on test items. Thus, the present study aims to replicate previous findings and extend them by investigating the extent to which gaze data recorded during learning can predict later test performance. The following research questions arise from these considerations:

RQ1: Can previous findings be replicated by using machine learning algorithms to predict performance on conceptual physics test items (TUG-K) [34] based on gaze data recorded during task solving?

H1: Correct and incorrect solvers of TUG-K test items can be distinguished using machine learning algorithms trained on gaze data collected during task solving.

RQ2: Can machine learning algorithms also predict performance on the same test items using gaze data recorded during the preceding learning phase?

H2: Gaze data collected during the learning phase contain sufficient predictive information to allow machine learning algorithms to distinguish between correct and incorrect solvers on subsequent physics test items.

We also investigate the features that are relevant for predicting participants' performance on physics test items. Consequently, we posit the following explorative research question:

RQ3: Which features are relevant for predicting task-solving performance when using the gaze data recorded on the learning material and when using the gaze data recorded on the physics test item as input for the ML algorithm?

III. METHODS

To answer our research questions, an experimental design was employed to collect gaze data from participants while they engaged with learning materials with MERs and subsequent test items on kinematic concepts.

A. Participants and procedure

The sample consisted of 88 ninth-grade students from two secondary schools. Participants conducted the study in groups of up to 15 students per session, under the supervision of trained researchers. The researcher first explained the study's purpose and procedures to the participants, as well as how eye-tracking works.

Each participant was sitting in front of a laptop with a 15-inch screen, equipped with an eye-tracking device, specifically the Tobii Spark remote system. The device operates with a sampling frequency of 60 Hz, offering a precision of 0.26° rms and an accuracy of 0.45° under optimal conditions [41]. Before participants began the learning environment and tasks, the eye-tracking device was calibrated for each individual using the Tobii Eye-Tracking Manager. During this calibration process, participants followed a dot that moved around the screen, allowing the eye-tracking device to accurately track their eye movements. After that, the supervisor guided them to the learning environment, which had been installed as a desktop application beforehand. Participants first viewed a learning slide on kinematics (see Fig. 1). They then completed two tasks based on this slide. The next learning slide was subsequently presented with 2 more tasks, and so on until all 6 instructional slides and all 12 tasks had been

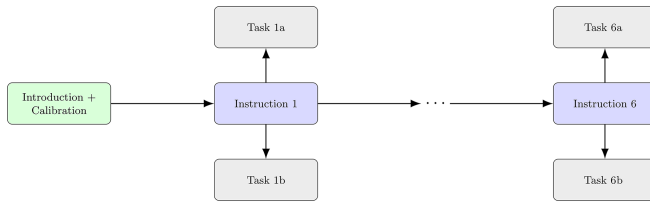


FIG. 1. Schematic flow of the study: Each instructional block (learning phase) was followed by two topic-related test items from the TUG-K (testing phase).

completed by every participant. After completion, the file was saved, and the eye-tracking data were exported for further analysis. The eye-tracking data collected during the study were then analyzed to identify specific eye movement metrics associated with conceptual understanding in the field of kinematics. Fixations were detected using an I-VT algorithm [42].

B. Learning environment

The learning environment was created using HTML and CSS to offer an interactive and engaging platform for the participants. In the learning environment, students navigated using a button to continue without being able to jump back to previous slides. The learning material [43] was based on conventional textbooks and, as shown in Fig. 2, was designed with multiple representations—including text, formulas, and graphs. The students were first given a learning slide on a specific topic, such as “accelerated motion.” Following each learning slide, which introduced specific kinematics concepts, two related test items from

the standardized TUG-K instrument [34] were presented to assess learners’ understanding of the just-covered content. The students were asked to formulate their answers freely and provide an explanation instead of clicking on an answer option as intended with the original TUG-K design. In general, no response options were displayed. Only for those items that, per design of the original instrument, included graphical answer options, students were instructed to enter the corresponding letter (e.g., A, B, C, ...) into the text field together with their explanation. All TUG-K items used as task items in this study are shown in Table II. We selected TUG-K items representing foundational, well-scoped kinematics concepts and designed the learning slides to match these targets. The instructional materials were thus tailored to the chosen items to maximize content alignment.

Performance in the test items was coded in the form of a binary variable (0 = incorrect, 1 = correct) and used as the target variable for prediction. Throughout this process, an eye-tracking device recorded their eye movements. To accurately link the data with the presented slides, we captured coordinates, timestamps of eye movement, as well as the final answer and its corresponding slide number.

C. AOI creation and definition

Areas of Interest (AOIs) refer to specific regions of a stimulus (such as a learning slide) that are particularly relevant for analysis and form the foundation of eye-tracking measurements. For instance, a whole graphical representation can serve as one large AOI or, for more detailed analyses, multiple specific parts of a graphic can be defined as separate AOIs to capture more fine-grained

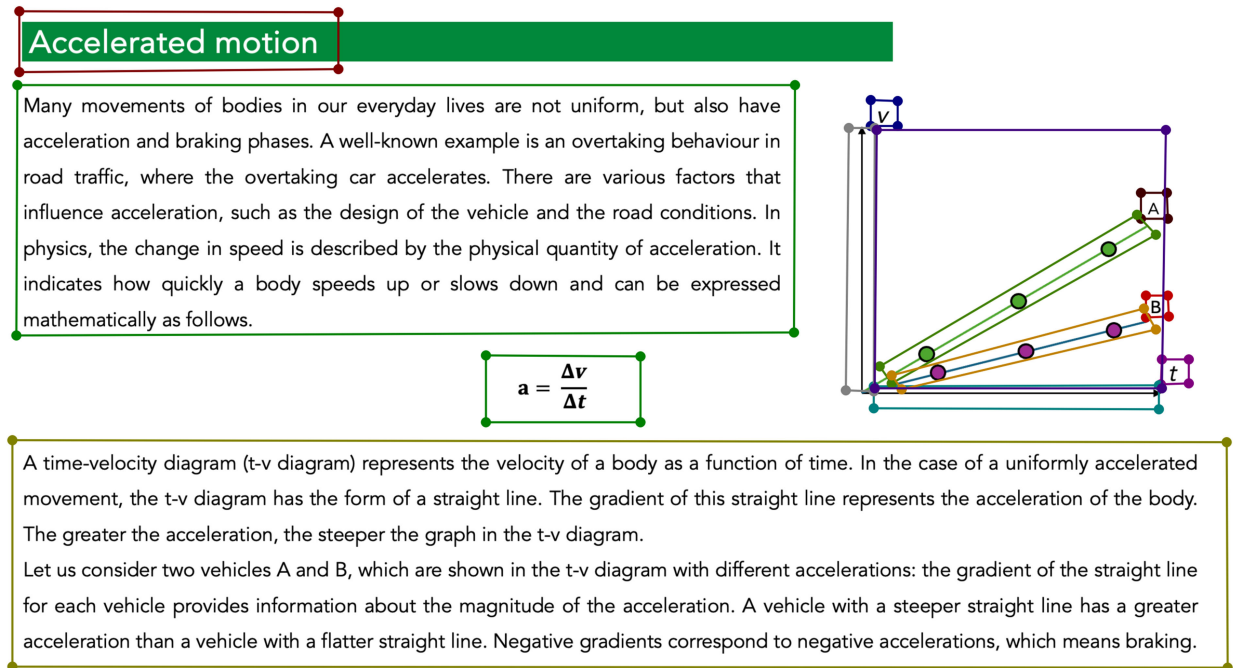


FIG. 2. Learning material with defined AOIs.

TABLE II. Overview of the tasks with corresponding item numbers according to the TUG-K numbering. The table reports the proportion correct for each task and indicates whether it was included in the analysis.

Study order	TUG-K item number	Solution probability	Included
1	8	25/85	Yes
2	3	29/85	Yes
3	13	41/85	Yes
4	11	8/85	No
5	18	53/85	Yes
6	17	8/85	No
7	2	44/85	Yes
8	21	45/85	Yes
9	12	7/85	No
10	14	28/85	Yes
11	7	14/85	No
12	6	12/85	No

aspects of visual attention. These regions can include key points on the graph, such as data points, labels, and axes. We systematically set up AOIs for all test items and learning slides, taking into account relevant representations such as formulas, text, and graphs, as well as graph-specific AOIs relevant for task solving, as shown in Figs. 2 and 3. This systematized structure follows the framework proposed by Becker *et al.* [10], which distinguishes between macrolevel and microlevel AOIs. Macrolevel AOIs include the main components of each item: the question text, the diagram area, and, where applicable, the graphical answer options. These regions reflect the general structure shared by all test items and were therefore applied uniformly across them to ensure consistency. The same systematics was used for the learning slides, analogous to Fig. 2. Microlevel AOIs, in contrast, involve a finer segmentation of the diagram itself. These include the x and y axes, axis labels, the graph line (function curve), specific points referenced in the question, and subregions of the graph that are either relevant or irrelevant for solving the task. These subregions were identified by experts in physics education based on whether they contributed meaningfully

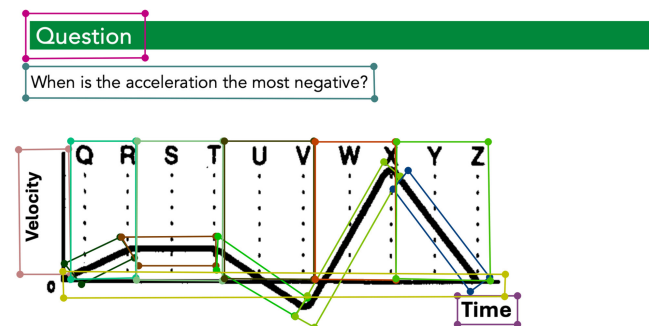


FIG. 3. Example test item (TUG-K; item 2) with defined AOIs.

to the mathematical solution (e.g., areas used for calculating slope). As shown in Fig. 3, we assigned an AOI to every interval in the graph area and treated the curve itself as a separate AOI within each interval. Thus, the interval is labeled “x-z,” whereas the corresponding curve segment is labeled “x-z-line.” A transition from the interval AOI to its curve segment is therefore coded as “transition_x-z_to_x-z_line.”

To ensure comparability across different test items, we applied a standardized tagging system to all AOIs. Each AOI was labeled accordingly using categories such as text, graph, or option. This approach enabled us to extract and compare gaze behavior in a structured manner across all learning slides and test items. We chose this systematized AOI structure with the aim of maximizing the predictive performance of our model, as previous studies have shown that specific gaze patterns—particularly in solution-relevant areas—can distinguish between correct and incorrect responses [44]. The eye-tracking metrics were then calculated based on participants’ gaze patterns within these defined AOIs. An overview of the metrics used in this study is presented in Table I [30]. The AOIs were created using the publicly available software Labelme, which can be used to generate any type of polygon [45].

D. Classification

We used SVMs as the classification technique in this study because it has been shown to be effective in previous research for predicting task-solving success based on eye-tracking data [11]. For this, we trained the SVM using the eye-tracking metrics and the performance on selected TUG-K test items of the participants. First, we aimed at reproducing the findings of previous studies where the metrics were defined on the same stimuli and used for predicting the task-solving performance based on gaze data (RQ1). Second, we aim to transfer this approach by using the same metrics, but defined on the instructional slides shown before, to predict task-solving performance based on gaze data during learning (RQ2). In addition, we analyzed feature importance, which indicates the extent to which each feature contributes to the model’s classification decisions (RQ3).

E. Data analysis

To address the first two research questions, we focus on analyzing the eye-tracking metrics collected during the testing phase and the learning phase to investigate to what extent these metrics predict students’ task-solving performance.

To evaluate the performance of our models across varying training set sizes, we implemented a comprehensive pipeline that includes data preprocessing, feature selection, dimensionality reduction, and model optimization. Initially, a Random Forest classifier was employed to identify the 20 most important features from each dataset, and these were subsequently used to train the models. While this feature selection was conducted on the entire dataset prior to splitting the data into training and test sets—which may lead to a slight

overestimation of performance—the effect is likely minimal, as performance metrics were averaged across multiple test items and slides of learning material. Importantly, the same procedure was applied consistently during learning and testing to ensure comparability. We then filtered the datasets to exclude instances where the “Total Fixation Duration” was less than a specific threshold set at 2 s, as such instances were considered insufficient to provide meaningful insights. Additionally, participants who completed fewer than two tasks were excluded, as their data were considered insufficiently processed and unsuitable for further analysis. Next, we split the data into training and testing sets using stratified sampling to maintain consistent class distributions across both sets. The proportion of data allocated for training was varied incrementally from 10% to 90%, allowing us to assess the model performance across different training sizes. To address slight class imbalances and reduce the dimensionality of the data, the preprocessing pipeline integrated standard scaling, principal component analysis (PCA), and the synthetic minority oversampling technique (SMOTE). PCA was applied to retain 95% of the variance, effectively reducing dimensionality while preserving most of the information. Since the selected datasets were already nearly balanced due to preselection (see results), SMOTE was specifically used to correct minor residual imbalances without introducing significant drawbacks from synthetic data generation.

For the model training, we selected an SVM classifier, which was optimized through hyperparameter tuning using a Randomized Search with fivefold cross-validation. The hyperparameters tuned included the number of components retained by PCA, the regularization parameter CC, and the kernel type and gamma parameter for the SVM. The randomized search was conducted over 100 iterations to identify a suitable model configuration. This number was chosen to ensure an efficient trade-off between computing time and model performance.

The model performance was evaluated on the test set using both the F1-score and the Matthews correlation coefficient (MCC).

The F1-score is the harmonic mean of precision and recall, balancing false positives and false negatives. It is particularly useful for imbalanced datasets, ensuring that both the ability to correctly identify positive cases (recall) and the reliability of positive predictions (precision) are considered. F1 ranges from 0 to 1, where higher values indicate better classification performance. The MCC is a more comprehensive metric that takes all four elements of the confusion matrix (true positives, true negatives, false positives, and false negatives) into account. Unlike the F1-score, it provides a balanced assessment even when class distributions are skewed [46]. MCC ranges from -1 (complete misclassification) to 1 (perfect classification), making it particularly valuable for evaluating models on imbalanced datasets [46].

Although the interpretation of performance metrics varies by context, an F1 score of 0.7 is often seen as acceptable, and scores above 0.8 are commonly interpreted as reflecting good performance in eye tracking [47,48]. For the MCC, scores above 0.4 are typically viewed as moderate, while values over 0.6 suggest good predictive quality [47]. For each training size, we evaluated the F1 score and MCC separately for each TUG-K item (dataset) within the respective scenario (learning or testing) and then reported the mean and standard deviation across all TUG-K items. Thus, the standard deviation reflects variability between these questions.

This approach allows for a robust comparison between the learning and testing phases by aggregating across multiple items per scenario.

IV. RESULTS

The analysis is based on the performance of 85 participants (3 were excluded due to predefined criteria mentioned in the section “Data analysis”) across 12 tasks, recorded in a binary format (1 = correct, 0 = incorrect). Descriptive statistics revealed a performance range from 8.2% to 62.3% across tasks, indicating that some test items were significantly more challenging. A linear regression showed no significant increase or decrease in students’ performance over the course of the tasks ($p = 0.976$), suggesting that participants’ performance remained consistent. For the SVM analysis, only tasks with a balanced success rate were considered to avoid bias introduced by using imbalanced data. Thus, we excluded items with a correctness rate below 25% in order to maintain the interpretability and robustness of the analysis. This threshold avoids severe class imbalance, where majority-class baselines can appear deceptively strong, and prevents inflated performance estimates. It also improves comparability across items by keeping difficulty within a reasonable range. The final set of seven items (see Table II) exhibited adequate performance levels and formed the basis for all subsequent analyses.

A. Prediction based on test items

In this section, we examine the results of the first research question, in which we analyze the gaze data on the test item slides, aiming to replicate the results of previous research [11]. First, we discuss the key metrics using an example test item (Fig. 3), show statistical significances, and report the predictivity using the F1 and MCC values on this slide. We then present the results aggregated across all item slides to determine an average prediction quality.

The analysis of feature importance (Fig. 4) revealed that some features contributed more strongly to the model’s ability to differentiate between correct and incorrect responses.

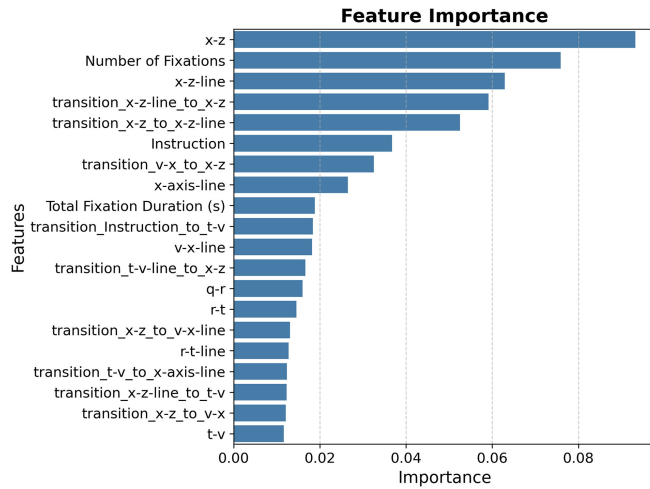


FIG. 4. Feature importance for example test item (see Fig. 3; TUG-K item 2).

The feature “x-z” (see AOI definition) showed the highest importance (Importance = 0.093), followed by “Number of Fixations” on the entire item (Importance = 0.076) and “x-z-line” (Importance = 0.063). These features are the most relevant for predicting differences between performance groups, though further analysis is needed to clarify their role and determine any causal relationships.

In addition, a Mann-Whitney U test was performed to identify significant differences between the two performance groups (performance = 0 and performance = 1). The test revealed that several numerical features, such as “x-z” ($p < 0.001$, $U = 224.0$), showed significant differences ($p < 0.05$). A complete overview of the first features can be found in Table III. The combination of feature importance and statistical significance suggests that visual metrics such as fixation and transition are associated with performance differences on this item and may help in predicting task performance.

The comparison of the model performance with different training sizes shows a clear positive correlation between the training size and the model performance, measured by the Matthews correlation coefficient (MCC) and the F1 score. Figure 5 shows that the MCC starts at almost zero for a training size of 10% and increases, reaching a peak value of 0.79 ± 0.14 at 90%. The F1 score shows a similar pattern, starting at 0.5 (random classification) and peaking at 0.89 ± 0.07 for 90% of the data. The standard deviations of the results show no trend with increasing training size.

In the next step, we consider all test items for which there is a balanced ratio between correct and incorrect solvers. For these items, the accuracy metrics are analyzed again across all training variables. In this context, the standard deviation represents the fluctuations between the individual measurements but not the variability during training for each item.

TABLE III. Mann-Whitney U test for task-solving.

Feature name	U statistic	p value
x-z	224.00	<0.001
transition_x-z-line_to_x-z	310.50	<0.001
transition_x-z_to_x-z-line	342.00	<0.001
x-z-line	412.50	<0.001
transition_x-z_to_v-x	489.50	<0.001
Number of Fixations	459.50	<0.001
transition_t-v-line_to_x-z	676.50	<0.001
transition_v-x_to_x-z	574.00	0.002
transition_Instruction_to_x-z	672.50	0.006
Total Fixation Duration (s)	598.00	0.008
transition_t-v_to_t-v-line	625.50	0.008
transition_t-v-line_to_x-z-line	758.50	0.008
transition_x-z-line_to_v-x-line	717.50	0.01
transition_y-axis_to_x-z	681.50	0.01
transition_x-axis-line_to_t-v	1127.50	0.01
transition_x-z_to_v-x-line	680.00	0.016
transition_v-x-line_to_x-z-line	694.00	0.017
transition_t-v_to_x-axis-line	1105.50	0.02
transition_x-axis-time_to_x-z-line	799.50	0.028
transition_x-axis-line_to_q-r	799.50	0.028
transition_v-x_to_q-r	1037.00	0.034

The results in Fig. 6 show that the model performance averaged over all balanced items again increases with increasing training size. For smaller training sizes (e.g., 30%), the values of the metrics considered—F1 score, MCC, and accuracy—are close to random probability. This indicates that the model is not able to recognize meaningful patterns with limited training data. At even smaller sizes (0.1 and 0.2), some metrics drop to zero, which is due to training sets that contain only one class or no correct responses, making meaningful classification impossible.

With a larger training size, the averaged model performance improves significantly. For example, with a training

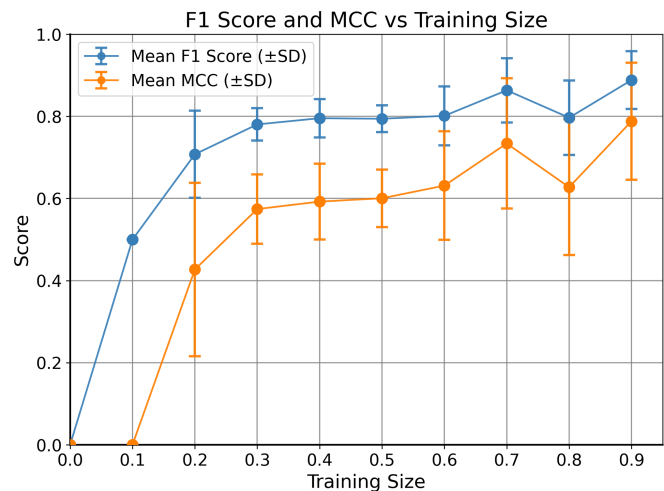


FIG. 5. F1 and MCC for different training sizes on example test item (see Fig. 3; TUG-K item 2).

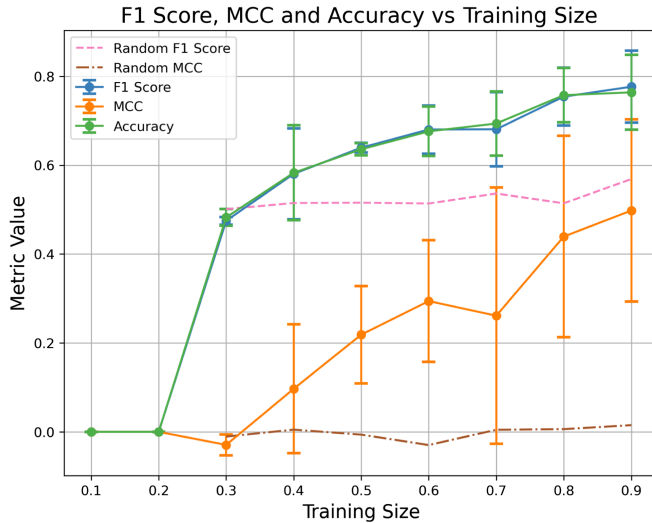


FIG. 6. F1 and MCC for different training sizes across all items.

size of 80% or more, a high reliability of predictions is achieved, characterized by an F1 score of above 0.7 and an MCC of 0.4 on average across all items. However, the standard deviations of the metrics show a high variance, which indicates that the prediction quality can vary widely for a fixed training size depending on the task item.

Figure 6 also shows the results of a permutation test, in which the target labels were randomly permuted 100 times while preserving the original class distribution. This approach removes any meaningful association between features and labels, allowing us to estimate how well the model would perform if there were no learnable structure in the data. The metrics remain almost constant regardless of the training size and fluctuate around random probability. The comparison between permutation results and actual model evaluations shows that the model delivers significantly better predictions with larger training sizes than with random classification. This indicates that the model is able to capture meaningful patterns beyond what would be expected under random labeling. While we report the permutation test as a suitable baseline in the context of our pipeline, a majority-class predictor reflects a different, complementary perspective on model performance under class imbalance.

B. Prediction based on learning

In this section, we examine the predictive performance of the SVM model with respect to the binary task-solving performance, which is evaluated based on the gaze data from learning slides. The approach is analogous to testing, as described in the methodology section. First, as in the previous chapter, we analyze a single example, show the most important features relating to this slide (Fig. 2) and the predictivity measured with the same metrics, and then again present the average model performance across all learning slides.

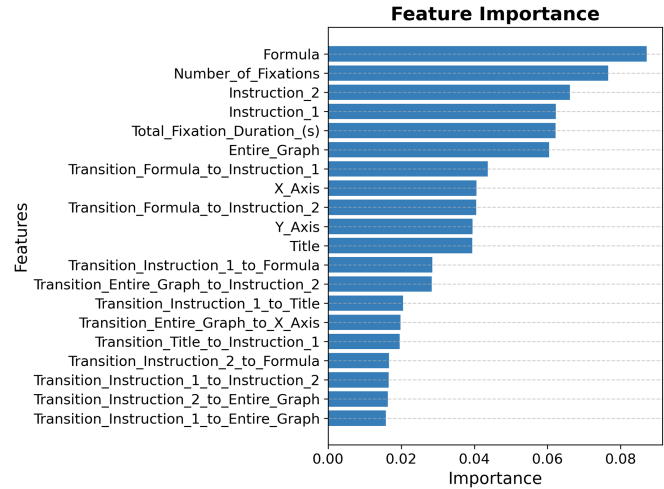


FIG. 7. Feature importance for learning on the example slide.

The feature formula (see AOI definition) showed the highest importance in Fig. 7 (Importance = 0.087), followed by “Number of Fixations” (Importance = 0.077) and “Instruction 2” (Importance = 0.066).

The Mann-Whitney U -test revealed that several features exhibited statistically significant differences between correct and incorrect solvers, with a significance level set at $p < 0.05$. Specifically, the analysis found that the Number of Fixations ($p = 0.035$), the transition from Instruction 1 (first textual representation in Fig. 2) to Formula ($p = 0.029$), the transition from Formula to Instruction 1 ($p = 0.044$), the transition from Formula to Instruction 2 (second textual representation in Fig. 2) ($p = 0.008$), the transition from the Overall Graph to Instruction 1 ($p = 0.044$), and the time spent on the Formula ($p = 0.011$) all showed significant differences between the two groups (see Table IV).

The analysis of the model performance (Fig. 8) on the example learning slide shows that both the F1 score and the MCC increase with increasing training size. The F1 score reaches its highest value at a training size of 0.9 (0.671 ± 0.115) while the standard deviation is largest at smaller training sizes, especially at 0.3, with a standard deviation of ± 0.240 . The MCC follows a similar trend, increasing from 0.0 at minimum training size to a peak value of 0.471 ± 0.218 at 0.9.

Figure 9 illustrates the F1 score and the MCC for both the testing and learning scenarios across all items. It is noteworthy that the F1 score for testing reaches a high value within the initial training sizes compared to the learning curve. At a training size of 0.9, the model achieves an F1 score of 0.776 ± 0.081 for testing, while the score for learning is slightly lower at 0.733 ± 0.170 . Additionally, the training curve of the F1 score for learning is observed to be below that of testing for all values and remains within the range of random probability up to a training size of 0.7.

A similar trend is observed for the MCC score. For learning, the MCC starts at 0.0 ± 0.0 and increases to

TABLE IV. Mann-Whitney U test for learning.

Feature name	U statistic	p value
transition_formula_to_instruction_2	602.5	0.008
formula	597.5	0.011
transition_instruction_1_to_formula	646.5	0.029
Number of Fixations	646.0	0.035
transition_formel_to_Instruktion_1	670.5	0.044
transition_entire_graph_to_instruction_1	678.0	0.044
Total Fixation Duration (s)	683.0	0.076
transition_y_axis_to_formula	800.5	0.106
instruction_1	703.5	0.112
entire_graph	704.5	0.114
transition_formula_to_y_axis	924.5	0.15
transition_instruction_1_to_title	741.5	0.164
transition_entire_graph_to_instruction_2	1020.0	0.164
transition_title_to_formula	840.5	0.17
transition_y_axis_to_title	840.5	0.17
transition_instruction_1_to_y_axis	820.5	0.187
transition_instruction_2_to_formula	746.0	0.189
transition_instruction_1_to_entire_graph	758.5	0.196
transition_x_axis_to_instruction_1	785.0	0.217
transition_formula_to_entire_graph	772.5	0.22

0.577 ± 0.375 at a training size of 0.9. The standard deviations for both metrics in the learning phase indicate greater variability compared to the testing scenario.

C. Relevant features for prediction in both contexts

To analyze the most important features for the prediction of testing and learning and find similarities between them, the features were systematically divided into categories based on keywords in the names. The following categories were defined: Graph Features (e.g., elements of graphs or axis labels), Fixation Features (e.g., total fixation number and duration), Text Features (e.g., instructions or headings), Formula Features (mathematical expressions or formulas),

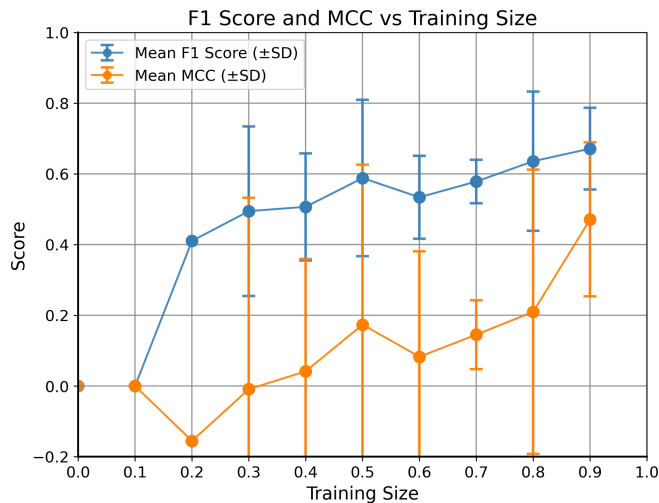


FIG. 8. F1 and MCC for different training sizes during learning.

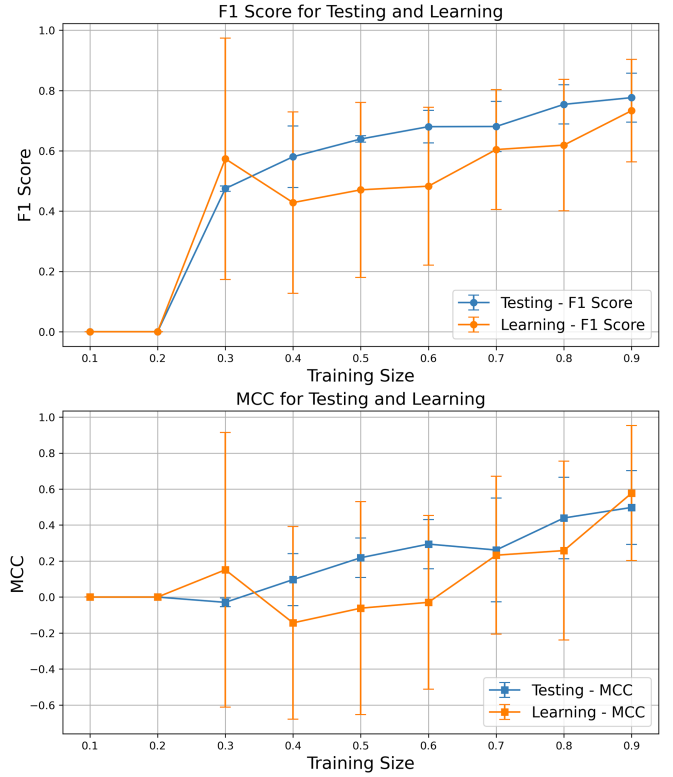


FIG. 9. Comparison of model performance for testing and learning across all items.

and Option Features, which are specific to testing where possible solution options are shown.

First, the average importance per category and context was calculated, and then, to better understand their role in prediction, the sum of all present features was considered. The average importance was calculated for both contexts of the respective categories, which are shown in Table V. The fixation features, which represent the total time on the slides, have similarly high average importance values in both contexts: 0.008 in the learning context and 0.009 in the testing context. The graph features have comparatively low values in both contexts: 0.008 in the learning context and 0.009 in the testing context. The text features have an average importance value of 0.021 for learning and 0.016 for testing.

The Formula Features occur only in the learning context and have an average Importance value of 0.0122. In contrast, the Option-Features occur only in the testing and have an average importance value of 0.0556.

TABLE V. Average feature importance.

Category	Learning	Testing
Fixation	0.049	0.056
Formula	0.012	...
Graph	0.008	0.009
Text	0.021	0.016
Option	...	0.055

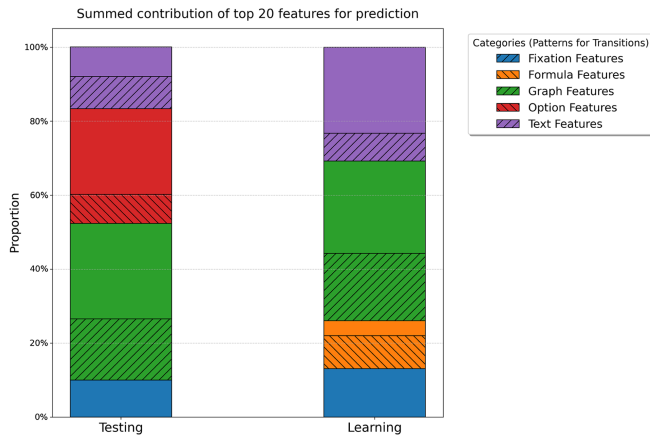


FIG. 10. Summed contribution of top 20 features for testing and learning.

Next, the top 20 features for each test item were then identified based on their importance in order to analyze the frequency and distribution of the categories. To assess the relative importance of the categories, a weighted analysis was performed. For this purpose, the sum of the importance of all associated features was calculated for each category in order to determine their relative contribution to the overall prediction performance across all datasets. The results are visualized in a stacked bar chart (Fig. 10) comparing the weighted importance of both the testing and learning phases.

Figure 10 shows that Graph Features have the highest relevance in both scenarios when summing up the 20 most relevant features per category. They account for 44.5% of the weighted importance in testing and 43.2% in learning, making them the dominant category regarding predicting the task-solving performance in both contexts.

Text Features follow with 17.5% of the weighted importance for testing and a significantly higher contribution of 30.8% for learning. However, this shows a stronger presence of this category in the learning context, with its contribution being almost twice as high as during the testing phase. This difference makes text features the second most important category in learning.

Fixation Features, which measure the total visual attention, show similar contributions in both scenarios, with 10.0% in testing and 13.1% in learning.

Furthermore, we analyzed which categories, weighted by their importance and frequency, contribute most frequently to model prediction, distinguishing between features related to dwell times and transition features. This showed that dwell time features play a dominant role in both scenarios, with 68.9% of the weighted importance in the testing context and 65.4% in the learning context. In comparison, transition features contribute less to prediction, with 31.1% in testing and 34.6% in learning, as shown by the patterned areas in Fig. 10. This suggests that dwell time on specific AOIs contributes more strongly to the

model's ability to distinguish between performance levels, whereas transitions appear to play a relatively smaller role. These insights may inform feature selection in similar modeling contexts.

V. DISCUSSION

A. Findings related to the research questions

In this work, we analyzed the predictivity of eye-tracking metrics in two contexts: one for testing and one for learning slides. The aim was to predict task-solving performance on items from the TUG-K in the area of kinematics based on the participants' gaze data (RQ1). In the first part, we replicated previous work, including that of Küchemann *et al.* [11] and, similar to them, observed that the accuracy metrics increase with increasing training size. Task-specific analyses of the AOIs showed that certain features exhibit significant differences between the tasks. These AOIs often cover areas relevant to the task solution, such as increased attention to specific graph areas (e.g., x-z area). Across multiple tasks, we calculated the same accuracy metrics, aggregated their means for the respective training variables, and reported the standard deviations. The general trend shows that the model accuracy for small training sizes is initially at a random level but increases gradually with increasing training size and reaches a maximum.

The F1 score of 0.776 ± 0.081 shows that the model has a good ability to make consistent predictions. This is in line with the literature, where an accuracy in the range of 70% was reported for similar approaches with SVM [11]. Another study in this field, which was based on a significantly higher number of features, reported a prediction accuracy of around 80% [49]. These results are in good agreement with our own results and emphasize the relevance of eye-tracking patterns in the classification of testing.

However, the moderate variance in the F1 score indicates that predictability varies between tasks, which could be due to differences in data quality, difficulty, or class separability. These findings support hypothesis 1 and replicate previous work by showing that correct and incorrect solvers of TUG-K items can be distinguished using machine-learning models trained on gaze data.

To address the second research question, we investigated whether performance in subsequent test items can be predicted from gaze data on learning slides. While solving test items that include MERs and learning with MERs may engage partially overlapping cognitive processes, we tested whether these similarities are sufficient for predictive modeling. Our results show that—similar to testing—certain features show significant differences between correct and incorrect performers. These differences are comparable to the results in the testing context, where relevant AOIs were associated with task solution. When analyzing feature importance, it is particularly noticeable that the

feature *formula* showed the highest discriminative power on the example learning slide. This indicates that visual attention on the formula area contributed most strongly to distinguishing between performance groups. However, elevated attention to formulas can reflect different cognitive states (conceptual engagement vs confusion/superficial checking), and our current design cannot disambiguate these. Therefore, we do not draw any conclusions at this point regarding its relevance for actual task-solving success. Neither the direction of this effect nor the underlying cognitive mechanism (e.g., conceptual engagement vs superficial attention) can be inferred from the current data. Clarifying this relationship requires further analysis and is therefore noted as a limitation of the present study.

However, aggregating the results across learning slides shows that the model cannot classify better than chance up to a comparatively high training size. Compared to the prediction of task-solving performance on test items based on gaze data collected during testing, the averaged F1 and MCC values for learning are consistently lower, although it was within the standard deviation. For instance, at a training size of 0.9, the F1 score for learning reaches 0.733 ± 0.170 , compared to 0.776 ± 0.081 for testing. Similarly, the MCC for learning at 0.9 is 0.577 ± 0.375 , while it is 0.498 ± 0.205 for testing. This suggests that gaze-based prediction tends to be more challenging for learning slides than for testing slides. The standard deviations for the learning phase are notably higher across all training sizes, highlighting greater variability in prediction quality between learning slides. In contrast, testing exhibits relatively smaller standard deviations, suggesting more consistent predictive performance across items. These differences indicate that testing may involve more clearly defined and consistent strategies, which are easier to map in gaze data, because the relevant information is spatially separated. In learning, however, the cognitive strategies appear to be more variable, which likely results in a greater variance in gaze metrics and makes prediction more challenging. Even though we explicitly designed the learning slides to align with the selected items, complete cognitive alignment between learning and problem-solving processes cannot be guaranteed; for example, part of the predictive signal may also reflect prior knowledge or existing representational and/or graph competencies.

Another difference between the two contexts lies in the discriminative features available, which may be less distinct for learning compared to task-solving. In testing, there are task-specific features, such as response options, that can clearly distinguish between correct and incorrect performers. Studies have shown that the chosen answer often receives increased visual attention, which is reflected in AOI features [50]. However, such specific AOIs are absent in the learning context, which could partly explain the lower prediction performance. Therefore, our results partially support hypothesis 2, indicating that predicting task

performance based on gaze data on learning material can be possible if the correct features are used.

The results of our exploratory analysis (RQ3) emphasize the central role of graph features in both contexts, which points to the universal importance of visual representations for learning and task solving. At the same time, the stronger weighting of text features in the learning context shows that textual information plays a greater role in the acquisition of new concepts. The dominance of dwell time features over transition features also makes it clear that the time spent on specific AOIs is a critical predictor of model performance.

B. Limitations and future research

Our study investigates the gaze-based prediction of task-solving success on specific TUG-K test items during testing and learning in the field of kinematics. However, there are some limitations and potential directions for future research.

A central limitation lies in the fact that the study focuses exclusively on the topic of kinematics. While we were able to draw on detailed data and domain-specific models here, the transferability to other physics topics is not trivial. Different subject areas could require completely different gaze patterns and cognitive strategies. For example, in the field of electrodynamics, other visually relevant areas, such as field representations or circuits, could play a central role. In contrast, in kinematics, the handling of graphs, the reading of slope values, or the visual formation of a slope triangle are crucial. These specific visual requirements of kinematics can be captured well by eye-tracking data, which may be less the case for other topics.

Moreover, our models were only trained on items with relatively balanced success rates. Although this enhances interpretability and reduces the risk of overestimating predictive performance under severe class imbalance, it leads to a selection of subtasks and thus to a possible systematic bias, as particularly difficult items, i.e., those that frequently occur in real learning environments, are underrepresented. Future studies should therefore use techniques for dealing with highly unbalanced data or include the entire range of tasks and adjust the selection criteria accordingly, e.g., via cost-sensitive learning and data augmentation techniques.

In addition, by removing the textual multiple-choice options on the TUG-K test items, we were unable to analyze classic misconceptions (distractors), which limits the targeted derivation and use of misconceptions. Furthermore, our analysis is based on aggregation across items, which means that intraitem effects are lost, such as why formally similar subtasks are solved differently.

When looking at the feature importances, causality cannot be derived from the weight-based feature importances only. Thus, longer fixation times on an equation do not necessarily mean better understanding but can also indicate difficulties. Experimental designs—like manipulation of the AOIs—are

therefore recommended for robust conclusions, causal modeling approaches, or explainable AI techniques such as SHAP values.

Another aspect concerns the sample size. Our study is based on data from a total of 88 participants, which is rather small in the field of machine learning. Smaller samples can limit the generalizability of the models, especially if the data are not sufficiently representative. Nevertheless, our study shows that predictive results can also be achieved with smaller datasets. This is in line with other research in the field of eye tracking, which has shown that reliable predictions are possible even with limited sample sizes [33]. Our results confirm this observation and emphasize that precisely designed studies and well-chosen features can play a central role in model performance. Methods such as data augmentation could help future research to further improve the robustness and generalizability of the models.

Finally, we also emphasize the temporal dimension of the prediction. Our analysis so far only considers static or aggregated gaze patterns. However, explicit consideration of the temporal dimension could provide deeper insights into learning and testing processes and improve the accuracy of the models. This approach would also have relevance for the development of adaptive learning systems that can respond to individual progress in real time.

VI. CONCLUSION

Our results show that the performance prediction of eye-tracking data for subsequent test items from the TUG-K during learning is, on average, lower than during the testing phase. This could be due to the fact that test items require more specific strategies that clearly distinguish between successful and less successful learners. In task solving, specific cognitive processes are activated that are more strongly reflected in the gaze patterns and therefore enable a higher predictive power.

Gaze patterns that occur during testing phases appear to be closely linked to the ability to understand and apply physical concepts. Adaptive Systems could utilize this to specifically identify learning gaps and tailor interventions accordingly.

In contrast, the lower predictivity during learning could indicate that learning processes involve less clearly structured strategies that can be directly transferred to conceptual tasks.

To improve prediction quality already in the learning phase, future work should explore more comprehensive temporal modeling (e.g., sequence models), combine gaze

with complementary process data such as keystrokes, and test item-dependent feature sets that capture visual strategies occurring while a concept is first encountered. Adaptive learning systems should furthermore consider supplementary methods such as feedback mechanisms, additional diagnostic tasks, or short tasks between learning phases. Such smaller tasks can help to increase understanding, provide insights into learning progress, and adapt the system's predictions to the current state of learning.

Our third exploratory analysis provides insights into the relevance of specific feature categories for predicting the performance in test items and learning scenarios. This may be relevant for the development of adaptive learning systems, where real-time analysis is required. To achieve a balance between computational efficiency and predictive performance, identifying the most informative features is essential.

The results suggest that features related to dwell time, especially those related to graphs, contribute most strongly to the model's classification performance in both contexts. Graph-related features appear to be particularly effective in distinguishing between high and low performers and should therefore be considered in the development of real-time prediction systems. At the same time, we found that transitions play a less important role in prediction in total. Even if they occur much more frequently in terms of number, they contribute less to the predictive power in both contexts. These insights are useful for reducing computational complexity in real-time applications without a major loss in accuracy.

These findings provide a basis for designing efficient and accurate adaptive, gaze-sensitive learning environments that focus on the most relevant features.

ACKNOWLEDGMENTS

This work was supported by LMUexcellent, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder. We thank Tobii for providing the eye-tracking hardware used in this study.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available because of legal restrictions preventing unrestricted public distribution. The data are available from the authors upon reasonable request.

- [1] V. Aleven, I. Roll, B. McLaren, and K. Koedinger, Help helps, but only so much: Research on help seeking with intelligent tutoring systems, *Int. J. Artif. Intell. Educ.* **26**, 1 (2016).
- [2] B. Jiang, M. Gu, and Y. Du, Recent advances in intelligent textbooks for better learning, in *AI in Learning: Designing the Future*, edited by H. Niemi, R. D. Pea, and Y. Lu (Springer, Cham, 2023), [10.1007/978-3-031-09687-7_15](#).
- [3] S. Järvelä, A. Nguyen, and A. Hadwin, Human and artificial intelligence collaboration for socially shared regulation in learning, *Br. J. Educ. Technol.* **54**, 1057 (2023).
- [4] Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediya Daskalova, Jeff Huang, and James Hays, WebGazer: Scalable webcam eye tracking using user interactions, in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)* (AAAI Press, New York, 2016), pp. 3839–3845.
- [5] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba, Eye tracking for everyone, in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV* (IEEE, New York, 2016).
- [6] Miriam Navarro, Álvaro Becerra, Roberto Daza, Ruth Cobos, Aythami Morales, and Julian Fierrez, VAAD: Visual attention analysis dashboard applied to e-Learning, in *Proceedings of the 2024 International Symposium on Computers in Education, SIIE, A Coruña, Spain* (IEEE, New York, 2024).
- [7] A. R. Strohmaier, K. J. MacKay, A. Obersteiner, and K. M. Reiss, Eye-tracking methodology in mathematics education research: A systematic literature review, *Educ. Stud. Math.* **104**, 147 (2020).
- [8] E. Alemdag and K. Cagiltay, A systematic review of eye tracking research on multimedia learning, *Comput. Educ.* **125**, 413 (2018).
- [9] K. Scheiter, C. Schubert, A. Schüler, H. Schmidt, G. Zimmermann, B. Wassermann, and T. Eder, Adaptive multimedia: Using gaze-contingent instructional guidance to provide personalized processing support, *Comput. Educ.* **139**, 31 (2019).
- [10] S. Becker, S. Küchemann, P. Klein, A. Lichtenberger, and J. Kuhn, Gaze patterns enhance response prediction: More than correct or incorrect, *Phys. Rev. Phys. Educ. Res.* **18**, 020107 (2022).
- [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 the 12th International Conference on Computer Supported Education—Volume 1: CSEDU* (SciTePress, 2020), pp. 36–46, ISBN 978-989-758-417-6, [10.5220/0009359400360046](#).
- [12] V. Ruf, A. Horrer, M. Berndt, S. I. Hofer, F. Fischer, M. R. Fischer, J. M. Zottmann, J. Kuhn, and S. Küchemann, A literature review comparing experts' and non-experts' visual processing of graphs during problem-solving and learning, *Educ. Sci.* **13**, 216 (2023).
- [13] D. Dzsofjan, K. Ludwig-Petsch, S. Mukhametov, S. Ishimaru, S. Kuechemann, and J. Kuhn, *The predictive power of eye-tracking data in an interactive AR learning environment*, in *Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the Proceedings of the 2021 ACM International Symposium on Wearable Computers* (Association for Computing Machinery, New York, NY, 2021), pp. 467–471, [10.1145/3460418.3479358](#).
- [14] 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).
- [15] W. M. Roth and K. Tobin, Cascades of inscriptions and the re-representation of nature: How numbers, tables, graphs, and money come to re-present a rolling ball, *Int. J. Sci. Educ.* **19**, 1075 (1997).
- [16] S. Ainsworth, DeFT: A conceptual framework for considering learning with multiple representations, *Learn. Instr.* **16**, 183 (2006).
- [17] K. R. Butcher, The multimedia principle, in *The Cambridge Handbook of Multimedia Learning* (2014), pp. 174–205, [10.1017/cbo9781139547369.010](#).
- [18] S. Ainsworth and A. T. Loizou, The effects of self-explaining when learning with text or diagrams, *Cognit. Sci.* **27**, 669 (2003).
- [19] W. Schnotz, Integrated model of text and picture comprehension, in *The Cambridge Handbook of Multimedia Learning*, 2nd ed., edited by R. E. Mayer (Cambridge University Press, Cambridge, England, 2014), pp. 72–103, [10.1017/CBO9781139547369.006](#).
- [20] S. Ainsworth, The multiple representation principle in multimedia learning, in *The Cambridge Handbook of Multimedia Learning*, 2nd ed., edited by R. E. Mayer (Cambridge University Press, Cambridge, England, 2014), pp. 464–486, [10.1017/CBO9781139547369.024](#).
- [21] E. Rexigel, J. Kuhn, S. Becker, and S. Malone, The more the better? A systematic review and meta-analysis of the benefits of more than two external representations in STEM education, *Educ. Psychol. Rev.* **36**, 124 (2024).
- [22] S. Malone, K. Altmeyer, M. Vogel, and R. Brünken, Homogeneous and heterogeneous multiple representations in equation-solving problems: An eye-tracking study, *J. Comput. Assist. Learn.* **36**, 781 (2020).
- [23] N. Ott, R. Brünken, M. Vogel, and S. Malone, Multiple symbolic representations: The combination of formula and text supports problem solving in the mathematical field of propositional logic, *Learn. Instr.* **58**, 88 (2018).
- [24] Marlit Annalena Lindner, Alexander Eitel, Benjamin Strobel, and Olaf Köller, Identifying processes underlying the multimedia effect in testing: An eye-movement analysis, *Learn. Instr.* **47**, 91 (2017).
- [25] R. E. Mayer, Cognitive theory of multimedia learning, in *The Cambridge Handbook of Multimedia Learning* (2014), pp. 43–71, [10.1017/cbo9781139547369.005](#).
- [26] M. A. Lindner, A. Eitel, J. Barenthien, and O. Köller, An integrative study on learning and testing with multimedia: Effects on students' performance and metacognition, *Learn. Instr.* **71**, 101100 (2021).
- [27] S. Malone, K. Altmeyer, M. Vogel, and R. Brünken, Homogeneous and heterogeneous multiple representations

- in equation-solving problems: An eye-tracking study, *J. Comput. Assist. Learn.* **36**, 781 (2020).
- [28] M. A. Lindner, A. Eitel, B. Strobel, and O. Köller, Identifying processes underlying the multimedia effect in testing: An eye-movement analysis, *Learn. Instr.* **47**, 91 (2017).
- [29] 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).
- [30] L. Hahn and P. Klein, Eye tracking in physics education research: A systematic literature review, *Phys. Rev. Phys. Educ. Res.* **18**, 013102 (2022).
- [31] H. Haider and P. A. Frensch, Eye movement during skill acquisition: More evidence for the information-reduction hypothesis, *J. Exp. Psychol. Learn. Mem. Cogn.* **25**, 172 (1999).
- [32] Q. Wang, Y. Zhu, L. Wei, and H. Deng, Visual attention pattern of middle school students during problem-solving in physics, *Mind Brain Educ.* **16**, 99 (2022).
- [33] V. Ruf, Y. Dinc, S. Küchemann, M. Berndt, S. Steinert, D. Kugelmann, and J. Kuhn, Comparison and AI-based prediction of graph comprehension skills based on the visual strategies of first-year physics and medicine students, *Phys. Rev. Phys. Educ. Res.* **20**, 020138 (2024).
- [34] R. J. Beichner, Testing student interpretation of kinematics graphs, *Am. J. Phys.* **62**, 750 (1994).
- [35] Géron, *Hands-On Machine Learning with SCIKIT-LEARN, Keras, and TensorFlow*, 2nd ed. (O'Reilly Media Inc., Canada, 2019).
- [36] S. Mozaffari, P. Klein, M. Al-Naser, S. Küchemann, J. Kuhn, T. Widmann, and A. Dengel, Quantifying gaze-based strategic patterns in physics vector field divergence, in *Proceedings of the Agents and Artificial Intelligence: 12th International Conference, ICAART 2020, Valletta, Malta, February 22–24, 2020, Revised Selected Papers 12* (Springer International Publishing, New York, 2021), pp. 465–481.
- [37] D. Ifenthaler and C. Widanapathirana, Development and validation of a learning analytics framework: Two case studies using support vector machines, *Technol. Knowl. Learn.* **19**, 221 (2014).
- [38] C. Verma, V. Stoffová, and Z. Illés, Prediction of students' awareness level towards ICT and mobile technology in Indian and Hungarian University for the real-time: Preliminary results, *Heliyon* **5**, e01806 (2019).
- [39] T. Kabudi, I. Pappas, and D. H. Olsen, AI-enabled adaptive learning systems: A systematic mapping of the literature, *Comput. Educ.* **2**, 100017 (2021).
- [40] A. Namoun and A. Alshantiti, Predicting student performance using data mining and learning analytics techniques: A systematic literature review, *Appl. Sci.* **11**, 237 (2020).
- [41] Tobii Pro Spark, eye tracker, <https://www.tobii.com/products/eye-trackers/screen-based/tobii-pro-spark> [accessed December 7, 2024].
- [42] D. D. Salvucci and J. H. Goldberg, Identifying fixations and saccades in eye-tracking protocols, in *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (Association for Computing Machinery, New York, NY, 2000), pp. 71–78.
- [43] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/4yys-22tc> for learning material (German).
- [44] 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).
- [45] K. Wada, Labelme: Image polygonal annotation with Python [Computer software], [10.5281/zenodo.5711226](https://doi.org/10.5281/zenodo.5711226).
- [46] S. Steinert, V. Ruf, D. Dzsojtjan, N. Großmann, A. Schmidt, J. Kuhn, S. Küchemann, and S. Veerappampalayam Easwaramoorthy, A refined approach for evaluating small datasets via binary classification using machine learning, *PLoS One* **19**, e0301276 (2024).
- [47] Kathrin Kennel, Shoya Ishimaru, Stefan Küchemann, Steffen Steinert, Jochen Kuhn, and Stefan Ruzika, Gaze-based prediction of students' math difficulties—A time dynamic machine learning approach to enable early individual assistance, *Int. J. Artif. Intell. Educ.* (2025), [10.1007/s40593-024-00447-5](https://doi.org/10.1007/s40593-024-00447-5).
- [48] Saleh Mozafari, Mohammad Al-Naser, Pascal Klein, Stefan Küchemann, Jochen Kuhn, Thomas Widmann, and Andreas Dengel, *Classification of Visual Strategies in Physics Vector Field Problem-Solving* (2020), pp. 257–267, [10.5220/0009173902570267](https://doi.org/10.5220/0009173902570267).
- [49] N. Rebello, M. Nguyen, Y. Wang, T. Zu, J. Hutson, and L. Loschky, Machine learning predicts responses to conceptual tasks using eye movements, presented at PER Conf. 2018, Washington, DC, [10.1119/perc.2018.pr.Rebello](https://doi.org/10.1119/perc.2018.pr.Rebello).
- [50] M. Sajka and R. Rosiek, Analysis of aspects of visual attention when solving multiple-choice science problems, in *Applying Bio-Measurements Methodologies in Science Education Research*, edited by I. Devetak and S. A. Glažar (2021), pp. 185–215, [10.1007/978-3-030-71535-9_10](https://doi.org/10.1007/978-3-030-71535-9_10).