International Journal of Artificial Intelligence in Education (2024) 34:931–973
https://doi.org/10.1007/s40593-023-00368-9

ARTICLE

Automated Gaze-Based Identification of Students' Strategies in Histogram Tasks through an Interpretable Mathematical Model and a Machine Learning Algorithm

Lonneke Boels[1,2] · Enrique Garcia Moreno-Esteva[3] · Arthur Bakker[1] · Paul Drijvers[1]

Accepted: 8 August 2023 / Published online: 22 September 2023

Abstract

As a first step toward automatic feedback based on students' strategies for solving histogram tasks we investigated how strategy recognition can be automated based on students' gazes. A previous study showed how students' task-specific strategies can be inferred from their gazes. The research question addressed in the present article is how data science tools (interpretable mathematical models and machine learning analyses) can be used to automatically identify students' task-specific strategies from students' gazes on single histograms. We report on a study of cognitive behavior that uses data science methods to analyze its data. The study consisted of three phases: (1) using a supervised machine learning algorithm (MLA) that provided a baseline for the next step, (2) designing an interpretable mathematical model (IMM), and (3) comparing the results. For the first phase, we used random forest as a classification method implemented in a software package (Wolfram Research Mathematica, 'Classify Function') that automates many aspects of the data handling, including creating features and initially choosing the MLA for this classification. The results of the random forests (1) provided a baseline to which we compared the results of our IMM (2). The previous study revealed that students' horizontal or vertical gaze patterns on the graph area were indicative of most students' strategies on single histograms. The IMM captures these in a model. The MLA (1) performed well but is a black box. The IMM (2) is transparent, performed well, and is theoretically meaningful. The comparison (3) showed that the MLA and IMM identified the same task-solving strategies. The results allow for the future design of teacher dashboards that report which students use what strategy, or for immediate, personalized feedback during online learning, homework, or massive open online courses (MOOCs) through measuring eye movements, for example, with a webcam.

✉ Lonneke Boels
l.b.m.m.boels@uu.nl

[1] Freudenthal Institute, Utrecht University, PO Box 85.170, 3508 AD Utrecht, the Netherlands

[2] Present Address: University of Applied Sciences Utrecht, PO Box 14007, 3508 SB Utrecht, Netherlands

[3] Faculty of Educational Sciences, University of Helsinki, PO Box 9, 00014 Helsinki, Finland

Keywords Eye-tracking · Computer in education · Histograms · Mathematica Classify Function · Problem-solving strategy · Graphs

The Challenge of Gaze-Based Strategy Identification in Statistics Education

Imagine students learning statistics on a laptop. They interpret histograms to solve a task. Assume that the webcam camera is good enough to track their eye movements when thinking about the task. It is imaginable that the eye movements of *all* students can be automatically recorded online[1] in the near future to document students' strategies. With available techniques, it is in principle possible to give automated feedback on students' task-specific solution strategies. In this imaginary situation, feedback on students' strategies can even be given before students answer. We see our article as a first step toward using gaze data as a source in, for example, an intelligent tutoring system (ITS) in *statistics education*.

Although techniques are available, there are still several challenges regarding the use of gaze data as a source in statistics education, before an ITS can be considered. The first is the availability of gaze data, as the use of eye tracking in statistics education is rare (e.g., Strohmaier et al., 2020). Recent reviews of eye-tracking studies in mathematics education found only four studies in statistics education (one out of 33 included studies, Lilienthal & Schindler, 2019; three out of 161, Strohmaier et al., 2020).

The second challenge is that the current usage of gaze data often addresses general pedagogical themes (e.g., metacognitive skills; Lai et al., 2013) instead of task-specific strategies teachers in statistics education are interested in (sometimes called didactics or domain-specific pedagogy). Most studies investigating students' strategies look at general strategies including planning and evaluation (e.g., Eivazi & Bednarik, 2010) or global scanning followed by local viewing (Van der Gijp et al., 2017). Other studies look at cognitive models such as visual working memory (e.g., Epelboim & Suppes, 2001). The number of studies that uncover task-specific strategies in mathematics in primary and secondary education (e.g., Lilienthal & Schindler, 2019; Strohmaier et al., 2020) and science (e.g., Garcia Moreno-Esteva et al. 2020; Klein et al., 2021; Kragten et al., 2015) is still relatively small but growing. For example, patterns in students' gazes indicating strategies have already been found in mathematical domains such as numbers (Schindler et al., 2021), arithmetic (Green et al., 2007), fractions (Obersteiner & Tumpek, 2016), proportional reasoning (Shayan et al., 2017), area and perimeter (Shvarts, 2017), Cartesian coordinates (Chumachemko et al., 2014), geometry (Schindler & Lilienthal, 2019), trigonometry (Alberto et al., 2019), and functions (e.g., parabola; Shvarts & Abrahamson, 2019). For mathematics and statistics teachers, such strategies are important as they can reveal students' knowledge of and deficiencies in this specific topic (cf. Gal, 1995).

Third, a challenge is that automation of strategy identification by using interpretable models or machine learning techniques in combination with gaze data, is even rarer in statistics education: Only one of the four studies in the previously mentioned review

[1] For most people, online means on a website. In eye-tracking research, however, online often refers to: real-time, hence, during the task solving process of the student. Here we refer to both meanings of online.

studies used a machine learning approach (Garcia Moreno-Esteva et al., 2016). None of these studies used an interpretable mathematical model. Our present study aims to address this third challenge by investigating how these two data science tools—an interpretable mathematical model and machine learning algorithms—can be used to automatically identify students' strategies on histograms based on gaze data.

Fourth, although the use of gaze data in ITSs is not new, the majority of studies on ITSs that use gaze data seem to focus on general skills such as engagement (e.g., D'Mello et al., 2012). This is in line with a review of research articles on artificial intelligence in education (AIED) in which an independent cluster of recent eye-tracking articles emerged that "include 'collaborative learning', 'engagement', 'video-based learning', and 'recommender system'" (Feng & Law, 2021, p. 293).

Fifth, many ITSs in mathematics and statistics education seem to focus on *procedural* knowledge—problems that can be solved by following a stepwise solving procedure such as solving a linear equation—although ITSs that focus on students' task-specific *strategies* do exist, also in statistics education (e.g., Tacoma et al., 2019). To the best of our knowledge, none of these seem to use gaze data as a source.

That said, ITSs that use gaze data for identifying visual-based task-specific strategies, as far as we are aware, do not exist yet in *statistics education*. Before such a gaze-based ITS can be considered and developed, we not only need to be able to link students' mathematical task-solving strategies to specific gaze patterns but also to automate the identification (or classification, as data scientists would say) of such strategies. In our previous, qualitative study, we inferred students' strategies from their gaze data. In the current study, we concentrate on automatization through the research question: *How can gaze data be used to automatically identify students' task-specific strategies on single histograms?*

The potential of automated identification of such strategies is to make large-scale, personalized feedback possible for online learning both in the initial stages of learning and during expertise development (Ashraf et al., 2018; Brunyé et al., 2019; Jarodzka et al., 2017; Hwang & Tu, 2021). This can make feedback in online courses or during homework more efficient and more accurate.

This article aims to show how the *identification* of students' task-specific strategies on histograms can be automated. We expect that this work can nurture the dialogue between experts in the field of data science algorithms—more specifically experts regarding interpretable mathematical models (IMM) and machine learning algorithms (MLAs)—and educational researchers. IMM and MLA experts may be more interested in how the IMM or MLA was or could be tailored to the specific application. Educational researchers may be more interested in using an MLA as it is, as a black box, and wonder what it provides them and how well it works. The advantage of an IMM for educational researchers is that it is transparent in how it exactly came to its decisions for individuals. We think this article can fuel the dialogue between IMM and MLA experts and educational researchers to keep the boundaries between disciplines permeable. At such boundaries, exciting new research can emerge.

In this article, we developed an interpretable mathematical model (IMM) and compared its results with a machine learning algorithm (MLA). We used these two methods from data science along with theories and insights from psychology research and neurosciences (e.g., on eye tracking, what gaze data can and cannot tell us, and the sensorimotor system), theories and insights from mathematics and statistics education research (e.g., on averages and histograms), and bring this to the world of human–computer interaction (in which, for example, the usability of an IMM or MLA is important). This means that we sometimes need to bridge worlds in terms of terminology, expectations, and explanations.

Our study is in line with the call for research focusing on methods for using measures of micro-level learning processes—including gaze data (Harteis et al., 2018). For the specific topic of histograms, our study also provides the level of detail that Peebles and Cheng (2001) referred to: "From […] eye-movement studies it is argued that there is a missing level of detail in current task analytic models of graph-based reasoning." (p. 1069). Yuan et al. (2019) showed that there is a need for searching for "visual cues that mediate the patterns that we can see in data, across visualization types and tasks" (p. 1).

Theoretical Background of the Tasks and the Use of Gaze Data

Estimating the Arithmetic Mean from Histograms

Developing students' statistical literacy, reasoning, and thinking is an important goal of education (Ben-Zvi et al., 2017). Statistical literacy is especially important in the world of "big data" and alternative truths (Burrill, 2020). Most adults will be data consumers, making decisions based on data collected by others (Gal, 2002). Statistical data in tables are not always clear. Messages can be clearer if these data are presented in more aggregated forms in graphical representations—including dotplots, boxplots, and histograms—that stress some aspects of the data (e.g., variability) and leave out other information (e.g., the exact measurements). Students, however, find it difficult to correctly interpret histograms.

A review of students misinterpreting histograms revealed that many of their difficulties stem from not understanding the statistical key concept of data (Boels et al., 2019a, b, c). The key concept of data includes an understanding of what, how many, and how variables and their values are depicted in a histogram. Despite many carefully designed interventions to tackle misinterpretations (e.g., Kaplan et al., 2014), students' difficulties with histograms remain (e.g., Cooper, 2018). We, therefore, decided to use eye tracking to study in depth how students interpret histograms (Boels et al., 2018, 2019a, 2022a, b).

Strengths and caveats in students' knowledge can be revealed by asking them to estimate averages from data in different representations (e.g., histogram, dotplot, case-value plot; cf. Gal, 1995). Estimating the mean can be seen as a prerequisite for assessing variability, as the variation in data is compared to a measure of center (e.g., standard deviation from the mean). Furthermore, our students are familiar with the mean, but not so much with variability. Therefore, in a previous eye-tracking study, students were asked to estimate the mean from various—but univariate—statistical graphs in 25 items (e.g., Boels et al., 2022a, b). In the present article, we re-use gaze data from a subset of this previous study containing all five single histogram items.

Historical examples show that the mean has emerged from estimating representative values for a dataset through compensation and balance (Bakker & Gravemeijer, 2006). Students exhibit minimal difficulty in estimating the mean from case-value plots (Cai et al., 1999), unless zero is one of the measured values (Boels et al., 2022a, b). Most students know how to calculate the arithmetic mean from raw data (e.g., Konold & Pollatsek, 2004). In a study with various items—including finding the "average" allowance from a histogram—five approaches were found for solving the items: average as (1) mode, (2) algorithm, (3) reasonable, (4) midpoint, and (5) balancing point (Mokros & Russell, 1995). Students often (implicitly) use the mean of frequencies in a histogram (cf. Cooper, 2018). The latter is incorrect when applied to histograms but correct for finding the mean from a case-value plot, and can be seen as finding the horizontal line that makes all bars of equal height by using compensation.

The weighted estimation of the mean in a histogram is the balance or gravity point of the graph (e.g., Mokros & Russell, 1995). This mean can be found by taking the range or spread of the data in the histogram into account together with the height of the bars. For this approach, it is not necessary to read off frequencies on the vertical axis. We call this approach a histogram (interpretation) strategy or correct strategy. An estimation of the mean in a histogram *with equal bin widths* can also be computed by multiplying the frequency or percentage (height of the bar) with the middle value of that bar, adding the results over all bars, and dividing this by the *sum* of the frequencies. No students in the previous study used this approach. Instead, all that used a computational approach added all frequencies and divided this sum by the number of bars. This would be a correct strategy if the height of each bar was representing weight and the number of bars was the number of measured weights (as in a case-value plot). Therefore, this count-and-compute strategy is incorrect for histograms.

In our previous study, we found several strategies for estimating the mean from a histogram based on students' visual search strategies (cf. Goldberg & Helfman, 2011) inferred from their gaze patterns (see the Empirical Background of the Re-Used Data section). A visual search strategy can be part of a task-specific strategy. People use these strategies to get "from an initial problem state to a desired goal state, without knowing exactly what actions are required to get there (Newell & Simon, 1972)" (Van Gog et al., 2005, p. 237). As the debate between Lawson (1990) and Sweller (1990) illustrates, there are different opinions on what strategies are. In our study on graph interpretation, students' strategies typically consist of (1) visually searching for the relevant information, (2) making inferences based on this information, and in some cases (3) verifying the inference; see also the section Theoretical Interpretation of Students' Gaze Patterns.

Given our focus on what eye tracking and data science tools (IMM, MLA) can provide to educational researchers, we now first discuss the theoretical background of using eye tracking. In the Research Approach section, we elaborate on the data science tools used (IMM and MLA).

Use of Gaze Data

There are multiple reasons for using gaze data to identify students' strategies. First, eye-movement patterns (e.g., order of fixations[2] or saccades) are online, real-time measures that may allow for more adequate feedback than feedback on answers only. Moreover, feedback on strategies can be provided earlier during the problem-solving process (e.g., Gerard et al., 2015; Mitev et al., 2018), although strategy feedback can also be on answers (e.g., Tacoma et al., 2019). Second, low-accuracy eye tracking—for example, through webcams—is expected to be a standard option for computers in several years (e.g., Kok & Knoop-Van Campen, 2022), which would make it possible to give feedback to large groups of students. Third, gaze data are direct motor data that are almost impossible to manipulate. This makes measuring eye movements more reliable than, for example, thinking-aloud protocols (e.g., Van Gog et al., 2005). In addition, younger students, novices, and sometimes even experts find it difficult to articulate their thinking process, are sometimes not aware of their thinking (e.g., Green et al., 2007) or might respond to what they think the interviewer expects or what is easily accessible (e.g., Wilson, 1994).

The implicit assumption here is that eye movements reflect cognitive processes. Spivey and Dale (2011) state: "Our most frequent motor movements—eye movements—are sure to play an important role in our cognitive processes. […they] provide the experimenter with a special window into these cognitive processes." (p. 551). It is indeed generally assumed that gaze data can provide evidence of conceptual actions, however with some caveats (e.g., Radford, 2010). First, the relationship between eye movements and cognitive processes is not straightforward (e.g.,Kok & Jarodzka, 2017; Russo, 2010). In addition, not every eye movement is part of a student's strategy (e.g., Anderson et al., 2004; Schindler & Lilienthal, 2019). Furthermore, one could argue that students' fixations on the screen do not indicate where they looked, as people also observe through their peripheral vision (Lai et al., 2013). Nevertheless, in our items, focused vision is needed for locating detailed information (e.g., locating a bar, or reading a specific number on the horizontal axis). As the fovea has the greatest acuity (sharpness; Wade & Tatler, 2011), locating a number on an axis is only possible with foveal vision, and peripheral vision most likely guides our attention to it and to the bars (cf. Kok & Jarodzka, 2017). Therefore, we can infer that the fixation on the screen is what the student is looking at.

Choosing what eye-movement measures to use is a methodological decision. According to a review study (Lai et al., 2013), the measures used most often were temporal (e.g., total fixation duration, time to first fixation, total reading time), followed by count (e.g., fixation count). The least used measures were spatial (e.g., scanpath, fixation position, order of AOIs). Goldberg and Helfman (2010)

[2] "A fixation is a period of time during which a specific part of [we use *graph on the computer screen* here] is looked at and thereby projected to a relatively constant location on the retina. This is operationalized as a relatively still gaze position in the eye-tracker signal implemented using the [Tobii] algorithm." (Hessels et al., 2018, p.22). A period of time during which a relatively fast switch of gaze between two fixations occurs is called a saccade.

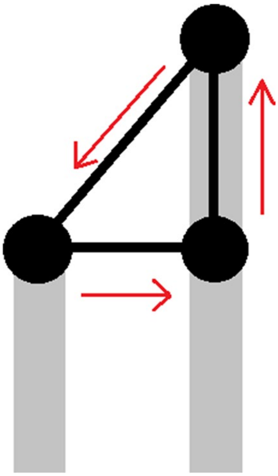

Fig. 1 Stable triangular scanpath that was interpreted as an AA. *Note.* Circles are fixations (places where students looked), arrows indicate the direction of saccades (fast transitions between two fixations), redrawn after Shayan et al. (2017, p. 175). Picture not included in the CC-license.

stated that “with appropriate task design and targeted analysis metrics, eye-tracking techniques can illuminate visual scanning patterns hidden by more traditional time and accuracy results” (p. 71). Scanpaths can reveal learning in more detail (Hyönä, 2010). Tai et al. (2006), therefore, advise using spatial measures such as scanpaths in problem-solving research.

Studies using students’ scanpaths for identifying strategies are rare, and often use the *sequence* of AOIs (e.g., Garcia Moreno-Esteva et al., 2018) or scanpaths that are aggregated over time and fixations (e.g., in heatmaps, Schindler et al., 2021). Up to now, scanpaths mostly require qualitative inspection or analysis of the eye-movement data (e.g.,Alemdag & Cagiltay, 2018; Susac et al., 2014)—especially when looking for task-specific strategies. Figure 1 provides an example of such a scanpath (a sequence of fixations and saccades). Qualitative analysis is both time-consuming and harder to objectify. In this article, we, therefore, use the raw scanpath data to identify students’ strategies.

In studies using angles and direction of saccades in educational settings (e.g., Dewhurst et al., 2018), scanpaths are often compared on multiple or all AOIs.[3] In our previous, qualitative study, we took a new approach in using the perceptual form (e.g., vertical gaze pattern) of the gazes on *one* AOI only (graph area)—the one that was found particularly relevant for students’ task-specific strategies (see also the following section, Inferring an Attentional Anchor from Gaze Data). This perceptual form consists of angles and direction of saccades that are roughly *aligned*. So far, we have not found any other study in education that uses the alignment of saccades. For more details, see the Research Approach section (Students’ Strategies). A possible advantage of looking at saccades over fixations or order of AOIs is that it may be less sensitive to spatial offsets (e.g., Jarodzka et al., 2010).

[3] In addition, this study is about the influence of task difficulty on scanpaths. It does not consider the kind of task-specific strategies we are aiming for.

Inferring an Attentional Anchor from Gaze Data

For our theoretical interpretation of the *perceptual form* of students' gaze patterns on the graph area (horizontal or vertical line segments), we draw upon insights from theories on enactivism and embodied cognition. According to these theories, cognition arises from interaction with the environment (e.g., Rowlands, 2010). The focus of an actor's interaction with this environment is called an attentional anchor (AA) (Hutto & Sánchez-García, 2015). An AA is "a real or imagined object, area, or other […] behavior of the perceptual manifold that emerges to facilitate motor-action coordination" (Abrahamson & Sánchez-García, 2016, p. 203). Other behavior of the perceptual manifold, for example, includes students gesturing a horizontal line when explaining how they made all bars equally high in a case-value plot strategy. The AAs found in our previous research (Boels et al., 2022a) facilitated students' imagined actions (strategies for finding the mean)—regardless of the strategies' correctness.

Examples of AAs in motor action can be found in research on high school trigonometry where students coordinate the movement of the left hand to describe a circle and their right hand to describe a sine graph (Alberto et al., 2019). Another example is the manipulation of two bars that are proportional to each other, see Fig. 1 (e.g., Shayan et al., 2017). Students needed to keep the bars green, which occurred when the bars had a fixed ratio of, for example, 1: 2 (unknown to the students). They dragged both bars up to find various points where both are green. Students had different strategies for finding these points. In one strategy, gaze fixation is on the right-hand bar in the middle, which is mathematically relevant, as this bar is twice as high as the left-hand bar (Fig. 1). As this imagined triangle emerges to facilitate the coordination of the motor action, it is an example of an AA.

Theoretical Interpretation of Students' Gaze Patterns

Although enactivism often assumes manipulation of an environment, there are indications that people's sensorimotor systems are also activated in situations without physical manipulation (e.g.,Fabbri et al., 2016; Lakoff & Núñez, 2000; Molenberghs et al., 2012). In the retrospective stimulated recall interviews, students talked about the graphs as if manipulation were possible. For example, student L10 refers to chopping and flattening all bars (hence, using compensation, Bakker & Gravemeijer, 2006), see the excerpts below. This student describes sensorimotor actions, namely: breaking up the longest bar into pieces that are then divided over the shorter bars, resulting in a horizontal line along the top of the now equally high bars. This would be a correct strategy for finding the arithmetical mean in a different type of graph (namely, a case-value plot, e.g.,Cai et al., 1999; Yuan et al., 2019). An imaginary horizontal line segment is used for coordinating this imagined action. Gaze data show this imaginary segment in the form of a stable scanpath indicating the focus of interaction of this student, see Fig. 2. We, therefore, interpret this segment as an AA. Gazes on Item06 indicate that this AA was also visible before Item20.

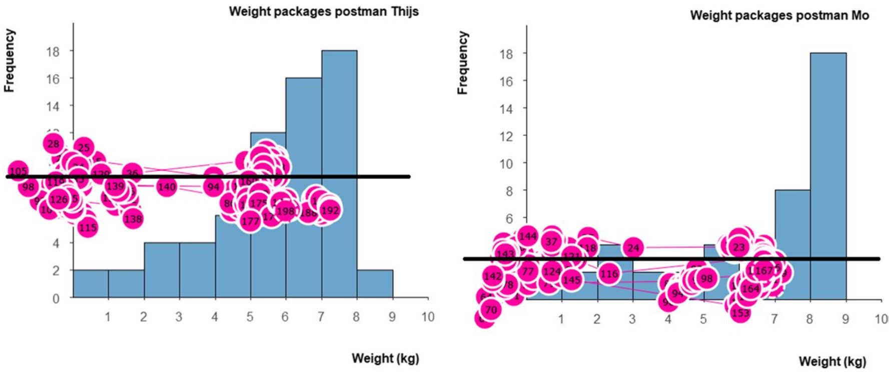

Fig. 2 Part of the stable scanpath of student L10 on Item06 (left) and Item20 (right). *Note.* The stable scanpaths reveal the horizontal line segment along which the student looks on Item06 (left; here superimposed on the figure for the reader) and Item20 (right). Circles indicate fixations and thin lines indicate saccades. As the weight value is on the horizontal axis, so is the actual mean. However, from the eye movements of this student, we can conclude that the mean of the frequency is estimated, instead of the mean weight. The interview data support this conclusion. This stable scanpath, therefore, indicates an incorrect (case-value plot interpretation) strategy

> L10: I looked at the graph itself [Item06] first and then I kind of looked at the axes, how is it constructed, and then I looked at the question and then I looked again at the frequency, how to group it. That was it in my opinion.
> R1: And were you doing that here the same way you did with those other [previous] questions? Chop it into pieces?
> L10: Yes
> […]
> R1: You said five here [Item20].
> L10:Yes, because I thought the weight would be on the left side. So, if you flattened it all out, between 4 and 6 would be the imaginary [horizontal] line.

Research Approach

In a previous study, we collected and qualitatively analyzed students' gaze and stimulated recall data and classified these into groups (Table 1) of students using the same strategy (Boels et al., 2022a). The present study consists of three phases: (1) analysis of gaze data on *one* AOI (that contains the stable scanpath) through a random forest MLA; (2) construction of a separate interpretable mathematical model (IMM) based on the gaze data and using insights from qualitative research; and (3) comparison of the results of the MLA with the IMM and with the results of the previous qualitative study, as the MLA is a baseline to compare our IMM to. The most important information of the previous study is presented in the following section. Next, the first two phases of the present study are explained in more detail. As we aim to bridge different research fields we also provide background information on the data science tools used and the considerations that have

Table 1 Strategies, percentage of trials (correct strategy in bold), $N = 50$ per item (see also Boels et al., 2022a)

Item	Histogram strategy	Case-value plot strategy	Count-and-compute strategy	Unclear
Item01	**48**%	30%	20%	2%
Item02	**46**%	38%	12%	4%
Item06	**38**%	42%	20%	0%
Item19	**44**%	42%	14%	0%
Item20	**46**%	34%	16%	4%

guided our choices, A comparison of the results is made in the section Results of Applying an MLA and IMM.

Empirical Background of the Re-Used Data

Participants

This study re-uses data from an eye-tracking study with 50 Grades 10–12 students of a Dutch public[4] city high school (Boels et al., 2022a, b). All participants are Dutch pre-university track students (15–19 years old; mean age 16.3; all with normal or correct-to-normal vision; 23 males, 27 females).

Eye-Tracking Apparatus

The gaze data that are used as input for the IMM and the MLA were collected with a Tobii XII-60 eye-tracker with a sampling rate of 60 Hz that was placed on an HP ProBook laptop between the laptop's 13-inch screen and keyboard (Fig. 3). A chin rest was used to reduce data loss and improve the accuracy of the gaze data. Furthermore, a 9-point calibration on the screen was used. The Tobii Pro Studio 3.4.5 software recorded in real-time where people looked at the screen by using harmless infrared light to detect their gaze. Data loss was minimal (7.2% on average) and none of the students (averaged over all trials[5]) or items (averaged over all students) went over the exclusion point of 34%. The mean accuracy is 56.6 pixels (1.16°) with the highest accuracy on the graph area (mean 13.4 pixels or 0.27°). The average precision (0.58°; RMS-S2S; Holmqvist et al., 2023) is considered good. For other measures of gaze data quality, see Boels et al. (2022a, b). We, therefore, did not exclude any student, although, for some specific trials, data loss could come close to or even over this exclusion point (e.g., student L39 and L32 had 27.5% and 46.0% data loss respectively on their trial of Item01). Some data loss is normal, due to blinking, wearing glasses or make-up, epicanthic eyes, or students looking above or

[4] In the Netherlands private schools are rare and in general there is no difference with state schools in terms of students' results.

[5] In mathematics education we usually talk about an item, task or problem. In eye-tracking research, a series of gazes of one student solving one such item is called a trial.

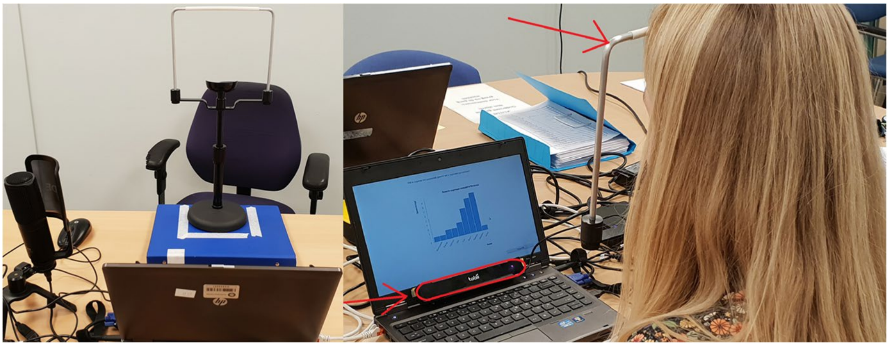

Fig. 3 Set-up of the experiment. *Note.* The red arrows in the right-hand picture point at the eye tracker (bottom left, see the red oval) and chin rest apparatus (top middle). The person in the picture is not a participant. Pictures are not included in de CC-license.

below the screen while thinking. Another reason for not excluding students is that this would not be representative of a future gaze-based feedback application where real-time data collection and processing would occur.

Tasks

The data on five histogram items are used and analyzed in the present article, see Fig. 4. The question for all five items was: "What is approximately the mean weight of the packages [name of postal worker] delivers?"

The students verbally estimated the mean (Table 2); their answer was coded as correct or incorrect. Answer correctness can differ from strategy correctness, due to, for example, underestimation of the mean, even if a correct strategy was used to locate the mean.

Students' Strategies

Qualitative data were collected through expert judgment on students' strategies on the items (Table 1), which in turn was based on (1) videos of students' gaze data on the items; (2) interview data when available; (3) students' answers. Three common strategies were identified (Table 1): a histogram strategy (Fig. 5—a correct strategy that reads off the estimation on the horizontal weight axis), a case-value plot strategy (Fig. 2—a strategy that would be correct for a case-value plot but is incorrect for finding the mean from a histogram as it returns the mean frequency, read on the vertical frequency axis), and a count-and-compute strategy (an incorrect[6] strategy that, for example, adds the height of the bars, hence the frequencies, and divides by the number of bars—resulting in a kind of zig-zag pattern of horizontal and vertical

[6] Although a correct variant of this strategy is in theory possible, we did not find such correct variant in the gaze data, nor in students' explanations.

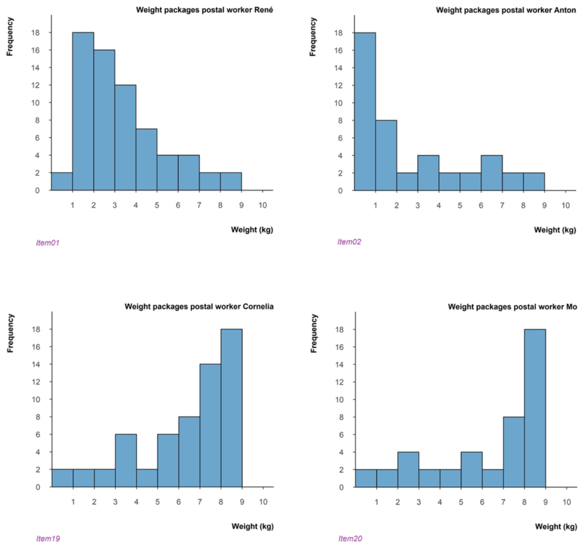

Fig. 4 Graphs (all single histograms) used in Item01, 02, 06, 19 and 20. *Note.* All items are translated into English and item numbers are added for the reader's convenience

gazes, see Boels et al., 2022a for more details). Both the case-value plot strategy and count-and-compute strategy relate to the same misinterpretation: interpreting the histogram as a case-value plot (Boels et al., 2019a; Cooper, 2018); the difference is whether students estimated (case-value plot strategy) or calculated (count-and-compute strategy) the mean. Hence, almost all strategies can be attributed to one of two classes: one in which students correctly interpreted the graph as a histogram and one in which students incorrectly interpreted the graph as a case-value plot.

Table 2 Answers given by the students, $N=50$ (see also Boels et al., 2022a)

Item	Correct answer	Answer range correct answers [a]	Average of given answers	Number of students correct	Percentage of students correct
Item01	3.3	2.2–4.4	5.6	20	40%
Item02	2.7	1.6–3.8	3.8	19	38%
Item06	5.7	4.6–6.8	6.9	25	50%
Item19	6.4	5.3–7.5	6.9	34	68%
Item20	6.3	5.2–7.4 [b]	6.7	17	34%

[a] Experts were asked to answer these items as well. Based on these results as well as students' preference for whole numbers, the answer range was set to ± 1.1 for all items

[b] If the answer 7.5 had been included, 18 students would have answered correctly. Furthermore, 10 students answered 5 for Item20

A second coder coded 10% (25 trials). The interrater reliability of the coding in Table 1 measured with a Cohen's Kappa of 0.62 is considered substantial (Landis & Koch, 1977). Four out of five disagreements involved the second coder choosing a count-and-compute strategy and the first coder choosing one of the other strategies. If this coding is aggregated to correct (histogram strategy) and incorrect (all others)—as used as input for training the machine learning algorithm—agreement goes to 22 out of 25 trials, which corresponds to a Cohen's Kappa of 0.73 (substantial).

We used a code for a correct (bold) or incorrect strategy (all other, mostly misinterpreting the histogram as a case-value plot) as input in the training phase of the MLA and compared these to the results of the IMM and of the testing phase of the MLA (we explain later how the testing was done relative to the training). The eye movements belonging to correct strategies were mainly vertical and answers were read on the horizontal axis, as the data in a histogram are positioned along this axis (Fig. 5).

In contrast to most eye-tracking studies, in the qualitative study, we looked at the *perceptual form* of the scanpath, for example, the vertical gaze pattern (Fig. 5), and refer to this as a stable scanpath if it includes multiple aligned fixations and saccades along this scanpath and was explicitly mentioned by at least some students as being relevant for their strategy (e.g., Boels et al., 2018, 2019a, 2022a, b). This vertical line is formed by looking back and forth between the balance point of the graph on the horizontal axis and the height of the bars as a weighting factor. Incorrect strategies contained mainly horizontal gaze patterns and searching for the answer on the vertical (frequency) axis—hence using incorrect data—and leveling all bars (Fig. 1 in the section on the Theoretical Interpretation of Students' Gazes).

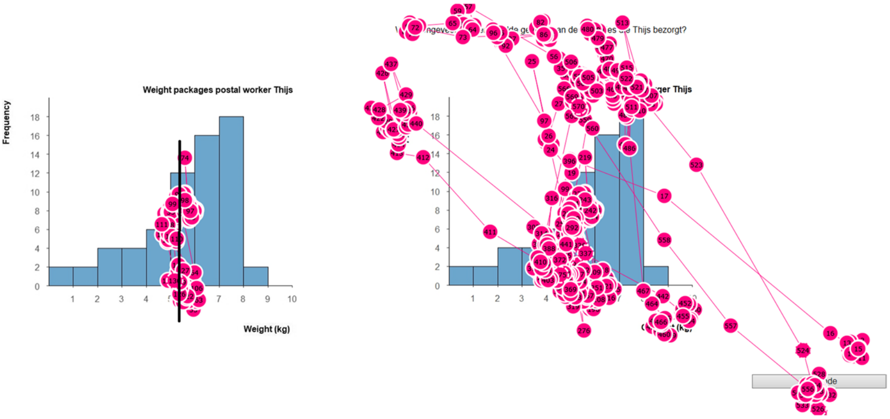

Fig. 5 Part of a stable vertical scanpath (left) and all gazes (right) on Item06. *Note.* The left figure reveals the vertical line segment (left: superimposed for the reader) in student L26's gazes on Item06 (right: all gazes). Circles indicate fixations, and thin lines indicate saccades: fast transitions between two fixations. The left figure is translated for the reader's convenience. These gazes indicate a correct (histogram interpretation) strategy

Interpretable Mathematical Model and Machine Learning Algorithm

Considerations for Analyzing Eye Movements with a Machine Learning Algorithm

The use of an MLA for analyzing gaze data is still uncommon (e.g., Kang et al., 2020) and even more unusual for educational use (e.g.,Brunyé et al., 2019; Mitev et al., 2018). For example, in a review of eye tracking in medical education, the use of an MLA is described in only two out of 33 studies (Ashraf et al., 2018). In many studies, areas of interest (AOIs) are used. AOIs are predefined areas of the item that are judged by the researchers as being distinct from each other and relevant to the strategy. Eye movements on these AOIs or between them are recorded. Typically, the information is often reduced to a single measure for each AOI—for example, whether an AOI was visited or not. In our study, we used raw gaze data on one large AOI to study the perceptual form of the scanpath. Examples of AOI usage in ML studies are the order in which AOIs are visited, or the number of fixations on the areas of interest (e.g., Garcia Moreno-Esteva et al., 2020; Najar et al., 2014). In other studies, temporal measures are used, such as the total duration for fixations on an AOI or mean duration per fixation (e.g., Voisin et al., 2013). Schindler et al. (2021) used heatmaps of students' gazes in an MLA, thus discarding the order of fixations in the scanpath pattern. However, in most eye-tracking studies, no MLA is used at all (e.g.,Strohmaier et al., 2020; Van Gog & Jarodzka, 2013). In some studies, an interpretable (vector) model was made from the raw data (e.g., Dewhurst et al., 2012). The use of multimodal data—including eye-movement data—in combination with an MLA is an emerging line in educational research (Järvelä et al., 2019). What is new in our study, to the best of our knowledge, is that we feed the MLA (and IMM) with the raw gaze data to identify students' task-specific strategies.

An advantage of supervised MLAs for analyzing gaze data is that they provide a rather generic approach. However, the disadvantage of many MLAs is that they are effectively like black boxes that do not reveal how they identify the results from the data, and, hence, what analytical model emerges (e.g.,Guidotti et al., 2018; Kuhn & Johnson, 2013; Lakkaraju et al., 2019; Rudin, 2019). Therefore, it is unknown what gaze data patterns are used by the MLA for data classification.

Several solutions are suggested in the literature to overcome this disadvantage. More transparency can be created by (1) model or global explanations, (2) outcome or local explanations, or (3) model inspection or differential explanations (e.g.,Guidotti et al., 2018; Lakkaraju et al., 2019). Explainable means that humans can understand how the MLA reached its decision (e.g., Doshi-Velez & Kim, 2017). An alternative is to create an (4) interpretable model directly from the data, sometimes after first using MLAs to understand what is relevant in the data (e.g., Rudin, 2019), or (5) use white-box techniques such as models that are made *a priori*. The disadvantage is that (5) is based on human assumptions, not on data; this risks relevant information in the eye-movement data being overlooked (e.g., Villagrá-Arnedo et al., 2017).

In our case, (1) model explanation could involve trying to extract the general rules that the MLA uses to decide what strategy a student uses (for examples from weather forecasting, see McGovern et al., 2019). Outcome explanations (2) might involve trying to extract why student A is identified as having strategy z (for clinical examples, see Krause et al., 2016; for an example with birds, see Rudin, 2019). Model inspection (3) could be understood as finding out how sensitive the model is to variations in the data.

As (2) is even more complex to achieve, we decided to make an interpretable model instead (4). An interpretable model is a model that captures the most important characteristics of the strategy, in a way that can be understood by human beings and that is transparent (Rudin, 2019). An interpretable model is often a mathematical and logical model (e.g.,Hancox-Li, 2020; Lakkaraju et al., 2019; Molnar, 2019); in our case, it consisted of a set of rules that approximately describes the stable scanpath of the gazes. We call this our interpretable mathematical model (IMM).

Considerations for the Construction of an Interpretable Mathematical Model

For automated strategy identification, an analytic model of this strategy is needed. The previous section explained why an interpretable mathematical model (IMM) seems to be a good candidate for this. In short: it is transparent in what and how characteristics of students' gaze patterns are used to identify[7] students' strategies. To detect task-specific strategies and ensure that the IMM is usable—an evaluation criterion, see the Methodological Evaluation Criteria section—we use the idea of an attentional anchor (AA) as described earlier to search for a task-specific perceptual form of the gazes (stable scanpath; see the section Theoretical Interpretation of Students' Gaze Patterns). The advantage of using an AA for constructing an IMM is that it is both task-specific (for each topic and task a different perceptual form is expected) and generalizable (it has been found already for topics and tasks in various mathematical domains).

Construction of an Interpretable Mathematical Model

In the second phase, we constructed an interpretable mathematical model (IMM) based on attentional anchors (AA) found in a qualitative analysis of students' strategies in a previous study (Boels et al., 2022a). Two AAs were found: an imaginary horizontal line and an imaginary vertical line. We tried several ways of capturing these two strategies mathematically, with varying success. The best model we found relies on the following algorithm, which is based on saccade lengths and angles. The cut-off values below were found empirically by testing many values close to a slope of 1. The starting point 1 for the slope followed from insights about the two AAs from the qualitative study. For each participant, we transform the sequence of saccades on the graph area into a sequence of -1, 0, and 1 values (Table 3):

[7] In data science, the common term is predict or classify. In educational science the term infer (students' strategies from gaze data) is also common.

Table 3 Example of applying the algorithm for the IMM

Step in algorithm	Result
1–3	{-1, -1, 0, 0, -1, 1, 1, 1}
4	{{-1, -1}, {0, 0}, {-1}, {1, 1, 1}}
5	{-1, 0, -1, 1}
6	{-1, -1, 1}
7	-1
8	0

A) If a saccade is less than 200 pixels long (Euclidean distance), map it to 0.
B) If a saccade is at least 200 pixels long and the absolute value of the slope of the saccade line is greater than or equal to 0.875 (or $^{7}/_{8}$), map it to 1 (these saccades are considered vertical).
C) If a saccade is at least 200 pixels long and the absolute value of the slope of the saccade is less than 0.875, map it to -1 (these saccades are considered horizontal).

Our algorithm continues as follows:

D) Split the -1, 0, 1 valued sequence into subsequences of identical consecutive values.
E) Delete the duplicates in each subsequence of the sequences and join the subsequences.
F) Remove the “0” cases; this is equivalent to disregarding saccades that are “short”.
G) Add up the elements of our sequence.
H) If the total is negative or 0, we replace the sequence with 0; if the total is positive, we replace the sequence with 1. This counts the number of runs of consecutive “long” horizontal or vertical scans, and based on the total, determines whether there were more horizontal or vertical sets of long scans. Replacing the sequence with a value of 0 indicates that the scanning is “mostly horizontal,” if we disregard short saccades and regard consecutive long scans with similar slopes as a single “scanning run”.

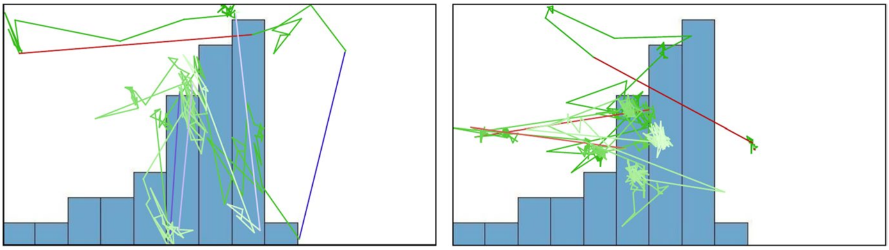

Fig. 6 Examples of horizontal and vertical eye movements that were counted in the IMM. *Note.* Red—long horizontal—gazes correspond to value 1 in step 7 of the IMM, and blue—long vertical—gazes correspond to 0 in step 7. Lighter colors indicate later occurrence. Green saccades are less than 200 pixels (hence disregarded). Left is an example of a correct strategy for Item06 (more blue vertical gazes), right is an example of an incorrect strategy (more red horizontal gazes). Readers are referred to the online enlargement of this figure for (subtle differences in) the coloring

An illustration of what we do is given in the following graphs, in which the scans that are considered horizontal are in red and the vertical ones in blue (Fig. 6, colored version online). The green saccades are shorter than 200 pixels. The color becomes lighter as the sequence progresses, to get a sense of the order in which the saccades occurred. In the left-hand graph, there are more blue lines than red ones, indicating a correct strategy for finding the mean in a histogram. In the right-hand graph, there are only red lines, so this scanning is regarded as horizontal. The 200 pixels saccade length cut-off point was fixed, thus not scaled to the width and height of the graph area (AOI). The size of the AOI in Fig. 6 is indicated with a black rectangle (not present in the item) and is the same for all histogram items. As the AAs are similar on all five items, we constructed one IMM for all items. This means that the IMM is a more general model compared to the random forest models as the latter is different for each item.

Supervised Machine Learning Algorithm

A machine learning algorithm[8] (MLA) automates building an analytical model and can provide a baseline to compare our IMM with. An MLA is a computer program that improves with experience (Kersting, 2018; Mitchell et al., 1990; Molnar, 2019). An MLA is not explicitly programmed to use any particular input features. 'Features' here refers to variables constructed from input data.

An MLA can be supervised or unsupervised. Being supervised means that the training cycle of the program is fed with, for example, the correctness of the strategy; in unsupervised learning, only the gaze data would be given to the program during the training cycle and the program might infer correctness information by itself. As we previously identified two groups in our qualitative study, we wanted to see if the MLA could identify those students correctly. That calls for a supervised MLA. During the training cycle, see Fig. 7, an MLA is fed with the raw gaze data as well as a classification code for the already identified strategy (0 for incorrect, 1 for correct). After this learning or training cycle, the trained MLA identifies the strategy of *other* students (or trials) that were not part of the training set.

To provide a baseline for our IMM, we compared it with the random forest MLA. We used Mathematica's implementation of random forest in the 'Classify Function' (version 13.2.1; WRI, 2020) with default parameters. The Classify Function automates and optimizes the data preparation process. For example, it automatically handles the different lengths of input vectors of *x*–*y*-pairs (it normalizes input features).

The Classify Function initially choose random forest as the best MLA for each of all five items in a *previous* version of the software that we started our analyses with. In the newest version (13.2.1) another MLA (logistic regression) provided slightly better results for several of our items and random forests for others. We report the results for random forests for several reasons, including consistency and that the MLA is only a baseline for our IMM. For all further analysis, we seeded this random forest to make sure that

[8] In the media, machine learning (ML) and artificial intelligence (AI) are often regarded as synonyms, but there is a difference. Interested readers are referred to Kersting (2018).

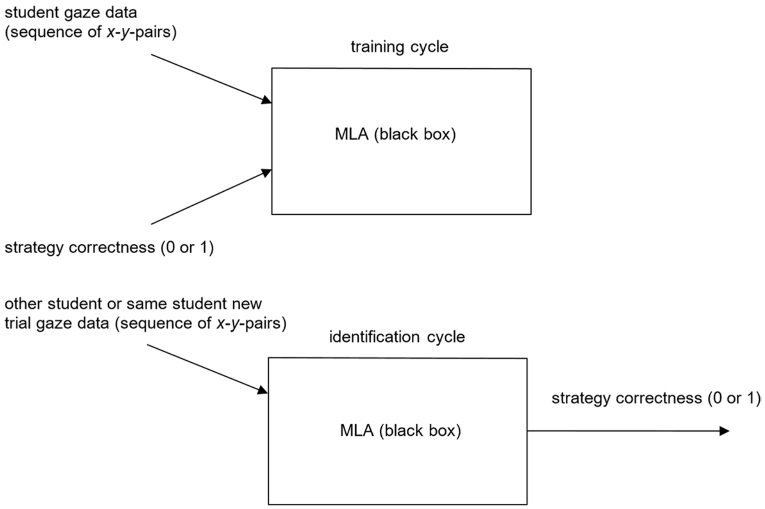
student gaze data
(sequence of x-y-pairs)
training cycle
MLA (black box)
strategy correctness (0 or 1)
other student or same student new
trial gaze data (sequence of x-y-pairs)
identification cycle
MLA (black box)
strategy correctness (0 or 1)

Fig. 7 Training and identification cycles of a supervised MLA

our results are reproducible. An advantage of the Classify Function in the Mathematica software is that users do not need to deal with the details of machine learning methods. The Classify Function is used “as is” and automates many aspects of the methodological stack. Specifically, to give an example, it automates the selection of the machine learning method and the preparation of data (consisting of only the—temporally ordered—*x*- and *y*-coordinates of fixations on the AOI graph area) so the selected method can be applied to it. Note that timestamps are not provided to the MLA, so the MLA does not ‘know’ that the data is *temporally* ordered or that data from other AOIs are removed. The downside of our approach is that we know little about how the Classify Function is handling our data. For example, data preparation and feature selection is all hidden in the software and we, therefore, consider it a black box, even though the MLA it initially selected—random forest—is itself known for its possibilities for feature extraction (e.g., through a feature importance plot). Our data, which are continuous, are by themselves difficult to interpret (*x*- and *y*-coordinates). Ultimately, the problem is that it is impossible to know whether the classification is based on gazes that are typical for the task-specific strategy at hand (cf. Kuhn & Johnson, 2013). In addition, a first step of (3) was performed by using the random forest MLA trained on one item for classifying strategies on another item and vice versa for all item-pairs.

We underline that we used the MLA as a baseline for our IMM. We see it as a tool. Educational researchers may be more interested in how well such an MLA performed and what it can provide for them. MLA experts may be more interested in the details of the ML method. It is like users of an electric screwdriver being interested in how well it works and designers of such screwdrivers being interested in the details of how this screwdriver was assembled and could be optimized and whether better ones exist. We would like to emphasize that our article is not a report about research into machine learning methods. It is a report of a study of cognitive behavior which uses machine learning tools to analyze its data. Hence, the purpose of our research was not to conduct an in-depth investigation into which machine learning methods would work best for our data but to see how well our IMM performed compared to an MLA.

An important prerequisite for ML analysis is that the dataset is large enough and contains enough information for the MLA to identify (classify) students' strategies. For this purpose, we decided to do a 'sanity check' and see if the MLA would be able to identify students' answer correctness. As 50% of the students answered Item06 correctly (Table 2), an MLA with an accuracy of about 50% would not be better than throwing a coin. Fifty percent accuracy could also be reached by identifying all students having a correct strategy. Therefore, for the prerequisite of enough information to be met, the accuracy of this answer identification needs to be well above 50% when using balanced data as in Item06. We decided to set it to 70% or above, as this is a low-stake classification problem. Such a sanity check can be used as a first step before proceeding to the manual determination of students' strategies through qualitative research, or when qualitative research is still in progress. To judge the performance of an MLA, other metrics also need to be considered; see the Methodological Evaluation Criteria section. Therefore, we first trained the MLA on students' answers (and then identified other students' answers) before we retrained the MLA on students' strategies (and then identified other students' strategies).

Random Forest Machine Learning Analysis

For the machine learning analyses, we used the random forest MLA that is implemented in the Mathematica software as described in the previous section, both for training the MLA—with a subset of students—and for the classification of the remaining students by the trained MLA. The aim of our ML analyses was to set a baseline for an IMM. We started our analysis with Item06 from the original 25 items in the qualitative study (e.g., Boels et al., 2022a, b). The first reason for choosing this item is that 50% of the students answered this item correctly (Table 2) which is—in theory—an ideal situation for MLA. If it did not work with this item, we would not expect the MLA to work on other items. The second reason was that we expected students to have settled on a strategy by the sixth item of the original study—in line with our observations from the qualitative study—possibly making the identification of strategies easier than for the first two items. The third reason was that this graph is skewed to the left, making large horizontal eye movements—part of the incorrect strategy—more likely to be explicit than in graphs that are skewed to the right (e.g., Fig. 8).

In the present study, we used a supervised MLA. The MLA is fed with the raw data of the gazes: the *x*- and *y*-coordinates of the eyes for selected timestamps and the correctness of the answers (0 = incorrect, 1 = correct). As the stable scanpath (see Students' Strategies section) occurred only in the graph area, we did not use other AOIs here. Note that we are interested in *task-specific* strategies, not in general reading or viewing strategies. Both the previous qualitative study and another study (Lyford & Boels, 2022) suggested that reading axes was not a relevant part of such a strategy and would add noise when used in an MLA. Examples of other AOIs were the horizontal label (title of the horizontal axis), vertical label, horizontal axis, vertical axis, graph title, question, and 'next' button. We filtered and prepared

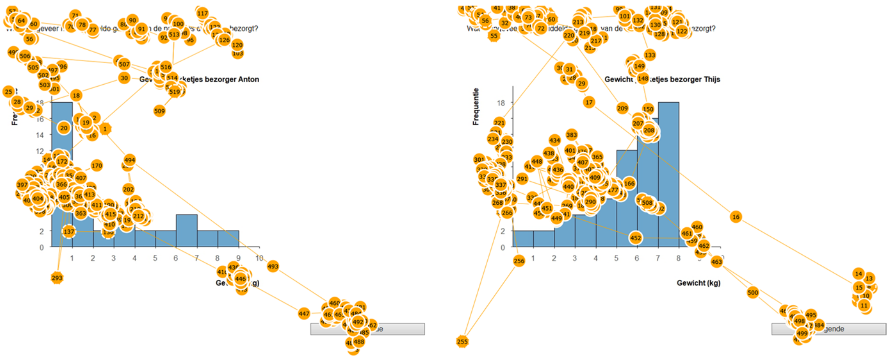

Fig. 8 Example of all gaze data on Item02 (left) and Item06 (right) of student L27. *Note.* This student incorrectly answered six and eight, respectively. The incorrect strategy is visible by the many horizontal saccades going from the left-hand side of the graph to the middle or right-hand side and the absence of vertical gazes going from the top or middle of bars to the bottom of the graph (see also Fig. 2 and 5 for comparison). The horizontal saccades in the gaze cloud on the *graph area* in Item02 (left) are smaller than in Item06 (right). On both graphs, this student applied an incorrect strategy, even though for the first items (e.g., Item02) this student looked at the titles of the axes (Looking at statistical graphs (e.g., concentrations of greenhouse gases from 0 to 2005), experts tend to spend more time on AOIs that help them understand the data in the graph (title, legend, axes) compared to novices (Harsh et al., 2019). The gaze pattern of student L27 on Item02 indicates that attending axes and graph titles might not be enough)

the data as follows. First,[9] we removed all data that fell outside the computer screen (all rows with *x*- and *y*-coordinates outside the range 0–1366 horizontally and 0–768 vertically). Then, we calculated new *y*-coordinates as 768 minus the original values, as the coordinate system is upside down in the Tobii software compared to the Cartesian plane commonly used in mathematics and preferred in Mathematica[10] so that a *y*-coordinate of 700 in Tobii indicates a position close to the bottom of the screen. In Mathematica, we selected the data that Tobii indicated as being within the graph area (signposted by a 1 in the column 'AOI graph' in the dataset) and then selected the coordinates that were in the graph area (pixels 500 to 1100, horizontally and pixels that were originally between 190 to 525, vertically, for all items, Fig. 9). We also removed few bad data (start and end line of the gazes on an item, as well as incomplete data due to data loss as described elsewhere). The order of the gaze data was kept in the input file but without timestamps.

The tool we used (the Classify function in Mathematica version 12.1) initially automatically chose random forest (considered to be a high-performance model, Kuhn & Johnson, 2013) as the best MLA for our continuous gaze data (WRI, 2020).

[9] Even before that, some data cleaning was partially done by hand, as, due to the huge amount of data, the Tobii software could not deliver the data in one database with all students and items together. Instead, all data were delivered per student (25 trials per student), whereas we wanted to have all data per item (50 trials per item). Moreover, the original dataset contained several empty lines that needed to be removed.

[10] In version 13.2.1, the vertically flipped coordinate system used by the Tobii software now can also be used.

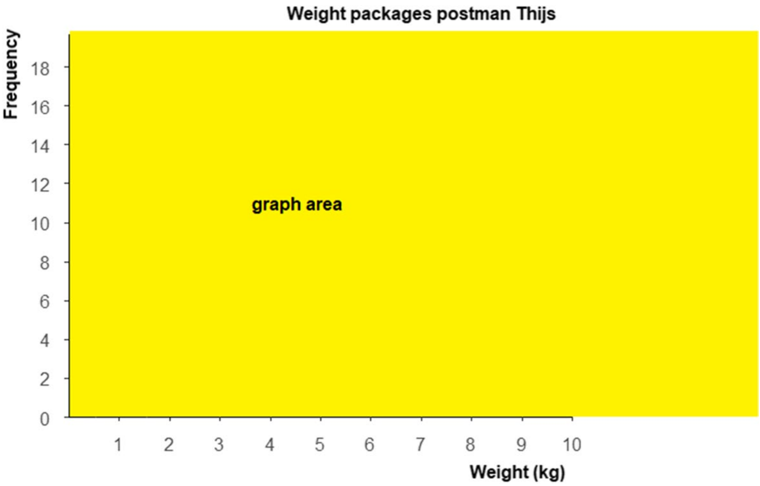

Fig. 9 Example of the AOI graph area (yellow rectangle) in Item06. *Note.* The size and place of this AOI are the same for all five items

During the review process, we made once more all analyses with the newest version of Mathematica, 13.2.1. Instead of the random forest, the software now suggested that logistic regression performs slightly better in most cases, and random forest in some. However, our analyses that are based on random forest still hold as the conclusions that can be drawn from these analyses with logistic regression are the same. As the MLA is intended as a baseline for our IMM, we, therefore, report the results of random forest for all items in the remainder of this article. We seeded the random forest and prescribed it as method in follow-up analyses (see code line below). Although random forest is known for its explainability, the way it is embedded in the software made us consider it a black box. This ML model is not to be confused with the IMM we constructed and was described in the section Construction of an Interpretable Mathematical Model. The following code line was used for obtaining a trained classifier function in follow-up analyses (some detailed codes can be found in Appendix):

```
Classify[{datarow1 → class1, datarow2 → class2...datarow50 → class50}, Method →"RandomForest"]
```

Here, each data row is a sequence of *x*–*y* pairs—the coordinates of the gaze locations on the screen, in pixel units, keeping the original order of the fixations—and each class is either 0 (incorrect strategy) or 1 (correct strategy). Classify returns a classifying function. After this training cycle, a list of data rows is fed to this Classify Function to obtain the algorithm's classification of the data. It returns a list of zeros and ones: the algorithm's identification of students' strategy correctness.

The ML analysis consisted of two steps: (1) verification of whether prerequisites were met (see the section Supervised Machine Learning Algorithm) and (2) identification of students' strategies. At each step, the MLA was re-trained. We repeated these two steps for all five items. As the results of the second step were above our expectations, we decided to take an extra step: training the MLA with one item and testing the MLA with another; see the Results of Applying an MLA and IMM section. We did this for all item pairs. In addition, we performed several cross-validation procedures as described in the Methodological Evaluation Criteria section.

Methodological Evaluation Criteria

Validity, reliability, and causality are important methodological evaluation criteria in educational research. Different terminology and definitions are used in data science and human–computer interaction research for evaluating MLA results. In this section, we compare these terminologies.

First, in educational research, validity "concerns whether we really measure what we intend to measure" (Bakker & Van Eerde, 2015, p. 443). In educational research, threats to the validity are internal and external, for example, maturation of subjects between measurements, subject selection effects on results, loss of subjects, changes in instrumentation, and so on (Eisenhart & Howe, 1992). This mostly applies to how the data were collected and what can be inferred from them. Regarding the data collection, as all data were collected in one session, maturation does not apply. Also, we did not exclude any subjects from the dataset, hence, there is no loss of subjects. We did not change our instrumentation, and so on. What does apply is that we selected subjects from pre-university track students in Grades 10–12 only, from one school, and only those who volunteered (which is inevitable). However, we observed the same phenomena (strategies) in different subjects in previous studies with secondary school teachers (Boels et al., 2019b) and with university students (Boels et al., 2018). Moreover, qualitative research does not seek to generalize from sample to population but from variation in the data to the phenomenon (Levitt, 2021). Regarding what can be inferred from the data: we combined the gaze data with the results of cued recall. Cued recall means that students were shown their gazes (the cue) and asked to explain what strategy they used (cf. Van Gog et al., 2005). This ensures that the data collection was valid. For the MLA and IMM, validity can be understood as whether these actually measure the phenomenon (strategies). As we discuss in the Research Approach section, this is true for the IMM by its design, but we cannot be sure about the MLA. However, the results of the IMM suggest that the phenomenon can be measured from the gaze data.

Second, reliability in educational research is about "independence of the researcher[s' judgment]" (Bakker & Van Eerde, 2015, p. 443) or small variation in outcomes. The reliability of a method entails that its results can be reproduced with the same population with comparable items (e.g., Golafshani, 2003). Reliability, in this sense, can be understood as the MLA results having about the same and sufficient accuracy as the results of the qualitatively identified strategies and of those of an IMM. Another way to look at the reliability of MLAs in this sense is by comparing results on different items (e.g., train the MLA with data of one item and identify students' strategies for another). When there is sufficient overlap, all are considered to identify the same phenomenon. This, in turn, makes the MLA results more reliable. Reliability here refers to what data scientists sometimes call the performance of the MLA.

In human–computer interaction research, reliability (also) involves the safety of the system, downtime, and consistency in the results (e.g., Bosnić & Kononenko, 2009; Webb et al., 2020), which is not relevant to us as we are not building an application.

Data scientists, however, define "reliability of classification as an estimated probability that the (single) classification is in fact the correct one" (Kukar & Kononenko, 2002, p. 219). To avoid confusion, we will, therefore, use *performance* when evaluating our results. Performance is also checked through cross-validation (e.g., Berrar, 2019) which involves applying the trained MLA to unseen data. We used several cross-validation procedures. First, we used a procedure often applied in statistical research—jackknife—which is a form of resampling (e.g., Efron & Stein, 1981) which means that the answers or strategies for all 50 students are identified in an iterative process by the MLA based on learning from the other 49 students. Since there are 50 students who can be left out one at a time, there are 50 ways to do this and the 50 results are averaged. Second, we performed a leave-one-out cross-validation (LOOCV) which means that data from 49 students are used as training data, and the strategy of the 50th student is classified. This is done 50 times until all students' data are used as test data once. Furthermore, we used a stratified fivefold cross-validation which means that the data are split into groups of ten students. Stratified means that in each group of ten students, the number of students with a correct strategy is roughly the same. Then, the MLA is trained with 40 students and classifies the strategies of the remaining 10. This is repeated five times until data from all groups are used once as test data.

An MLA's performance can be measured in different ways; through accuracy, through a confusion matrix (e.g., comparing the results of the MLA with the results of the qualitative coding), and through a ROC[11] plot (Fawcett, 2006; see the Results of Applying an MLA and IMM section) that gives an idea of the true positives and false negatives rates. Accuracy is expressed as a percentage of correctly predicted or identified cases (e.g., Afonja, 2017). As explained earlier, for our supervised MLA in this low-stake situation, we regard an accuracy of 70% or higher as good. In addition, we consider 80% or above as very good, and 90% or more as excellent. However, accuracy can be misleading. For example, if only ten percent of the students used the correct strategy, and the MLA identifies all strategies as incorrect, the accuracy would be 90%, but this identification would not be valid. Therefore, accuracy should be used with precaution. In a confusion matrix, counts for true and false positives and negatives are reported separately (see the Results of Applying an MLA and IMM section and Appendix). From this matrix, sensitivity and specificity can be calculated using the following formulas (cf. Kuhn & Johnson, 2013), which give a better idea of how the MLA is performing:

$$Sensitivity = \frac{\text{\#samples qualitatively coded correct } \textit{and} \text{ by MLA identified as correct (strategy)}}{\text{\#samples qualitatively coded having a correct strategy}}$$

This formula is often shortened to (e.g., Fawcett, 2006):

$$Sensitivity = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} = \frac{\text{true positives}}{\text{total positives in true class}} = \frac{TP}{P}$$

[11] ROC stands for Receiver Operating Characteristic. Originally it was used to judge how well a specific Receiver Operated meaning how well it was picking up enemy signals (radar). In our case, the plot visualizes how well the signal (in our case: true positives) is detected compared to false negatives.

In this second formula, it is not immediately clear what is considered to be the 'true' class, whereas in the first it is clear that we took the results of the qualitative study as the 'true' results and the MLA results as the hypothesized class. For example, for Item01 and the IMM, $TP = 14$ and

$P = 14 + 10$, (Table 6), therefore, $sensitivity = \frac{14}{24} = 0.583\ldots$ which is rounded to 0.58 (Table 5, Results of Applying an MLA and IMM section).

$$Specificity = \frac{\text{\#samples qualitatively coded incorrect } \textit{and} \text{ MLA identified as incorrect (strategy)}}{\text{\#samples qualitatively coded having an incorrect strategy}}$$

Similarly to the above formula for sensitivity, the formula for specificity is often shortened to:

$$Specificity = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} = \frac{\text{true negatives}}{\text{total negatives in true class}} = \frac{TN}{N}$$

Third, in educational research, measuring is only valid "if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure" (Borsboom et al., 2004, p. 1061). Data science does not use the word validity. With the MLA, we intend to measure students' strategy correctness based on their gazes. Therefore, variations in students' gazes should produce variations in the classification by the MLA. Furthermore, if an IMM can be understood by human beings—and describes the observed phenomenon accurately—it adds to the validity of both the model and the MLA (e.g.,Doshi-Velez & Kim, 2017; Rudin, 2019) as well as to its usability (e.g., Guidotti et al., 2018). Usability is another criterion from human–computer interaction literature. In the section Construction of an Interpretable Mathematical Model, we described how an IMM that meets these criteria can be constructed.

Fourth, causality implies that a change in the results of the MLA is due to a change in the real system (Doshi-Velez & Kim, 2017). Although not every eye movement is part of the task-solving strategy (e.g., Schindler & Lilienthal, 2019), eye movements and strategies do associate (e.g., Kok & Jarodzka, 2017). Therefore, instead of causality, we use association, meaning that if students perform a specific pattern of gazes, they use a specific strategy. If there is an association between a gaze pattern and a strategy, the accuracy of identifying this strategy by the IMM and the MLA will be sufficient or better.

Results of Applying an MLA and IMM

The details of the results, including confusion matrices, can be found in Appendix (Tables 8, 9, 10, 11, 12, 13, 14 and 15).

Machine Learning Algorithm Results

Supervised Machine Learning with Students' Answers' Correctness

The identification accuracy of the MLA for students' answers for the first item we looked at, Item06, turned out to be 88% (very good) with a jackknife cross-validation procedure. As we aimed to identify strategies, not students' answers, we needed accuracy to be at least 70%. This criterion is met for all items when using version 13.2.1 (Table 4). Based on the confusion matrices (Appendix Tables 10 and 11) sensitivity and specificity were calculated (Appendix Table 8).

Next, we trained the MLA with gaze data on one item and then identified answers on another item, see Appendix (Table 13) for the results. Accuracies varied between chance level (32%) to well above (70%).

Supervised Machine Learning with Students' Strategies' Correctness

The accuracy of the random forest MLA for identifying students' strategies in the first item we looked at, Item06, turned out to be 86% with a jackknife cross-validation procedure. This result is considered very good. Consistency of the MLA was tested through various procedures such as jackknife, leave-one-out cross-validation, and fivefold cross-validation (see section Methodological Evaluation Criteria). Overall, the MLA correctly identifies 71% to 88% of students' strategies (jackknife cross-validation) for five different (but all histogram) items (Table 4). Percentages for correct strategies in the qualitative study (Table 1) varied between 38 and 48% (or 52%– 62% when reversed) and strategies identification results are all well above these chance levels with jackknife. When applying other cross-validation procedures, results drop and vary from around chance level (38%) to good (78%) for leave-one-out cross-validation to around chance level (56%) to good (74%) for fivefold cross-validation. We consider these results a baseline for the IMM.

Table 4 Accuracies of the IMM and of the random forest MLA after cross-validation

Validation procedure	Classification of:	Item01	Item02	Item06	Item19	Item20	Overall
Jackknife	Answers (mean)	83%	80%	88%	86%	82%	84%
Jackknife	Strategies (mean)	71%	71%	86%	88%	83%	80%
Leave one out	Answers	66%	68%	74%	60%	58%	65%
Leave one out	Strategies	60%	38%	64%	78%	48%	58%
fivefold	Answers	62%	62%	62%	58%	68%	62%
fivefold	Strategies	56%	56%	64%	74%	62%	62%
IMM	Strategies	62%	70%	84%	70%	72%[a]	72%

Accuracy is expressed as a percentage of correctly predicted or identified cases (e.g., Afonja, 2017). The results of the qualitative study are treated as the positive case. [a]With a small adjustment of the slope, this could go up to 74%

Next, we trained the MLA with gaze data on one item and then identified strategies on all other items. Accuracies varied between around chance level (52%) to quite well above that level (80%) see Appendix (Table 14). Specifically interesting are the results when gazes on Item01 are used as training data (Fig. 10). Testing this trained random forest MLA for identifying strategies on other items resulted in accuracies that vary between good (72%) and very good (80%) which suggests that the MLA has the potential to generalize beyond a specific item, even when shapes and skewness of the histograms differ. See also the ROC plot (Fig. 10).

Based on the confusion matrices (Appendix Tables 12 and 13), specificity and sensitivity (e.g., Kuhn & Johnson, 2013) were calculated (Table 5). Sensitivity (identification of correct strategies) is low to acceptable and lower than the low to excellent specificity (identification of incorrect strategies) after cross-validation. An excellent specificity is favorable for a future application that seeks to provide feedback to this particular group of learners. In practice this could mean that only a few students who used an incorrect strategy will be missed (Q-incorrect, MLA-correct, type I error) which we consider most important for feedback. In addition, some more students who used a correct strategy would get feedback implying that they used an incorrect strategy (Q-correct, MLA-incorrect, type II error) while they were not. As such feedback would hint at the correct strategy, this feedback could make them think once more and then conclude that their strategy was correct, which is not a problem.

The confusion matrices (see Appendix), provide further insight into how well the results of the MLA align with the results of the qualitative coding. Differences between the qualitative coding of the strategies and the MLA results can be due to what data scientists call ground truth noise in the data. Ground truth, here, is what the strategies actually "are" (in the real world, independent of coding). Educational researchers would explain this noise as inconsistencies, inaccuracies (e.g., due to

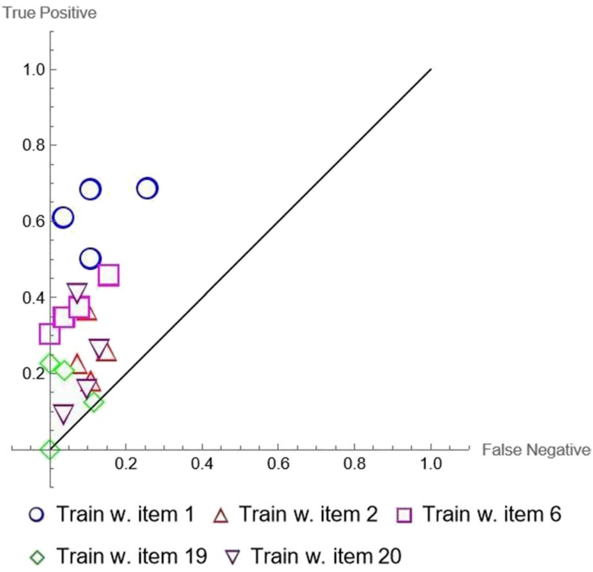

Fig. 10 ROC plot for strategies with train-test item pairs for random forest. *Note.* Ideally, for educational use, points should be concentrated in the upper left corner of the plot and close together for all items

merging two different but similar strategies in one code) or errors in the qualitative coding (as coders usually do not fully agree on the qualitative codes), or as noise in the gaze data (e.g., not every fixation or saccade on the graph area being part of the strategy). Differences can also be used to reconsider qualitative coding.

Results from the Interpretable Mathematical Model

With the IMM described earlier, we can correctly identify 62% to 84% of students' *strategies* (Table 4). From the confusion matrices (Tables 6 and 7) that compare the results of the IMM with the results of the qualitative study (Q), sensitivity and specificity can be calculated (Table 5). Sensitivity varies between low and good; specificity varies between acceptable and excellent, see also Fig. 11.

Comparison of IMM and Random Forest MLA results

The accuracy results of the IMM are quite close to the results of the random forest MLA with the jackknife cross-validation. In addition, the overlap between the IMM and MLA in identifying students' strategies before cross-validation was good and varied between 66 and 82% (Appendix Table 15). The results of the IMM are better than the MLA after cross-validation. Moreover, the results of both the IMM and the MLA indicate that strategies might be clearer in Item06, which is in line with what we qualitatively found. In the ROC plot (Fig. 11) the MLA results after cross-validation are compared to the IMM results. The

Table 5 Sensitivity and specificity of the MLA and the IMM when classifying strategies

Cross-validation procedure	Metric	Item01	Item02	Item06	Item19	Item20
Leave one out CV	Sensitivity	0.58	0.45	0.21	0.64	0.39
Leave one out CV	Specificity	0.62	0.32	0.90	0.89	0.56
fivefold CV	Sensitivity	0.50	0.50	0.21	0.59	0.57
fivefold CV	Specificity	0.62	0.61	0.90	0.86	0.67
IMM	Sensitivity	0.58	0.46	0.74	0.36	0.43
IMM	Specificity	0.65	0.89	0.90	0.96	0.96

Calculation of sensitivity and specificity is not possible for jackknife. CV = cross-validation

Table 6 Confusion matrices for Item01, 02, and 06 (IMM)

	Item01		Item02		Item06	
	Q-correct	Q-incorrect	Q-correct	Q-incorrect	Q-correct	Q-incorrect
IMM-correct	14	9	10	3	14	3
IMM-incorrect	10	17	12	25	5	28

N = 50 per item

Table 7 Confusion matrices for Item19 and 20 (IMM)

	Item19		Item20	
	Q-correct	Q-incorrect	Q-correct	Q-incorrect
IMM-correct	8	1	10	1
IMM-incorrect	14	27	13	26

N = 50 per item

plot shows that the IMM performs better. Altogether, this is considered a very good result, as it is not possible to know what parameters our MLA is using. As the accuracy of strategy identification of the MLA and the IMM is sufficient or above, the association is also sufficient or above.

The IMM is a more general model than the random forest model as the latter is different for each item. Given the MLA as a baseline for the IMM, we consider the current IMM likely to be a main component of a more general and precise model explaining gaze behavior during the kind of cognitive tasks considered in this article. The fact that the MLA, when trained on one item, relatively successfully predicts performance in a different item can be considered as evidence that the gaze data contain information about cognitive behavior on this task, even though this behavior is not explained by the MLA.

As this article aims to provide proof of principle, we did not further optimize the IMM. We chose the IMM that had the best overall performance for all items among the models we tried. When aiming to use the model in an application, one way to refine the IMM could be to split the model into two models: one only for identifying incorrect strategies (with two possible outcomes: the strategy

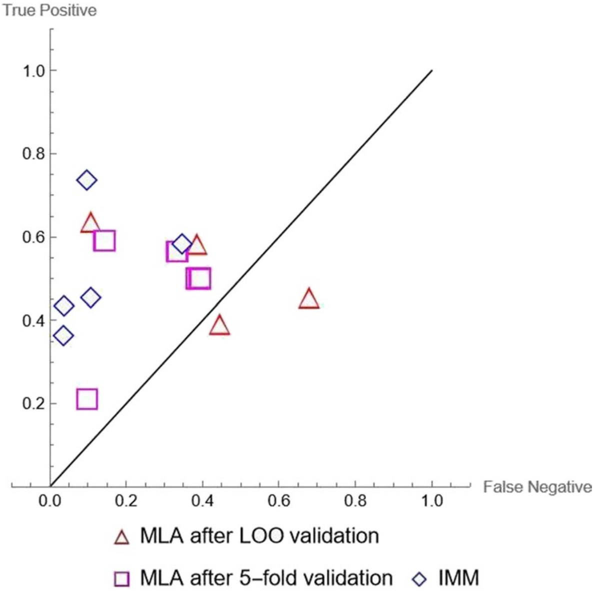

Fig. 11 ROC plot for strategy identification (Fawcett, 2006) in which "the point (0, 1) represents perfect classification" (p. 862). Random forest is the MLA the IMM is compared with. *Note.* Ideally, for educational use, points should be concentrated in the upper left corner of the plot and close together for all items. One triangle is hidden behind the square in the lower left corner. The MLA provides a baseline for the IMM. Although the IMM worked well, the plot shows there is room for improvement

is incorrect[12] or unknown) and one only for correct strategies. Combined, both models would have four options for strategy identification: correct strategy, incorrect strategy, unknown, or contradicting outcomes. Unknown means that the strategy is unknown, contradicting outcomes would require an extra rule for deciding what strategy it is. Another idea for optimizing these two models—besides adjusting the slope that distinguishes between horizontal and vertical gazes to obtain better results for specific items—is to adjust the saccade length in the model. In the current model, only saccades of at least 200 pixels are considered for both horizontal and vertical gazes. The saccade length for the vertical gazes could be scaled (shortened) to the size of the AOI. The relatively small height of the graph area (335 pixels) as opposed to the width (600 pixels) would justify such an adjustment. A possible improvement of the IMM could also be to incorporate alignment of saccades (relevant for the most common strategies). As the aim of the IMM in this research is not to find the best model but to show that the perceptual form of the stable scanpath can be captured by a model, we did not try to further optimize this model.

In addition, the performance of both IMM and machine learning might be influenced by students switching their strategy during or in between trials, making strategy identification less clear. There is some evidence in the interview data (e.g., Boels et al., 2022a, b and below) that students' strategies were influenced by Item13 to Item18 (with dotplots) which were designed to scaffold students in applying a correct strategy (e.g., Lyford, 2017). Although this did not result in a higher number of correct strategies, it might have changed the strategies on a more subtle level.

> L22: Yes then [Item19] I was going to change my approach a little bit. Then I started doing it a little bit similar to what I did with the dots. So then here is about more and there is again less, so then it will be somewhere here in between.

Conclusions and Discussion

Automated identification of students' strategies is a prerequisite for targeted intelligent feedback. The present study took on this challenge by providing an example of automated identification of students' *task-specific* strategies on single histograms based on gaze data. This could allow for future applications such as real-time feedback based on gaze data, for example, collected through webcams (e.g., Knoop-Van Campen et al., 2021).

We analyzed a set of gaze data in three phases: (1) an analysis of raw gaze data on *one* AOI (that contained the relevant scanpath) through an MLA that provided a baseline for the second step; (2) the construction of a separate interpretable mathematical model (IMM) using the same gaze data and insights on the perceptual form

[12] This incorrect strategy could also be further divided into a case-value plot interpretation strategy and a count-and-compute strategy not further discussed here.

of this stable scanpath from previous qualitative research; and (3) an evaluation of the results by comparing the performance of the MLA, the IMM and the overlap between the two. The IMM outperformed the MLA in several cases.

The MLA (phase 1, random forest as implemented in the software Mathematica Classify Function) has the advantage that it can process raw gaze data (*x*- and *y*-coordinates). It has the disadvantage that it is a black box in that it does not explain how it reached its decision for an individual student. The results of the random forest MLA after cross-validation provided our baseline for the IMM (phase 2). The IMM performs well (62% to 84% accuracy) with the advantage of being transparent for individual decisions and theoretically meaningful. The overlap between the results of the IMM and MLA is sufficient (phase 3).

Although these results are encouraging, an issue in data science is whether an MLA trained on one item is able to identify students' strategies for similar but different items. We, therefore, trained the MLA on one item and then tested it on all other items, and repeated this until all items were used once as training items. The accuracy results varied between around chance level and well above. The latter indicates very good performance. Combined, these results indicate that the IMM and MLA do describe the same phenomena—strategies—and that these strategies can be derived from students' gaze data.

What is new in our approach is that the filtering and preparation of the gaze data for the IMM and MLA are based on one AOI that contains the perceptual form of the gazes, instead of using, for example, the number of transitions between AOIs. This perceptual form is a stable scanpath indicating the student's focus of interaction and is interpreted as an attentional anchor (AA, e.g., Abrahamson & Sánchez-García, 2016). In our previous research (Boels et al., 2022a) we found that this perceptual form is indicative of students' strategies.

A prerequisite for the approach used is that a stable scanpath has been found in the gaze data. Furthermore, it requires gaze data to be classified into groups of students using the same strategy. Both prerequisites were met for our study. A further prerequisite for ML analysis is that the dataset is large enough and contains enough information for the MLA to identify (classify) students' strategies.

A limitation of our study is that we had gaze data on only fifty students. We alleviated this limitation by showing it worked for five items with differently shaped histograms, by using a resampling approach (jackknife cross-validation), and by training the MLA with gaze data from one item and then having it identify strategies for all other items. For future research, collecting data from a larger and different population is recommended. Another limitation of our study is that we use one item type (single graphs) and five variants of one graphical representation (histograms). It would be interesting to apply our approach to other domains, following the three phases described above for each new topic. Once the IMM is optimized and the MLA is trained, both an IMM and an MLA could be implemented. The MLA might be more accurate (see results of the jackknife cross-validation) but is computationally more complex. The IMM is easier, faster, provides insight into the relevant part of the gaze pattern that its decision was based on, and can deal with partial data. This allows feedback on strategies before an answer is given. The agreement between the IMM and MLA can be used as an indication of how certain the strategy identification was.

From a methodological perspective, a first contribution of this study is that our approach is both item-specific and generalizable. It is important that both IMM and MLA are item-specific, as the mathematical strategies of students are specific to an item type. Our IMM and MLA meet this requirement since the stable scanpath is specific to a given item; a stable scanpath refers to the *perceptual form* of this scanpath (e.g., a horizontal line, triangle, or point), not the sequence of AOIs. We expect that a stable scanpath can be found in gaze data on items in various domains (e.g., Strohmaier et al., 2020). It is also important that this method is generalizable, and, therefore, suitable for other mathematical domains. First, IMMs and MLAs (e.g., Rudin, 2019)—such as our set of rules for the IMM—are general methods. Moreover, our approach is also theoretically generalizable in the sense that educational researchers aim for "how and why the studied events occurred (or not)" (Yin, 2013, p. 326). The studied events are students' strategies and are observed as stable scanpaths that indicate students' focus of interaction with the item.

It could be argued that our approach is not very generalizable as *x*- and *y*-coordinates are sensitive to, for example, scaling, the position of the graph on the screen, and the shape of the histogram. However, the same could be argued for AOIs, as AOIs are *x*- and *y*-coordinates binned into categories by researchers, and, therefore, are also item-specific. Furthermore, coordinates can be rescaled which adds to their generalizability. In addition, we showed that for five differently shaped histograms, the IMM performed above chance level to good and the MLA performed at chance level to very good after cross-validation. In addition, the *same* IMM was used for all five items which makes it a more generable model than the MLA that was initially retrained for each item. Also adding to this generalizability is that, for several items, we have successfully trained the random forest MLA with this one item and then tested it for all other items.

A second methodological contribution is that *all* gaze data on one AOI are used, in contrast to methods that use aggregated data such as the *transition* from one AOI to another (e.g., Garcia Moreno-Esteva et al., 2020). Unlike heatmaps that are produced afterward (e.g., Schindler et al., 2021), raw gaze data offer the possibility of real-time feedback.

Third, we show that it is possible to *automatically* identify students' task-specific strategies from their gaze patterns. As soon as a stable scanpath is found, we think this can be grasped in an IMM as well as through an MLA, hence, be automated. Scanpaths are found for tasks in various mathematic domains: numbers (Schindler et al., 2020), arithmetic (Green et al., 2007), proportional reasoning (Shayan et al., 2017), area and perimeter (Shvarts, 2017), Cartesian coordinates (Chumachemko et al., 2014), geometry (Schindler & Lilienthal, 2019), trigonometry (Alberto et al., 2019), parabola (Shvarts & Abrahamson, 2019), statistical graphs (Boels et al., 2022a, b), and more (Lilienthal & Schindler, 2019; Strohmaier et al., 2020). Therefore, we believe that a similar approach can be used for other topics.

Fourth, using raw gaze data opens the possibility of implementation in an online (e.g., Cavalcanti et al., 2021) and adaptive tutoring system (e.g., Scheiter et al., 2019) with real-time feedback. The IMM can process raw data directly without the (black-box) preprocessing that the MLA performs and the results are straightforward to interpret.

Fifth, studying the differences between the results of the IMM and MLA on the one hand, and the qualitative coding on the other hand have the potential to improve the qualitative coding. Whenever the three methods lead to a different outcome, a closer inspection of the gaze patterns on the item, combined with interview data (if available), may lead to new insights for qualitative coding. This would combine the best capacities of people and machines, as suggested by Van de Schoot (2020). Sixth and final, our approach offers a new road for replicating results from a qualitative study.

From a theoretical perspective, this study shows that an AA can be used as a theoretical lens to search for a stable scanpath that reflects a mathematical strategy that is meaningful to the students. These stable scanpaths can be linked to the idea of an AA as follows. In a retrospective recall, students talked about an imagined action. This imagined action is—according to students—coordinated by an imaginary mathematical object: a horizontal or vertical line. As an AA is an existing or imagined object or area that emerges to facilitate or coordinate sensorimotor actions (Abrahamson & Sánchez-García, 2016), we also interpret these lines as an AA.

In addition, the manipulation of an imaginary object manifest in the gaze data could suggest links between theories of mental processes and embodied cognition. The AA was previously found when students interact with the environment. In our items, students cannot physically manipulate the graph. Nevertheless, gaze data show a stable scanpath indicating scanning of this imaginary object corresponding to a concrete location on the graph area on the screen (Boels et al., 2022a, b). We believe this reveals that cognitive processes can also be embodied and that eye movements can be a manifestation of both the perception and the action.

This article may fuel the dialogue between educational researchers and data science experts. An advantage of our IMM is its interpretability. MLAs may be experienced as a black box, and educational researchers may focus on how well it performs rather than how it performs. As educational researchers, we wondered what the application of data science tools to our data would bring us. It is important to promote the dialogue between educational researchers and MLA experts to keep boundaries between disciplines permeable. At such boundaries, exciting new research can emerge.

Future research into finding stable scanpaths for applying this method might, for example, concern geometry, such as the Pythagoras theorem or the cosine rule. In calculus, one might consider interpreting the slope or direction field when learning to solve differential equations. To allow task-specific gaze patterns to emerge an alternative way of introducing a topic could be considered (e.g., Janßen et al., 2020). Finally, the domain of graphical or diagrammatic literacy could be a future line of research. Several examples can be found in the literature of students having difficulties with graphs in mathematics (e.g., difficulties with complex line graphs, Carpenter & Shah, 1998; overgeneralization of linearity, Leinhardt et al., 1990; misreading of graphs, Roth

& Bowen, 2001; inadequate strategies, Tai et al., 2006), but also in science education (e.g., Kragten et al., 2015). All these domains have in common that spatial patterns may play a role.

Another direction for future research could be to improve the IMM and investigate the apparent trade-off between its sensitivity and specificity. We know that students used several strategies but it is unclear whether and how this is visible in this trade-off. In addition, a possible improvement of the IMM could be to tailor it to each item. Furthermore, the alignment of saccades could be included in the IMM (important for the most common strategies, see Boels et al., 2022a).

Future research might also focus on the appearances and changes in students' strategies over time. By using an IMM and an MLA, online automated feedback becomes possible on students' strategy, in some cases maybe even before students give their answers. This might make online feedback in massive online courses, online teaching, and homework more accurate and efficient. Another possibility would be to provide teachers with a dashboard on students' strategies (e.g., Knoop-Van Campen et al., 2021). The agreement between an MLA and an IMM could then be used to provide a measure of how reliable the strategy identification is. A prerequisite is the availability of cheap equipment for measuring eye movements. We expect more exact measuring of eye movements will be available for consumer computers in the near future, for example, through webcams. Whether this will be implemented in software and used by consumers will also depend on ethical discussions about privacy, fairness, bias, et cetera.

Appendix

In this Appendix, we provide additional code and all results of the IMM, random forest ML analysis and cross-validation procedures.

Mathematica Code used to find the Best Method for the Data

```
pickMethod[answerTrainingData[[item]],answerData[[item]],method, seed, performanceGoal]

tob = DateObject[]
beforeValidationResults =
          Monitor[Table[{pickMethod[answerTrainingData ⟦item⟧ , answerData ⟦item⟧ ,
                    Automatic, "1234", Automatic], pickMethod[strategyTrainingData ⟦item⟧ ,
                    strategyData ⟦item⟧ , Automatic, "1234", Automatic]},
             {item, 1, 5}], item]
DateObject[] - tob
Clear[tob]
```

Overview of the Results for Sensitivity and Specificity

For the results of strategy classification, see Table 5 in the article.

Table 8 Sensitivity and specificity of the random forest of answers

Validation procedure	What	Item01	Item02	Item06	Item19	Item20
Leave one out	Sensitivity	0.26	0.32	0.72	0.82	0.00
Leave one out	Specificity	0.90	0.90	0.76	0.13	0.88
5-fold	Sensitivity	0.11	0.05	0.40	0.68	0.12
5-fold	Specificity	0.94	0.97	0.84	0.38	0.97

Confusion Matrices of Answers

Table 9 Confusion matrices of answers for all items, random forest MLA, after LOOCV

	Item01		Item02		Item06		Item19		Item20	
	Qc	Qi	Qc	Qi	Qc	Qi	Qc	Qi	Qc	Qi
MLAc	5	3	6	3	18	6	28	14	0	4
MLAi	14	28	13	28	7	19	6	2	17	29

$N = 50$ per item. The results of the qualitative study (Q) compared with the results of the MLA random forest, c = correct answer, i = incorrect answer. LOOCV = leave-one-out cross-validation. The qualitative study is treated as the positive case. For example, the number 5 in the top-left corner of Item01 stands for 5 students identified by both the qualitative study and the MLA as having a correct strategy

Table 10 Confusion matrices of answers for all items, random forest MLA, after 5-fold cross-validation

	Item01		Item02		Item06		Item19		Item20	
	Qc	Qi	Qc	Qi	Qc	Qi	Qc	Qi	Qc	Qi
MLAc	2	2	1	1	10	4	23	10	2	1
MLAi	17	29	18	30	15	21	11	6	15	32

$N = 50$ per item. The results of the qualitative study (Q) compared with the results of the MLA random forest, c = correct answer, i = incorrect answer. The qualitative study is treated as the positive case

Confusion Matrices of Strategies

Table 11 Confusion matrices of strategies for all items, random forest MLA, after LOOCV

	Item01		Item02		Item06		Item19		Item20	
	Qc	Qi	Qc	Qi	Qc	Qi	Qc	Qi	Qc	Qi
MLAc	14	10	10	19	4	3	14	3	9	12
MLAi	10	16	12	9	15	28	8	25	14	15

N=*50* per item. The results of the qualitative study (Q) compared with the results of the MLA, c = correct strategy, i = incorrect strategy. LOOCV = leave-one-out cross-validation. The qualitative study is treated as the positive case

Table 12 Confusion matrices of strategies for all items, random forest MLA, after 5-fold cross-validation

	Item01		Item02		Item06		Item19		Item20	
	Qc	Qi	Qc	Qi	Qc	Qi	Qc	Qi	Qc	Qi
MLAc	12	10	11	11	4	3	13	4	13	9
MLAi	12	16	11	17	15	28	9	24	10	18

N=*50* per item. The results of the qualitative study (Q) compared with the results of the MLA, c = correct strategy, i = incorrect strategy. The qualitative study is treated as the positive case

Results of the Analyses when Using one Item as Training Item

Table 13 Accuracy of answer prediction for all items for the random forest MLA

	Item01	Item02	Item06	Item19	Item20
Item01	x	64%	68%	46%	62%
Item02	70%	x	52%	38%	64%
Item06	58%	56%	x	62%	60%
Item19	42%	48%	58%	x	38%
Item20	62%	64%	52%	32%	x

N=*50* per item. The item in each row is used as a training item. The items in the columns are the test items. For example, when using Item01 as a training item, the accuracy of predicting students' answers on Item02 is 64%

Table 14 Accuracy of strategy prediction for all items for the random forest MLA

	Item01	Item02	Item06	Item19	Item20
Item01	x	72%	72%	80%	80%
Item02	66%	x	70%	58%	58%
Item06	66%	66%	x	62%	68%
Item19	60%	56%	64%	x	68%
Item20	52%	58%	62%	70%	x

Comparison of the Random Forest MLA and IMM results

Table 15 Confusion matrices of strategies for all items for the random forest MLA versus the IMM

	Item01		Item02		Item06		Item19		Item20	
	MLAc	MLAi	MLAc	MLAi	MLAc	MLAi	MLAc	MLAi	MLAc	MLAi
IMMc	20	10	11	13	11	6	7	2	7	4
IMMi	4	16	4	22	3	30	9	32	8	31

N = 50 per item. The results of the MLA random forest compared with the results of the IMM, c = correct strategy, i = incorrect strategy. The MLA before any cross-validation is treated as the positive case

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40593-023-00368-9.

Acknowledgements We thank Nathalie Kuijpers for checking the document on APA style, English, and conciseness, Gerben van der Hoek for the second coding, Rutmer Ebbes for his contribution to the qualitative pilot study, Alex Lyford for calculating accuracy and precision measures and Wim Van Dooren for his contribution to the qualitative study that is used as input for training the machine learning algorithms. We are very grateful to the reviewers who provided many helpful suggestions and feedback.

Author Contributions All authors contributed to the study's conception and design. Previous eye-tracking data collection including material preparation and qualitative analysis was performed by LB. The machine learning analysis and interpretable mathematical model code were done by EGME. The first draft of the manuscript was written by LB and EGME and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The research was funded with a Doctoral Grant for Teachers from the Dutch Research Council (NWO), number 023.007.023, awarded to Lonneke Boels. Any opinions, findings, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Dutch Research Council.

Data Availability The following data are already available for public (re)use but not for commercial use: Dataset 1. This dataset contains the design of the previous eye-tracking study, the raw gaze data of the students, etc.: https://doi.org/10.34894/WEKAYE. Dataset 3: This dataset contains the processed data for 12 tasks, including three of the histogram tasks used in the present article. https://doi.org/10.34894/7KNEOH.

Declarations

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

References

Abrahamson, D., & Sánchez-García, R. (2016). Learning is moving in new ways: The ecological dynamics of mathematics education. *Journal of the Learning Sciences, 25*(2), 203–239. https://doi.org/10.1080/10508406.2016.1143370

Afonja, T. (2017). *Accuracy paradox*. TDS. Retrieved June 19, 2020, from https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b

Alberto, R. A., Bakker, A., Walker-van Aalst, O., Boon, P. B. J., & Drijvers, P. H. M. (2019). Networking theories in design research: An embodied instrumentation case study in trigonometry. In U. T. Jankvist, M. van den Heuvel-Panhuizen, & M. Veldhuis (Eds.), *Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education* (pp. 3088–3095). Freudenthal Group & Freudenthal Institute, Utrecht University and ERME. https://hal.archives-ouvertes.fr/hal-02418076/

Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education, 125*, 413–428. https://doi.org/10.1016/j.compedu.2018.06.023

Anderson, J. R., Bothell, D., & Douglass, S. (2004). Eye movements do not reflect retrieval processes: Limits of the eye-mind hypothesis. *Psychological Science, 15*(4), 225–231. https://doi.org/10.1111/j.0956-7976.2004.00656.x

Ashraf, H., Sodergren, M. H., Merali, N., Mylonas, G., Singh, H., & Darzi, A. (2018). Eye-tracking technology in medical education: A systematic review. *Medical Teacher, 40*(1), 62–69. https://doi.org/10.1080/0142159X.2017.1391373

Bakker, A., & Gravemeijer, K. P. E. (2006). An historical phenomenology of mean and median. *Educational Studies in Mathematics, 62*, 149–168. https://doi.org/10.1007/s10649-006-7099-8

Bakker, A., & Van Eerde, D. (2015). An introduction to design-based research with an example from statistics education. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Eds.), *Approaches to qualitative research in mathematics education. Advances in mathematics education* (pp. 429–466). Springer. https://doi.org/10.1007/978-94-017-9181-6_16

Ben-Zvi, D., Makar, K., & Garfield, J. (Eds.). (2017). *International handbook of research in statistics education* (1st ed.). Springer. https://doi.org/10.1007/978-3-319-66195-7

Berrar, D. (2019). Cross-validation. *Encyclopedia of bioinformatics and computational biology, 1* (pp. 542–545). Academic Press. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Boels, L., Ebbes, R., Bakker, A., Van Dooren, W., & Drijvers, P. (2018). Revealing conceptual difficulties when interpreting histograms: An eye-tracking study. Invited paper, refereed. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics* (pp. 1–4). ISI/IASE. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_8E2.pdf

Boels, L., Bakker, A., & Drijvers, P. (2019a). Eye tracking secondary school students' strategies when interpreting statistical graphs. In M. Graven, H. Venkat, A.A. Essien, & P. Vale (Eds.), *Proceedings 43rd Annual Meeting of the International Group for the Psychology of Mathematics Education (PME-43), 2*, (pp. 113–120). Pretoria, South Africa. http://www.igpme.org/publications/

Boels, L., Bakker, A., & Drijvers, P. (2019b). Unravelling teachers' strategies when interpreting histograms: an eye-tracking study. In U. T. Jankvist, M. Van den Heuvel-Panhuizen, & M. Veldhuis (Eds.), *Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education* (pp. 888–895). Freudenthal Group & Freudenthal Institute, Utrecht University and ERME. https://hal.archives-ouvertes.fr/hal-02411575/document

Boels, L., Bakker, A., Van Dooren, W., & Drijvers, P. (2019c). Conceptual difficulties when interpreting histograms: A review. *Educational Research Review, 28*, 100291. https://doi.org/10.1016/j.edurev.2019.100291

Boels, L., Bakker, A., Van Dooren, W., & Drijvers, P. (2022a). *Secondary school students' strategies when interpreting histograms and case-value plots: an eye-tracking study* [Manuscript submitted for publication]. Freudenthal Institute, Utrecht University.

Boels, L., Bakker, A., Van Dooren, W., & Drijvers, P. (2022b). *Gaze, interview and other data of secondary school students when interpreting statistical graphs* [Datapaper in preparation] Freudenthal Institute, Utrecht University. See also the dataset: https://doi.org/10.34894/WEKAYE

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Bosnić, Z., & Kononenko, I. (2009). An overview of advances in reliability estimation of individual predictions in machine learning. *Intelligent Data Analysis, 13*(2), 385–401. https://doi.org/10.3233/IDA-2009-0371

Brunyé, T. T., Drew, T., Weaver, D. L., & Elmore, J. G. (2019). A review of eye tracking for understanding and improving diagnostic interpretation. *Cognitive Research: Principles and Implications, 4*(1), 7. https://doi.org/10.1186/s41235-019-0159-2

Burrill, G. (2020). Statistical literacy and quantitative reasoning: Rethinking the curriculum. In P. Arnold (Ed.), *New Skills in the Changing World of Statistics Education Proceedings of the Roundtable conference of the International Association for Statistical Education.* ISI/IASE.

Cai, J., Moyer, J. C., & Grochowski, N. J. (1999). Making the mean meaningful: An instructional study. *Research in Middle Level Education Quarterly, 22*(4), 1–24. https://doi.org/10.1080/10848959.1999.11670153

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied, 4*(2), 75–100. https://doi.org/10.1037/1076-898X.4.2.75

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence, 2*, 100027. https://doi.org/10.1016/j.caeai.2021.100027

Chumachemko, D., Shvarts, A., & Budanov, A. (2014). The development of the visual perception of the Cartesian coordinate system: An eye tracking study. In C. Nicol, P. Liljedahl, S. Oesterle, & D. Allan (Eds.), *Proceedings of the Joint Meeting of PME 38 and PME-NA 36, 2* (pp. 313–320). PME. Retrieved June 29, 2020 from: https://files.eric.ed.gov/fulltext/ED599779.pdf

Cooper, L. L. (2018). Assessing students' understanding of variability in graphical representations that share the common attribute of bars. *Journal of Statistics Education, 26*(2), 110–124. https://doi.org/10.1080/10691898.2018.1473060

D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies, 70*(5), 377–398. https://doi.org/10.1016/j.ijhcs.2012.01.004

Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., & Holmqvist, K. (2012). It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior Research Methods, 44*, 1079–1100. https://doi.org/10.3758/s13428-012-0212-2

Dewhurst, R., Foulsham, T., Jarodzka, H., Johansson, R., Holmqvist, K., & Nyström, M. (2018). How task demands influence scanpath similarity in a sequential number-search task. *Vision Research, 149*, 9–23.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv Preprint.* Retrieved May 19, 2020 from: arXiv:1702.08608v2 [stat.ML] https://arxiv.org/abs/1702.08608

Efron, B., & Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics, 9*(3), 586–596. https://www.jstor.org/stable/2240822

Eisenhart, M., & Howe, K. (1992). Validity in educational research. In M. LeCompte, W. Millroy, & J. Preissle (Eds.), *The Handbook of Qualitative Research in Education* (pp. 64–680). Academic Press. http://www.elsevier.com/wps/find/bookdescription.cws_home/674919/description#description

Eivazi, S., & Bednarik, R. (2010). Inferring problem solving strategies using eye-tracking: System description and evaluation. *Proceedings of the 10th Koli Calling International Conference on Computing Education Research* (pp. 55–61). Association for Computing Machinery.

Epelboim, J., & Suppes, P. (2001). A model of eye movements and visual working memory during problem solving in geometry. *Vision Research, 41*(12), 1561–1574. https://www.sciencedirect.com/science/article/pii/S004269890000256X

Fabbri, S., Stubbs, K. M., Cusack, R., & Culham, J. C. (2016). Disentangling representations of object and grasp properties in the human brain. *The Journal of Neuroscience, 36*(29), 7648–7662. https://doi.org/10.1523/JNEUROSCI.0313-16.2016

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letter, 27*, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Feng, S., & Law, N. (2021). Mapping artificial intelligence in education research: A network-based keyword analysis. *International Journal of Artificial Intelligence in Education, 31*, 277–303. https://doi.org/10.1007/s40593-021-00244-4

Gal, I. (1995). Statistical tools and statistical literacy: The case of the average. *Teaching Statistics, 17*(3), 97–99.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*(1), 1–25. https://doi.org/10.1111/j.1751-5823.2002.tb00336.x

Garcia Moreno-Esteva, E., White, S. L. J., Wood, J. M., & Black, A. A. (2018). Application of mathematical and machine learning techniques to analyse eye-tracking data enabling better understanding of children's visual-cognitive behaviours. *Frontline Learning Research, 6*(3), 72–84. https://doi.org/10.14786/flr.v6i3.365

Garcia Moreno-Esteva, E., Kervinen, A., Hannula, M. S., & Uitto, A. (2020). Scanning signatures: A graph theoretical model to represent visual scanning processes and A proof of concept study in biology education. *Education Sciences, 10*(5), 141. https://doi.org/10.3390/educsci10050141

Garcia Moreno-Esteva, E., White, S.L.J., Wood, J., & Black, A. (2016). Mathematical and computational modeling of eye-tracking data to predict success in a problem solving task. In *Proceedings of the 40th PME* (Vol. 1, p. 163).

Gerard, L., Matuk, C., McElhaney, K., & Linn, M. C. (2015). Automated, adaptive guidance for K–12 education. *Educational Research Review, 15*, 41–58. https://doi.org/10.1016/j.edurev.2015.04.001

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report, 8*(4), 597–606. https://doi.org/10.46743/2160-3715/2003.1870

Goldberg, J. H., & Helfman, J. I. (2010). Comparing information graphics: A critical look at eye tracking. *Proceedings of the 3rd BELIV'10 Workshop: BEyond Time and Errors: Novel EvaLuation Methods for Information Visualization* (pp. 71–78). Association for Computing Machinery. https://doi.org/10.1145/2110192.2110203

Goldberg, J., & Helfman, J. (2011). Eye tracking for visualization evaluation: Reading values on linear versus radial graphs. *Information Visualization, 10*(3), 182–195. https://doi.org/10.1177/1473871611406623

Green, H. J., Lemaire, P., & Dufau, S. (2007). Eye movement correlates of younger and older adults' strategies for complex addition. *Acta Psychologica, 125*(3), 257–278. https://doi.org/10.1016/j.actpsy.2006.08.001

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 93. https://doi.org/10.1145/3236009

Hancox-Li, L. (2020). Robustness in machine learning explanations: Does it matter? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372836

Harsh, J. A., Campillo, M., Murray, C., Myers, C., Nguyen, J., & Maltese, A. V. (2019). "Seeing" data like an expert: An eye-tracking study using graphical data representations. *LSE, 18*(3), 32. https://doi.org/10.1187/cbe.18-06-0102

Harteis, C., Kok, E. M., & Jarodzka, H. (2018). The journey to proficiency: Exploring new objective methodologies to capture the process of learning and professional development. *Frontline Learning Research, 6*(3), 1–5. https://doi.org/10.14786/flr.v6i3.435

Hessels, R. S., Niehorster, D. C., Nyström, M., Andersson, R., & Hooge, I. T. C. (2018). Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers. *Royal Society Open Science, 5*(8), 1–23. https://doi.org/10.1098/rsos.180502

Holmqvist, K., Örbom, S. L., Hooge, I. T. C., Niehorster, D. C., Alexander, R. G., Andersson, R., Benjamins, J. S., Blignaut, P., Brouwer, A-M., Chuang, L. L., Dalrymple, K. A., Drieghe, D., Dunn, M. J., Ettinger, U., Fiedler, S., Foulsham, T., Van der Geest, J. N., Witzner Hansen, D., Hutton, S., … Hessels, R. S. (2023). Eye tracking: empirical foundations for a minimal reporting guideline. *Behavior Research Methods, 55,* 364–416. https://doi.org/10.3758/s13428-021-01762-8

Hutto, D. D., & Sánchez-García, R. (2015). Choking RECtified: Embodied expertise beyond Dreyfus. *Phenomenology and the Cognitive Sciences, 14*(2), 309–331. https://doi.org/10.1007/s11097-014-9380-0

Hwang, G.-J., & Tu, Y.-F. (2021). Roles and research trends of artificial intelligence in mathematics education: A bibliometric mapping analysis and systematic review. *Mathematics, 9*(6), 584. https://doi.org/10.3390/math9060584

Hyönä, J. (2010). The use of eye movements in the study of multimedia learning. *Learning and Instruction, 20*(2), 172–176. https://doi.org/10.1016/j.learninstruc.2009.02.013

Janßen, T., Vallejo-Vargas, E., Bikner-Ahsbahs, A., & Reid, D. A. (2020). Design and investigation of a touch gesture for dividing in a virtual manipulative model for equation-solving. *Digital Experiences in Mathematics, 6*, 166–190. https://doi.org/10.1007/s40751-020-00070-8

Jarodzka, H., Holmqvist, K. & Nyström, M. (2010). A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 211–218). ACM.

Jarodzka, H., Holmqvist, K. & Gruber, H. (2017). Eye tracking in educational science: Theoretical frameworks and research agendas. *Journal of Eye Movement Research, 10*(1). https://doi.org/10.16910/jemr.10.1.3

Järvelä, S., Malmberg, J., Haataja, E., Sobocinski, M., & Kirschner, P. A. (2019). What multimodal data can tell us about the students' regulation of their learning process? *Learning and Instruction,* 101203. https://doi.org/10.1016/j.learninstruc.2019.04.004

Kang, J., Han, X., Song, J., Niu, Z., & Li, X. (2020). The identification of children with autism spectrum disorder by SVM approach on EEG and eye-tracking data. *Computers in Biology and Medicine, 120*, 103722. https://doi.org/10.1016/j.compbiomed.2020.103722

Kaplan, J. J., Gabrosek, J. G., Curtiss, P., & Malone, C. (2014). Investigating student understanding of histograms. *Journal of Statistics Education, 22*(2), 1–30. https://doi.org/10.1080/10691898.2014.11889701

Kersting, K. (2018). Machine learning and artificial intelligence: Two fellow travelers on the quest for intelligent behavior in machines. *Frontiers in Big Data, 1*. https://doi.org/10.3389/fdata.2018.00006

Klein, P., Becker, S., Küchemann, S., & Kuhn, J. (2021). Test of understanding graphs in kinematics: Item objectives confirmed by clustering eye movement transitions. *Physical Review Physics Education Research, 17*(1), 013102. https://doi.org/10.1103/PhysRevPhysEducRes.17.013102

Knoop-Van Campen, C. A. N., Kok, E., Doornik, R. V., Vries, P. D., Immink, M., Jarodzka, H., & Van Gog, T. (2021). How teachers interpret displays of students' gaze in reading comprehension assignments. *Frontline Learning Research, 9*(4), 116–140. https://doi.org/10.14786/flr.v9i4.881

Kok, E.M., & Knoop-van Campen, C. A. N. (2022). Using webcam-based eye-tracking to uncover reading strategies. *Presentation at the EARLI SIG 27 conference*, Southampton, UK.

Kok, E., & Jarodzka, H. (2017). Before your very eyes: The value and limitations of eye tracking in medical education. *Medical Education, 51*(1), 114–122. https://doi.org/10.1111/medu.13066

Konold, C., & Pollatsek, A. (2004). Conceptualizing an average as a stable feature of a noisy process. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 169–199). Springer. https://doi.org/10.1007/1-4020-2278-6_8

Kragten, M., Admiraal, W., & Rijlaarsdam, G. (2015). Students' learning activities while studying biological process diagrams. *International Journal of Science Education, 37*(12), 1915–1937. https://doi.org/10.1080/09500693.2015.1057775

Krause, J., Perer, A., & Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems,* 5686–5697. https://doi.org/10.1145/2858036.2858529

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Kukar, M., & Kononenko, I. (2002). Reliable classifications with machine learning. In T. Elomaa, H. Mannila, & H. Toivonen (Eds.), *European Conference on Machine Learning, Lecture Notes in Computer Science* (Vol. 2430, pp. 219–231). Springer.

Lai, M., Tsai, M., Yang, F., Hsu, C., Liu, T., Lee, S. W., . . . Tsai, C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review, 10*, 90–115. https://doi.org/10.1016/j.edurev.2013.10.001

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). Faithful and customizable explanations of black box models. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 131–138). AAAI/ACM. https://doi.org/10.1145/3306618.3314229

Lakoff, G., & Núñez, R. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being*. Basic Books.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Lawson, M. J. (1990). The case for instruction in the use of general problem-solving strategies in mathematics teaching: A comment on Owen and Sweller. *Journal for Research in Mathematics Education, 21*(5), 403–410. https://doi.org/10.2307/749397

Leinhardt, G., Zaslavsky, O., & Stein, M. K. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research, 60*(1), 1–64. https://doi.org/10.3102/00346543060001001

Levitt, H. M. (2021). Qualitative generalization, not to the population but to the phenomenon. *Qualitative Psychology, 8*(1), 95–110. https://doi.org/10.1037/qup000018

Lilienthal, A., & Schindler, M. (2019). Eye-tracking research in mathematics education: A PME literature review. In M. Graven, H. Venkat, A.A. Essien, & P. Vale (Eds.), *Proceedings 43rd Annual Meeting of the International Group for the Psychology of Mathematics Education (PME-43), 4* (pp. 62–62). Extended version of this article retrieved May 5, 2019 from https://arxiv.org/abs/1904.12581

Lyford, A., & Boels, L. (2022). Using machine learning to understand students' gaze patterns on graphing tasks. Invited paper: refereed. *Proceedings of the Eleventh International Conference on Teaching Statistics (ICOTS11)* (pp. 1–6). IASE. http://www.iase-web.org/Conference_Proceedings.php?p=ICOTS_11_2022

Lyford, A. J. (2017). *Investigating undergraduate student understanding of graphical displays of quantitative data through machine learning algorithms* [Doctoral dissertation, University of Georgia]. https://iase-web.org/documents/dissertations/17.AlexanderLyford.Dissertation.pdf

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent. Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society, 100*(11), 2175–2199. https://doi.org/10.1175/BAMS-D-18-0195.1

Mitchell, T. M., Buchanan, B., De Jong, G., Dietterich, T., Rosenbloom, P., & Waibel, A. (1990). Machine learning. *Annual Review of Computer Science, 4*, 417–433. https://doi.org/10.1146/annurev.cs.04.060190.002221

Mitev, N., Renner, P., Pfeiffer, T., & Staudte, M. (2018). Towards efficient human–machine collaboration: Effects of gaze-driven feedback and engagement on performance. *Cognitive Research: Principles and Implications, 3*, 51. https://doi.org/10.1186/s41235-018-0148-x

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education, 26*(1), 20–39. https://doi.org/10.2307/749226

Molenberghs, P., Cunnington, R., & Mattingley, J. B. (2012). Brain regions with mirror properties: A meta-analysis of 125 human fMRI studies. *Neuroscience & Biobehavioral Reviews, 36*(1), 341–349. https://doi.org/10.1016/j.neubiorev.2011.07.004

Molnar, C. (2019). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable.* Retrieved September 1, 2023 from https://christophm.github.io/interpretable-ml-book/index.html

Najar, A. S., Mitrovic, A., & Neshatian, K. (2014). Utilizing eye tracking to improve learning from examples. In C. Stephanidis & M. Antona (Eds.), *Lecture Notes in Computer Science 8514. Proceedings of the Eighth International Conference on Universal Access in Human-Computer Interaction part 2* (pp. 410–418). Springer. https://doi.org/10.1007/978-3-319-07440-5_38

Obersteiner, A., & Tumpek, C. (2016). Measuring fraction comparison strategies with eye-tracking. *ZDM, 48*, 255–266. https://doi.org/10.1007/s11858-015-0742-z

Peebles, D., & Cheng, P. C. (2001). Graph-based reasoning: From task analysis to cognitive explanation. *Proceedings of the Annual Meeting of the Cognitive Science Society, 23.* Retrieved February 5, 2021 https://escholarship.org/uc/item/9rz4r25j

Radford, L. (2010). The eye as a theoretician: Seeing structures in generalizing activities. *For the Learning of Mathematics, 30*(2), 2–7. https://www.jstor.org/stable/20749442

Roth, W.-M., & Bowen, G. M. (2001). Professionals read graphs: A semiotic analysis. *Journal for Research in Mathematics Education, 32*(2), 159–194. https://doi.org/10.2307/749672

Rowlands, M. J. (2010). *The new science of the mind*. The MIT Press.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Russo, J. E. (2010). Eye fixations as a process trace. In M. Schulte-Mecklenbeck, A. Kühberger, & R. Ranyard (Eds.), *Handbook of process tracing methods for decision research* (pp. 43–64). Psychology Press. https://doi.org/10.4324/9781315160559

Scheiter, K., Schubert, C., Schülera, A., Schmidt, H., Zimmermann, G., Wassermann, B., Krebsa, M.-C., & Edera, T. (2019). Adaptive multimedia: Using gaze-contingent instructional guidance to provide personalized processing support. *Computers & Education, 139*, 31–47. https://doi.org/10.1016/j.compedu.2019.05.005

Schindler, M., Schaffernicht, E., & Lilienthal, A. J. (2021). Identifying student strategies through eye tracking and unsupervised learning: The case of quantity recognition. In M. Inprasitha, N. Changsri, & N. Boonsena (Eds.), *Proceedings of the Forty-fourth Conference of the International Group for the Psychology of Mathematics Education*, *4* (pp. 9–16). Retrieved August 17, 2021 from https://www.igpme.org/wp-content/uploads/2022/04/Volume-4_final.pdf

Schindler, M., & Lilienthal, A. J. (2019). Domain-specific interpretation of eye-tracking data: Towards a refined use of the eye-mind hypothesis for the field of geometry. *Educational Studies in Mathematics, 101*, 123–139. https://doi.org/10.1007/s10649-019-9878-z

Shayan, S., Abrahamson, D., Bakker, A., Duijzer, A., & Van der Schaaf, M. F. (2017). Eye-tracking the emergence of attentional anchors in a mathematics learning tablet activity. In C. Was, F. Sansosti, & B. Morris (Eds.), *Eye-tracking technology applications in educational research* (pp. 166–194). IGI-Global. https://doi.org/10.4018/978-1-5225-1005-5.ch009

Shvarts, A., & Abrahamson, D. (2019). Dual-eye-tracking Vygotsky: A microgenetic account of a teaching/learning collaboration in an embodied-interaction technological tutorial for mathematics. *Learning, Culture and Social Interaction, 22*, 100316. https://doi.org/10.1016/j.lcsi.2019.05.003

Shvarts, A. (2017). Eye movements in emerging conceptual understanding of rectangle area. In B. Kaur, W. K. Ho, T. L. Toh, & B. H. Choy (Eds.), *Proceedings of the Forty-first Conference of the International Group for the Psychology of Mathematics Education, Vol. 1* (p. 268).

Spivey, M. J., & Dale, R. (2011). Eye movements both reveal and influence problem solving. In S. P. Liversedge, I. Gilchrist, & S. Everling (Eds.), *The Oxford Handbook of Eye Movements* (pp. 551–562). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199539789.013.0030

Strohmaier, A. R., MacKay, K. J., Obersteiner, A., & Reiss, K. M. (2020). Eye-tracking methodology in mathematics education research: A systematic literature review. *Educational Studies in Mathematics, 104*, 147–200. https://doi.org/10.1007/s10649-020-09948-1

Susac, A. N., Bubic, A., Kaponja, J., Planinic, M., & Palmovic, M. (2014). Eye movements reveal students' strategies in simple equation solving. *International Journal of Science and Mathematics Education, 12*(3), 555–577. https://doi.org/10.1007/s10763-014-9514-4

Sweller, J. (1990). On the limited evidence for the effectiveness of teaching general problem-solving strategies. *Journal for Research in Mathematics Education, 21*(5), 411–415. https://doi.org/10.2307/749398

Tacoma, S. G., Heeren, B. J., Jeuring, J. T., & Drijvers, P. H. M. (2019). Automated feedback on the structure of hypothesis tests. In U. T. Jankvist, M. van den Heuvel-Panhuizen, & M. Veldhuis (Eds.), *Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education* (pp. 2969–2976). Freudenthal Group & Freudenthal Institute, Utrecht University and ERME. https://hal.archives-ouvertes.fr/hal-02428867v1

Tai, R. H., Loehr, J. F., & Brigham, F. J. (2006). An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments. *International Journal of Research & Method in Education, 29*(2), 185–208. https://doi.org/10.1080/17437270600891614

Van de Schoot, R. (2020). Machines vervangen wetenschappers [Machines replace scientists]. Retrieved February 14, 2020 from https://www.rensvandeschoot.com/oratie/

Van der Gijp, A., Ravesloot, C. J., Jarodzka, H., Van der Schaaf, M. F., Van der Schaaf, I. C., Van Schaik, J. P. J., & Ten Cate, Th. J. (2017). How visual search relates to visual diagnostic performance: A narrative systematic review of eye-tracking research in radiology. *Advances in Health Sciences Education, 22*(3), 765–787. https://doi.org/10.1007/s10459-016-9698-1

Van Gog, T., & Jarodzka, H. (2013). Eye tracking as a tool to study and enhance cognitive and metacognitive processes in computer-based learning environments. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 143–156). Springer. https://doi.org/10.1007/978-1-4419-5546-3_10

Van Gog, T., Paas, F., Van Merriënboer, J. J., & Witte, P. (2005). Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied, 11*(4), 237–244. https://doi.org/10.1037/1076-898X.11.4.237

Villagrá-Arnedo, C. J., Gallego-Durán, F. J., Llorens-Largo, F., Compañ-Rosique, P., Satorre-Cuerda, R., & Molina-Carmona, R. (2017). Improving the expressiveness of black-box models for predicting student performance. *Computers in Human Behavior, 72*, 621–631. https://doi.org/10.1016/j.chb.2016.09.001

Voisin, S., Pinto, F., Morin-Ducote, G., Hudson, K. B., & Tourassi, G. D. (2013). Predicting diagnostic error in radiology via eye-tracking and image analytics: Preliminary investigation in mammography. *Medical Physics, 40*(10), 101906-01–101906-10. https://doi.org/10.1118/1.4820536

Wade, N. J., & Tatler, B. W. (2011). Origins and applications of eye movement research. In S. P. Liversedge, I. Gilchrist, & S. Everling (Eds.), *The Oxford Handbook on Eye Movements* (pp. 17–46). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199539789.013.0002

Webb, M. E., Fluck, A., Magenheim, J., Malyn-Smith, J., Waters, J., Deschêne, M., & Zagami, J. (2020). Machine learning for human learners: Opportunities, issues, tensions and threats. *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-020-09858-2

Wilson, T. D. (1994). The proper protocol: Validity and completeness of verbal reports. *Psychological Science, 5*(5), 249–252. https://doi.org/10.1111/j.1467-9280.1994.tb00621.x

WRI. (2020). *Mathematica [computer software]*. Wolfram Research. https://www.wolfram.com/mathematica/

Yin, R. K. (2013). Validity and generalization in future case study evaluations. *Evaluation, 19*(3), 321–332. https://doi.org/10.1177/1356389013497081

Yuan, L., Haroz, S., & Franconeri, S. (2019). Perceptual proxies for extracting averages in data visualizations. *Psychonomic Bulletin & Review, 26*(2), 669–676. https://doi.org/10.3758/s13423-018-1525-7

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.