RISTAL 4 / 2021

Research in Subject-matter Teaching and Learning

Volume 4 – Special Issue edited by Timo Leuders & Katharina Loibl

Citation:
Brunner, K., Obersteiner, A., Leuders, T. (2021). How prospective teachers detect potential difficulties in mathematical tasks – an eye tracking study [Special Issue]. *RISTAL, 4*, 109–126.

DOI: https://doi.org/10.23770/RT1845

ISSN 2616-7697

How prospective teachers detect potential difficulties in mathematical tasks – an eye tracking study

Kirsten Brunner, Andreas Obersteiner & Timo Leuders

Abstract

An important aspect of mathematics teachers' diagnostic competences is the ability to judge the difficulty of a mathematical task. The process of judging task difficulty includes the perception and interpretation of task characteristics that are potentially challenging for students. Such judgement processes are often quick and difficult to assess. Most previous studies described these processes on the basis of teachers' verbal reports. A more recent approach to tap into cognitive processes is eye tracking. However, there is no firm knowledge yet whether eye tracking allows for a reliable assessment of teachers' judgements of mathematical task difficulty. The present study aims at filling this gap. We asked $N = 55$ prospective mathematics teachers to judge the difficulty of 20 tasks on linear functions, some of which included characteristics that are well known to be challenging for students. Participants viewed the tasks on a computer screen while their eye movements were recorded with an eye tracker. Our analyses of various eye-tracking parameters suggest that "fixation duration", "fixation duration average" and "number of fixations" were the most reliable measures of participants' perception and interpretation processes across a set of tasks. These measures were also correlated with participants' judgement accuracy. Using qualitative analyses of two participants' eye-tracking data, we illustrate when and how they processed the relevant task characteristics. In conclusion, eye tracking may be considered a suitable method for assessing how teachers detect task difficulty. We discuss implications for the use of eye tracking in further research on teachers' diagnostic competences.

Keywords

Diagnostic competences, diagnostic processes, judgement, eye tracking, functions and graphs, typical student errors

1 Introduction

An essential aspect of teachers' diagnostic competences is their ability to correctly judge task difficulty (Corno, 2008; Stein, Grover, & Henningsen, 1996; Sullivan, Clarke, & Clarke, 2012). Accurate diagnostic judgements are an important element for adaptive teaching (Hardy, Decristan, & Klieme, 2019; Parsons et al., 2018). Judging task difficulty is fundamental for adapting instruction to students' learning objectives and their learning requirements (Bromme, 1981; Chapman, 2014; Stein et al., 1996). Diagnostic judgements are relevant in all subject domains including mathematics.

Previous studies have predominantly focused on the accuracy and quality of diagnostic judgements as the outcome of a judgement process (Karst, 2012; McElvany et al., 2009; Schrader, 1989). In contrast, little is known about the underlying cognitive processes, such as the perception and interpretation of relevant task characteristics. Loibl, Leuders,

DOI: https://doi.org/10.23770/RT1845

and Dörfler (2020) suggested a framework for investigating teachers' diagnostic judgements by cognitive modeling (DiaCoM). The framework distinguishes between internal diagnostic processes, which are not directly measurable, and external diagnostic behavior, which is observable. This distinction addresses the methodological challenge of generating theoretical knowledge about diagnostic processes. Most previous studies that did address diagnostic processes used think-aloud interviews (e.g., Philipp, 2018; Reinhold, 2018), which have limitations because they only measure processes that participants are aware of and able to verbalize (Ericsson & Simon, 1980; Waern, 1988).

The present study explores whether eye-tracking technology is a suitable method for gaining insight into the diagnostic process and can thus complement currently used methods. The specific aim of our study was to investigate whether eye tracking allows researchers to reliably assess perception and interpretation, which are essential components of the diagnostic process (Loibl et al., 2020), and to examine how these processes are related to diagnostic judgement accuracy.

In the following sections, we first elaborate on the notion of diagnostic competences and on the processes of teachers' diagnostic thinking. We also describe the theoretical assumptions of our study using the DiaCoM framework. We then review the literature on diagnostic judgementand present the state of discussion on the application of eye-tracking technology as a method in mathematics education research. We reflect the relevance of our approach for education research on teachers' diagnostic judgment processes in other disciplines. Finally, we review the literature on the typical challenges students face with solving mathematical tasks that include functions. Our central assumption is that although diagnostic judgements are relevant in all subjects, judging task difficulty will strongly depend on content-specific task characteristics and on one's subject matter knowledge about these task characteristics.

1.1 Diagnostic competences and processes

Diagnostic competences are an important facet of teacher competences. They include the ability to judge task difficulty. If teachers aim to support students adaptively, they need to be aware of students' challenges. Teachers' diagnostic competences include accurate judgements of students' competences as well as accurate judgements of learning material, such as task difficulty (Karst, 2012; Schrader, 1989).

The quality of diagnostic judgements is usually evaluated on the basis of judgement accuracy (Helmke & Schrader, 1987; Südkamp, Kaiser, & Möller, 2012). However, according to Artelt and Rausch (2014) the term diagnostic judgement includes both, process and result. To reach the result of a judgement, a teacher needs to perform a diagnostic judgement process. The DiaCoM framework (Loibl et al., 2020) aims to capture major components of diagnostic judgements in pedagogical contexts. This framework distinguishes between internal aspects of the process, which are not directly measurable (e.g., person characteristics and diagnostic thinking), and external aspects that are observable (e.g., diagnostic situations and individuals' diagnostic behavior). According to the

DOI: https://doi.org/10.23770/RT1845

framework, diagnostic thinking includes perceiving, interpreting and decision making[1]. The process of perception depends on individual knowledge and therefore cannot easily be separated from interpretation. For the purposes of this article, we use the term "perceiving" to refer to both processes.

Most previous studies assessed the process of diagnostic thinking through processes that participants are aware of and able to verbalize. For example, Ostermann, Leuders, and Nückles (2017) showed that prospective teachers' judgements of task difficulty improved after an intervention on specific pedagogical content knowledge. However, the analysis of teachers' verbal responses could not specifically identify whether improved judgements were due to increased perception of relevant task characteristics. Wildgans-Lang, Scheuerer, Obersteiner, Fischer, and Reiss (2020) addressed this issue by using log data and verbal responses in a computer-based learning environment, in which prospective teachers selected mathematical tasks and evaluated virtual students' responses. Again, it was not totally clear which task features prospective teachers considered when selecting tasks and making their judgements. Reinhold (2018) and Philipp (2018) both conducted think-aloud interviews in order to determine participants' diagnostic strategies. However, this method does not allow for firm conclusions about internal and possibly unconscious perception processes. The method of eye tracking could be suitable to collect data that allows exploring perception processes during diagnostic judgements.

1.2 Using eye-tracking technology to assess diagnostic processes

Eye tracking is a method to record individuals' eye movements in real time. In recent years, eye tracking was increasingly used in mathematics education research to assess cognitive processes (Strohmaier, MacKay, Obersteiner, & Reiss, 2020). According to Just and Carpenter (1980) the method is based on two fundamental assumptions: (1) immediacy, which means that the perceived information is processed immediately, and (2) eye-mind-correspondence, which states that the eye remains fixed on an area of the stimulus until the information from that area is processed. Although these assumptions have been under debate and may not apply in all situations, they seem reasonable for studying perception of visually presented mathematical tasks (Schindler & Lilienthal, 2019b). Thus, eye tracking may also be suitable for assessing diagnostic judgement processes in mathematical tasks.

Eye tracking devices measure eye fixations and eye saccades. Fixations are phases of relative eye stagnation, usually around 200–300 ms, during which the brain begins to process visual information received from the eyes (Holmqvist et al., 2011). Saccades are rapid eye movements from one fixation to another, usually around 30–80 ms (Holmqvist et al., 2011), during which vision is extremely limited without detailed perception (Matin, 1974). Saccades and fixations alternate permanently.

[1] Note that Blömeke, Gustafsson and Shavelson (2015) conceptualize these three processes somewhat differently as situation-specific skills.

DOI: https://doi.org/10.23770/RT1845

To analyse eye-tracking data, different measures of eye fixations and saccades are used to identify specific cognitive processes (Holmqvist et al., 2011). *Global* eye-tracking measures allow observing eye movements to the whole stimulus (Strohmaier, Lehner, Beitlich, & Reiss, 2019). In contrast, *local* eye-tracking measures are used to get more detailed information about eye movements on specific parts of the stimulus (Holmqvist et al., 2011). For this purpose, the stimulus is divided into specific areas of interest (AOI) and eye movements on these AOIs are evaluated individually.

The following three eye-movement parameters can be used for *global* as well as *local* eye-movement analyses: The first two are the *fixation duration* and the *fixation duration average*. The fixation duration is the sum of duration of all fixations of an item (e.g. a mathematical task as a whole or a certain area within this task); the fixation duration average is the fixation duration divided by the number of fixations of an item. Both the fixation duration and the fixation duration average (on the whole item or on an AOI) are used as indicators for the depth of processing, where longer fixation durations and longer fixation duration averages tend to be linked to deeper cognitive processing (Holmqvist et al., 2011). For example, Unema and Rötting (1990) showed that participants' fixations were longer when processing more difficult mental calculations compared to simpler ones. A third important eye-tracking measure is the *number of fixations*, which is the number of all fixations of an item. The number of fixations is an indicator of how important a participant judges the stimulus or the AOI. People fixate on an object more often when they consider the object to be important (Holmqvist et al., 2011). More difficult items induce, in general, more fixations (Rayner, Pollatsek, Ashby, & Clifton Jr, 2012).

1.3 Domain-specific diagnostic processes: tasks with functions

Judging task difficulty requires knowledge about typical student errors, which are specific to a domain. In the domain of functions and graphs, typical student errors are well documented in the mathematics education literature (Hattikudur et al., 2012; Nitsch, 2015; Russell, O'Dwyer, & Miranda, 2009). Three typical student errors in this domain are the *graph-as-picture error*, the *slope-height confusion* and the *confusion of slope parameters and the intersection with the x-axis*. These three errors are the focus of the present study. We explain each error in the following.

When committing the graph-as-picture error, students interpret a graph as a real or geometric representation of a situation without considering the algebraic relationships (Nitsch, 2015; Russell et al., 2009). Figure 1 shows a sample task that may trigger the graph-as-picture error. This graph shows the speed of a race car during its second round on a circuit. The question is how many bends the driver has to take during each round. Students may consider the graph to represent the circuit rather than the car's change in speed. If they do so, their response would be four bends instead of two bends.

DOI: https://doi.org/10.23770/RT1845

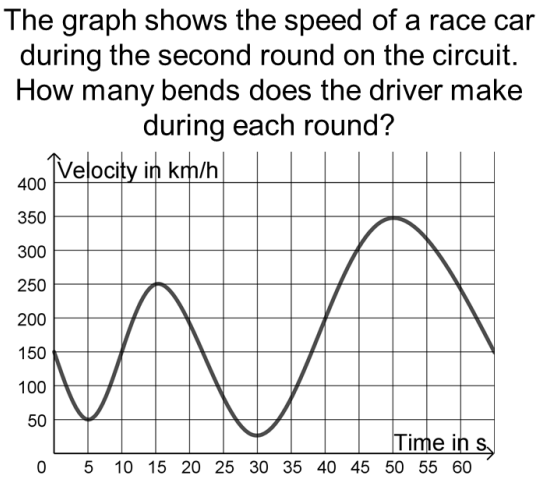

Fig. 1: Sample task for graph-as-picture error

The slope-height confusion means that students confuse the slope with the maximum value (the height) at a given point on the x-axis (Nitsch, 2015; Russell et al., 2009). The example in Figure 2 shows the distance covered by three runners. Students are asked to find the graph that represents the runner who was fastest at time t = 5 sec. Students who confuse the graph's slope with its height would choose the graph at the top (solid line) instead of the graph in the middle, which has the highest slope.

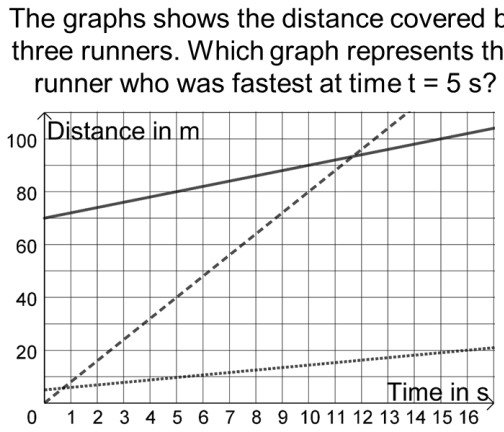

Fig. 2: Sample task for slope-height confusion

Confusion of slope parameters with intersection of the x-axis can occur in tasks in which students are presented with a linear graph and asked to determine or match the parameters in the linear equation $y = mx + b$. Rather than identifying m as the slope, students would confuse it with the intercept x (Nitsch, 2015). Figure 3 shows a sample task that may trigger the confusion of slope parameters with intersection of the x-axis. Students who confuse those two parameters would choose the first of the four given equations instead of the fourth, which would be the right answer.

DOI: https://doi.org/10.23770/RT1845

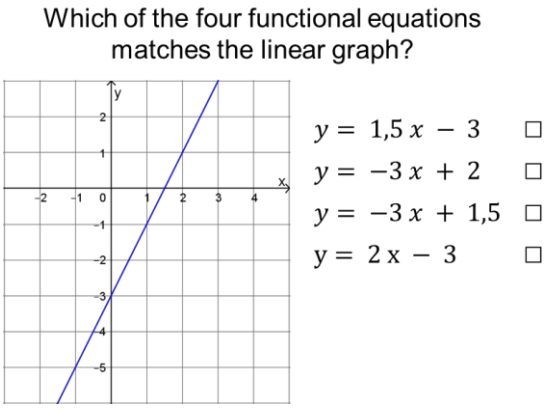

Fig. 3: Sample task for confusion of slope parameters with intersection of the x-axis

The three typical student errors described above are a suitable basis for developing tasks that may trigger these errors in students. Because the relevant task characteristics are visually perceptible, such tasks could be suitable for assessing judgement processes with eye tracking.

1.4 The present study

Our study aimed to assess diagnostic judgement processes while overcoming the methodological limitations of think-aloud interviews. We used the eye-tracking method, which has proven to allow assessing cognitive processes in prior research. Our major goal was to explore whether eye tracking is suitable to reliably assess the diagnostic process of perception (and interpretation see 1.1).

We constructed tasks on functions and graphs that either do or do not provoke one of the three well-documented typical student errors described above (see section 1.3). Functions and graphs are relevant in everyday life (newspapers, statistics, etc.) and they are an important topic of grade 8 and 9 mathematics curricula in Germany. The domain of functions and graphs also had the advantage that tasks could clearly be classified by visual task characteristics. Moreover, we aimed to develop tasks that were easy to understand and structured in a simple way, in order to reduce the complexity of eye-movement patterns. Most relevant for our eye-tracking method, it is possible to judge task difficulty through visual perception and correct interpretation of the relevant task characteristics. The tasks were designed such that there was always one area that provoked the typical student error (except for no-error tasks, see below) and another area that was relevant for determining the correct answer. Such a task design would allow us to identify relevant eye movements.

Our central assumption was that tasks that do provoke one of the typical student errors described above (hereafter: error tasks) can be differentiated from tasks that do not provoke any of the three student errors (hereafter: no-error tasks) on the basis of visually perceptible task characteristics. Error tasks and no-error tasks were created such that they differed in the relevant task characteristics only. Thus, error tasks should, in general, be more difficult for secondary school students to solve than no-error tasks.

DOI: https://doi.org/10.23770/RT1845

To analyse *local* eye movements, we defined "diagnostic AOIs" as the areas of error tasks that include relevant characteristics to provoke the respective typical student error. In order to judge task difficulty correctly, it was necessary to interpret the information from these AOIs correctly. As an example, the box in the sample task in Figure 4 displays the diagnostic AOI that reflects slope-height confusion. The AOI represents the position that students would pay most attention to when committing the slope-height confusion error, because they would look at the graph with the highest value (rather than the highest slope) at the time point t = 5 s.

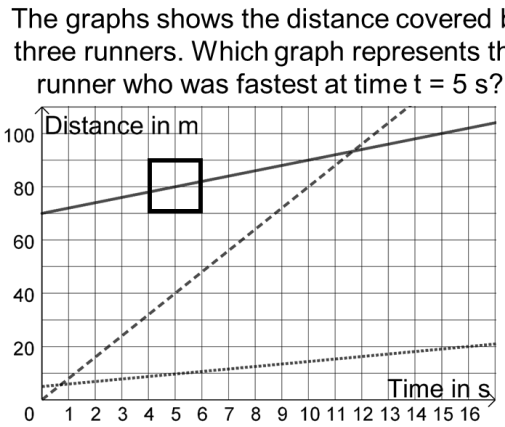

Fig. 4: Selected AOI in an exemplary graph

Participants were shown tasks on a computer screen and instructed to judge the difficulty of the presented task while their eye movements were recorded with an eye tracker. We addressed three issues:

First, we analysed if we can reliably detect the diagnostic process in participants' eye movements (assessed with the *global* eye-tracking parameters fixation duration, fixation duration average, and number of fixations). To that end, we analysed reliability coefficients of these scales across all items. To validate that the eye-tracking parameters were related to participants' perception of the relevant task characteristics, we compared error tasks and no-error tasks. We expected that error tasks, compared to no-error tasks, needed to be processed deeper and would thus require longer fixation durations. This assumption was in line with results from studies by Andrzejewska and Stolińska (2016) and Dewolf, Van Dooren, Hermens, and Verschaffel (2013), who found longer average fixation durations and a larger number of fixations during deeper cognitive processing. We also compared judgement accuracy between error and no-error tasks.

Second, we analysed if we could reliably detect participants' ability to identify task characteristics that provoke typical student errors with *local* eye-tracking parameters. We also investigated if these *local* eye-tracking parameters on the diagnostic AOIs correlated with participants' judgement accuracy. We expected that this would be the case because participants who show longer fixation durations and more fixations on the diagnostic AOI would be more likely to recognize the typical student error than participants who show shorter fixation durations and fewer fixations on the diagnostic AOI (Holmqvist et al., 2011; Unema & Rötting, 1990).

DOI: https://doi.org/10.23770/RT1845

Third, to improve our understanding of how participants made their judgements, we illustrate in a qualitative analysis of one sample item two participants' eye movements, their verbal reasoning and their judgements.

2 Methods

2.1 Sample

Participants were 51 prospective mathematics teachers (30 female, 21 male; age: *M* = 23.43 years, *SD* = 2.07). The majority of participants (73%) were in their 5th or 6th semester of teacher training at a university in Germany. According to their study programs and their self-reports, none of the participants had previous experience with typical student errors in function problems. However, we can assume that all participants had acquired some general content knowledge and general pedagogical content knowledge during their teacher training. We removed four participants from the original sample of 55 due to a technical error and data loss during the eye-tracking session. Participation was voluntary, and all participants were informed about the study procedure before the study began.

2.2 Tasks

We constructed 20 tasks about functions and graphs that did or did not provoke one of the three typical student errors described above (see section 1.3): graph-as-picture error, slope-height confusion, and confusion of slope parameters with intersection of the *x*-axis.

Tab. 1: Overview of the numbers of items per typical student error

Typical student error	Items
No-error tasks	4
graph-as-picture error	4
slope-height-confusion	4
confusion of slope parameters with intersection of the x-axis (multiple-choice task format)	4
confusion of slope parameters with intersection of the x-axis (open response task format)	4

Table 1 provides an overview of the 20 items used in this study. Four items were no-error tasks. Two of these items were similar to the graph-as-picture error tasks, and two were similar to the slope-height-confusion error tasks, except that the no-error tasks included different questions that did not provoke the respective errors. The two typical student errors graph-as-picture and slope-height confusion were represented with four items each. For the error confusion of slope parameters with intersection of the x-axis, there were eight items: four in multiple-choice format and another four with an open response format. Note that for these items, it was not possible to create corresponding

DOI: https://doi.org/10.23770/RT1845

no-error tasks because these errors can occur in any task of that type and their occurrence does not depend on the specific question. We designed the error tasks in a way that there was always just one area which provoked the typical student error.

Figure 5 shows an example of a no-error task. The graphs in this example show the completed distances of three runners as a function of elapsed time. The question is at which time point another runner passed the runner represented by the graph with the solid line. This example is a no-error task because none of the three typical errors seems to prevent finding the correct answer (10s). Figures 1, 2 and 3 (see 1.3) show samples for error tasks.

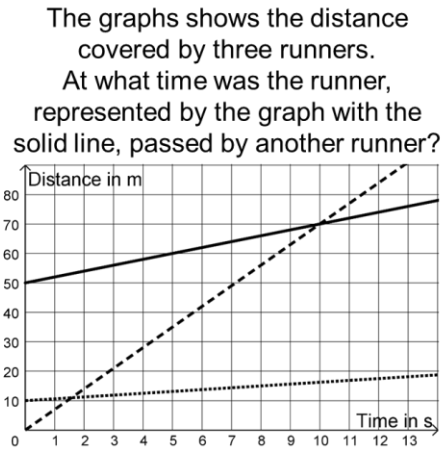

Fig. 5: Example of a no-error task

Our rationale was that error tasks would be more difficult for secondary school students to solve than no-error tasks, just because they include the relevant task characteristic, which may provoke the typical error. We validated this assumption empirically by presenting the tasks to a sample of $N = 110$ students of grade 8 from four secondary schools in Germany. As expected, the mean solution rates for error tasks were fairly low (M = 34%, SD = 14), while those for no-error tasks were extremely high (M = 98%, SD = 2); the difference in mean solution rates was large and highly significant, $t(18) = –8.55$, $p < .001$, Cohen's $d = 6.14$. Therefore, based on theoretical considerations and empirical evidence, we classified no-error tasks as "very easy" for secondary school students, and error tasks as "rather difficult". Participants' judgement of task difficulty was also consistent with this classification of task difficulty (see 2.5.2).

2.3 Equipment

We recorded participants' eye movements with a remote eye tracker (SMI 250) with a sampling rate of 250 Hz. The participants were seated at a distance of 65–70 cm in front of a 24-inch screen with a resolution of 1920 x 1080 px. Since the eye tracker is sufficiently robust towards head movements, we did not use a chin rest or other head fixations. However, we asked participants to avoid head movements as much as possible. For the data collection and task presentation, we used the manufacturer's software SMI iView X and SMI Experiment Center. For exporting the data, we used SMI BeGaze.

2.4 Procedure

DOI: https://doi.org/10.23770/RT1845

Before the eye-tracking session, we asked participants to complete a questionnaire about demographic information. Afterwards we informed the participants about the eye-tracking procedure and showed them a sample item. We used a five-point calibration to obtain the accuracy between markers and gaze position. The calibration was repeated until we achieved a deviation of less than 0.5°, which is the maximum deviation required for accurate calculation of fixation durations (Holmqvist et al., 2011). For three participants, a deviation less than 0.5° could not be achieved and a calibration of 0.6° was accepted. The 20 tasks were presented to the participants one by one and in the same order on the computer screen. The participants were instructed to judge the difficulty of the presented task on a scale from 1 to 4 (1 meaning very easy, 2 fairly easy, 3 fairly difficult, and 4 very difficult) and to give their responses verbally. There was no time limit. Once the participants had made their judgement, the experimenter ended the recording of the eye movements by pressing a button. After pressing the button, the task remained visible on the screen, so the participants were able to give a reason for their judgement. These responses were audio-recorded but eye movements were not recorded during this phase. For the purpose of this study, we focused on the analysis of eye movements and participants' judgements of task difficulty, and did not systematically analyse participants' verbal reasoning. At the end of the eye-tracking session, we asked the participants to solve the 20 tasks on paper to assess their ability to solve the tasks correctly. Participants solved 90% of the tasks correctly, which shows that the participants had sufficient knowledge of the content, and suggests that their eye movements were related to diagnostic processes rather than to difficulty in solving the tasks.

2.5 Measures

We focussed specifically on the diagnostic process of perception, which should be accessible through eye movements. We also analysed the relation between perception and decision making. The effect sizes of the following calculations are interpreted according to Cohen (1992).

2.5.1 Perceiving

We measured the process of perceiving on the basis of various eye-movement parameters. *Global* parameters were the total fixation duration (in ms) and the fixation duration average (in ms) as indicators for the depth of processing. The number of fixations is considered to be an indicator for the general importance of an object. *Local* parameters were those related to diagnostic AOIs. The parameters were the same as the *global* parameters but were measured for the AOI in relation to the total parameters per item and AOI size. The AOI size varied from task to task and the fixation duration differed from participant to participant. To create a scale across all tasks and all participants, we used the relative values as follows. The fixation duration on the AOI [%] and the fixation duration average on the AOI were divided by the size of AOI; the number of fixations on the AOI was divided by the total number of fixations.

2.5.2 Decision making

DOI: https://doi.org/10.23770/RT1845

Judgement of task difficulty was used as an indicator for decision making. Participants gave their judgements on a scale from 1 to 4 (1 = very easy, 4 = very difficult, see 2.4). A *t*-test for paired samples revealed a highly significant difference in judgement between error tasks (M = 2.37, SD = .38) and no-error tasks (M = 1.44, SD = .28), (t = 16.60, p < .001), with error-tasks being judged as significantly more difficult. The effect size was d = 2.55 and thus represented a large effect. Thus, participants' decisions were in accordance with our classification of error tasks being more difficult for secondary school students than no-error tasks (see 2.2), even though the difference in participants' judgements between error tasks and no-error tasks was less pronounced than the empirical difference. For the current analyses, we were interested in participants' judgement accuracy. For error tasks, judgements of difficulty 3 and 4 were coded as correct. For no-error tasks, judgements of difficulty 1 and 2 were coded as correct. For each task, participants received 1 point for a correct and 0 points for an incorrect judgement. Therefore, across all 20 tasks, the maximum score was 20. The internal consistency of this scale was acceptable with Cronbach's alpha = .60 (M = 10.75, SD = 2.91).

3 Results

3.1 Global eye movements

We first examined whether it was possible to establish reliable scales across all tasks for *global* eye-tracking parameters, which refer to the tasks as a whole. For all parameters, Cronbach's alpha was high: for fixation duration: .93 (M = 369.23 s, SD = 119.55); for fixation duration average: .97 (M = 5.28 s, SD = .90); for number of fixations: .93 (M = 1438.04, SD = 447.31). These results show that there is a high internal consistency in participants' *global* eye movements across all tasks.

Further, we were interested in whether there were differences between error tasks and no-error tasks regarding eye-tracking parameters. The descriptive statistics for these tasks are shown in Table 2. *T*-tests for paired samples revealed a highly significant difference in fixation duration (t = –4.29, p < .001), fixation duration average (t = –8.35, p < .001) and number of fixations (t = –2.38, p = .021) between error tasks and no-error tasks. Error tasks were fixated longer, had a higher fixation duration average, and a higher number of fixations. The effect sizes were between d = 0.27 and d = 0.56 and thus represented small to medium effects.

In sum, the data suggests that fixation duration, fixation duration average, and number of fixations are reliable eye-movement parameters. The data also shows that these parameters are sensitive to whether or not the task included the relevant task characteristic for the typical student error.

Tab. 2: Descriptive statistics for the global eye-movement parameters, for error tasks and no-error tasks

Scale	Task type	N	*M*	*SD*
Fixation duration [s]	Error-task	51	19.26	6.68
	No-error task	51	16.42	4.97

DOI: https://doi.org/10.23770/RT1845

Fixation duration average [ms]	Error-task	51	268	46
	No-error task	51	245	40
Number of fixations	Error-task	51	72.97	24.18
	No-error task	51	67.63	19.24

Note: M = mean; SD = standard deviation

3.2 Local eye movements

The reliabilities of the four *local* eye-movement parameters were calculated across the 16 error-tasks. Mean values and standard deviations are shown in Table 3. The Fixation duration on the AOI [%] in relation to size of AOI [%] showed a sufficiently high Cronbach's alpha value of .63. The reliability for fixation duration average [ms] on the AOI in relation to size of AOI [%] with .50 was acceptable (Wirtz, 2020). However, the reliability for the number of fixations on the AOI in relation to the total number of fixations was not acceptable: -.36.

To evaluate how local eye movements relate to participants' judgement accuracy, we calculated correlations between local eye-tracking parameters and average accuracy of judgements. The findings are displayed in Table 3. Judgement accuracy correlated significantly with relative fixation duration on the AOI (r = .32, p = .022). The longer a participant looked at the AOI the more accurate he or she judged task difficulties, this was a medium effect. There were no significant correlation between fixation duration average on the AOI and judgement accuracy. We did not calculate the correlation of number of fixations and judgement due to this scale's low reliability.

***Tab. 3**: Descriptive statistics for error-tasks: judgement accuracy, local eye-tracking parameters, and correlations with judgement accuracy*

	M	*SD*	Correlation with judgement accuracy
Judgement accuracy	.44	.18	
[1 = correct; 0 = incorrect]			
Fixation duration on the AOI [%] in relation to size of AOI [%]	9.85	3.69	.32*
Fixation duration average [ms] on the AOI in relation to size of AOI [%]	321.97	96.67	.08

Note: M = mean; SD = standard deviation; * indicates p < .05

3.3 Qualitative analyses

In order to obtain more detailed information about eye-movement patterns and, to illustrate what these data mean on the individual level, we explored the data in a qualitative way. Overall, eye-movement patterns were fairly similar across participants. In most cases, participants first read the text from beginning to end. Afterwards they looked at individual graph components in quick succession (axes, lines, equations).

DOI: https://doi.org/10.23770/RT1845

However, for the subsequent steps, we identified two diverse patterns: Some participants looked back and forth between different areas more frequently, while others focused their attention on a specific area, which was often the area that included information about the correct solution or the diagnostic AOI.

In the following, we contrast two participants' (A and B) eye movements on a task that provoked the slope-height confusion. The task is shown in Figure 2 (see 1.3). Both participants had the required content knowledge, since they solved the task correctly in the paper-pencil test. The two participants differ in their *local* eye movements (e.g. fixation durations; see heatmaps in figure 6), the accuracy of their judgements and their reasoning. While participant A gave a correct judgment, participant B did not.

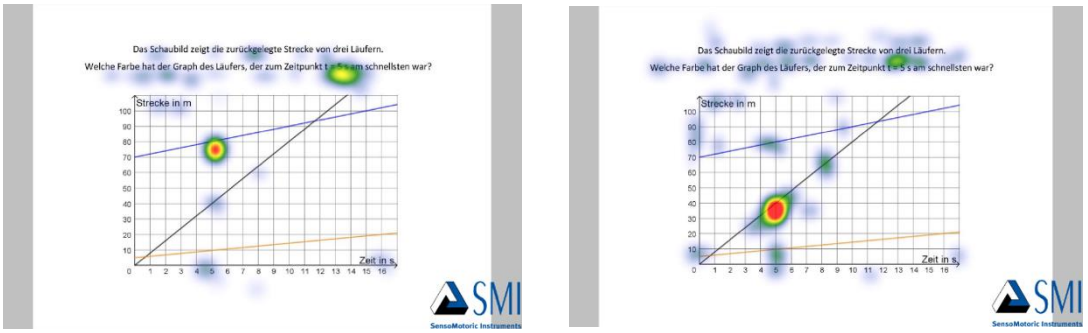

Fig. 6: Heatmaps of participant A (left) and participant B (right); red areas indicate long fixation durations

Participant A had a fixation duration of approximately 20% on the diagnostic AOI that is related to the typical student error, which is clearly visible in the heatmap (see red area in Figure 6 left). This participant also had the longest fixation duration average (450 ms) on that area. In addition, participant A looked at this area three times more often (number of fixations = 6) than at the area that is relevant for the correct solution (number of fixations = 2). The participant judged the task as difficult (which was correct) and referred in his reasoning to the typical error of slope-height confusion (which was also correct).

In contrast, participant B had a very short fixation duration of about 4% on the diagnostic AOI that is related to the typical student error. The red area in the heatmap (figure 6 right) shows that this participant had a high fixation duration of approximately 30% on the AOI that showed the correct solution. This participant also had the longest fixation duration average (490 ms) on the area that is relevant for the correct answer. In addition, participant B looked at this area five times more often (number of fixations = 11) than at the area showing the typical error (number of fixations = 2). The participant judged the task as being easy (which was wrong) and did not refer to the typical error in his or her reasoning.

In sum, participant A, who judged the task correctly and referred to the typical student error, looked longer and more often to the area which showed the typical student error than participant B who did not judge the task correctly.

DOI: https://doi.org/10.23770/RT1845

4 Discussion

Adaptive teaching requires accurate diagnostic judgements (Hardy et al., 2019; Parsons et al., 2018). An essential aspect is the ability to correctly judge task difficulty (Corno, 2008; Stein et al., 1996; Sullivan et al., 2012). The aim of this study was to explore whether eye tracking is a suitable method for assessing perception processes during diagnostic judgements. We were also interested in how eye-tracking parameters were related to the accuracy of the diagnostic judgement. Prospective mathematics teachers judged the difficulty of linear function tasks, some of which included characteristics that are well known to be challenging for students (error tasks), while others did not include these characteristics (no-error tasks). During their judgements, we recorded participants' eye movements.

The results showed high internal consistency of global eye-tracking parameters. Thus, global eye-movement parameters seem to be comparable across specific tasks and across the three typical error types considered in this study. This is in line with results of the study by Strohmaier et al. (2019) who also found high internal consistency for global eye-tracking data across several items on mathematical word problems.

Participants showed longer fixation durations as well as longer average fixations on error tasks than on no-error tasks. This result supports the assumption that global eye-movement parameters were sensitive to the presence of relevant task characteristics, and thus supports the validity of the measures. It also shows that judging error tasks requires deeper cognitive processing of the respective task characteristics. In addition, the number of fixations was higher for error tasks than for no-error tasks. This could mean that participants tried to extract as much information as possible from the task and that this process required more comparisons between the task characteristics in error tasks. Again, these results are in line with previous eye-tracking studies (e.g. Strohmaier et al., 2019; Unema & Rötting, 1990), which showed that participants made longer fixations while processing more difficult items compared to processing simpler items.

We identified local eye-movement parameters (fixation duration and fixation duration average on diagnostic AOI) that showed sufficient internal consistency and were thus available to analyse correlations with judgement accuracy. This result shows that these items can be used for further eye-tracking research in the area of diagnostic competences.

Only one out of the three local parameters was correlated to judgement accuracy. Participants who showed longer fixation duration on the typical error were significantly more accurate in their judgements. This suggests that local eye movements may indicate participants identifying relevant task characteristics and that such identification is connected with deeper cognitive processes. Although this interpretation of our data seems plausible, eye-tracking data do not provide sufficient information whether participants actually interpreted the identified task characteristics as provoking typical student errors. To address this issue, it would be worthwhile to systematically link participants' verbal reasoning with their eye movements. The present study provides an empirical

DOI: https://doi.org/10.23770/RT1845

foundation for such analysis by clarifying which eye-movement parameters can be used as reliable measures.

Our qualitative analysis showed that the ability to identify task characteristics relevant to task difficulty is reflected in eye movements. The two sample eye-movement patterns illustrated the different depth of processing of either the area that included the correct task solution or the diagnostic area that included the typical student error.

One reason why the local eye-movement patterns did not always show hight internal consistency could be that participants had no specific experience with typical student errors on function tasks. Low correlations of local eye movements with judgement accuracy may also be explained by the low variance in the sample. However, the qualitative analysis of two individuals showed that it might be possible to relate specific eye-movement patterns to judgement accuracy and participants' reasoning. Further research is certinaly needed in this regard. In further research, one could also increase the variance in participants' domain-specific pedagogical content knowledge, by assigning some participants to an intervention in which they receive instruction about typical student errors. This would also allow for causal inferences about the role that certain knowledge facets play for eye-movement patterns.

The findings of this study are highly specific in the sense that they relate to a very specific task type (graphical tasks related to mathematical functions). As such, our study contributes to current discussions about the use of eye tracking in mathematics educational research (Strohmaier et al., 2020). However, similar research could be conducted in another mathematical topic or in a different subject area as well. This would require a subject-specific analysis of relevant student errors and task types that reveal these errors. Using eye tracking requires topics and corresponding tasks which allow *localizing* certain pieces of information that constitute student thinking in a *graphical way* (Schindler & Lilienthal, 2019a). In mathematics, such a localization is often possible, for example in geometry tasks or – as in our current study – a graphical representation of a quantitative relation of variables (i.e., a mathematical function). In other subjects such as science, such localizations may be possible in graphical displays of natural phenomena or natural laws with which students have to generate descriptions or explations (Klein, Viiri, & Kuhn, 2019). In subjects which rely on the interpretation of texts, one could use reading tasks or text-picture combinations in which certain pieces of information generate difficulty or need to be integrated (Beitlich et al., 2014; Schreiter, Vogel, Rehm, & Dörfler, accepted). Furthermore, in order to test hypotheses on information processing, our study has demonstrated the necessity to systematically vary task situations in a controlled way and predict effects on eye movements. Such approaches could advance subject-specific theories on teachers' diagnostic judgment.

In conclusion, the present study suggests that eye-tracking technology may complement traditional methods for assessing participants' perception during diagnostic processes.

DOI: https://doi.org/10.23770/RT1845

Funding: This research is part of the graduate school "DiaKom", funded by the Ministry of Science, Research and the Arts in Baden-Wuerttemberg, Germany.

References

Andrzejewska, M., & Stolińska, A. (2016). Comparing the Difficulty of Tasks Using Eye Tracking Combined with Subjective and Behavioural Criteria. *Journal of Eye Movement Research*, *9*(3). https://doi.org/10.16910/jemr.9.3.3

Artelt, C., & Rausch, T. (2014). Accuracy of teacher judgments: When and for what reasons? In S. Krolak-Schwerdt, S. Glock, & M. Böhmer (Eds.), *The future of education research: Volume 03. Teachers' professional development: Assessment, training, and learning* (pp. 27–43). Rotterdam, Boston, Taipei: Sense Publishers.

Beitlich, J. T., Obersteiner, A., Moll, G., Mora Ruano, J. G., Pan, J., Reinhold, S., & Reiss, K. (2014). The role of pictures in reading mathematical proofs: An eye movement study. In Liljedahl, Peter (ed.) et al. (Ed.), *Proceedings of the 38th conference of the International Group for the Psychology of Mathematics Education and the 36th Conference of the North American Chapter of the Psychology of Mathematics Education: Vancouver, Canada, July 15-20,2014* (pp. 121–128). Belo Horizonte, Brazil: International Group for the Psychology of Mathematics Education.

Bromme, R. (1981). *Das Denken von Lehrern bei der Unterrichtsvorbereitung: eine empirische Untersuchung zu kognitiven Prozessen von Mathematiklehrern*: Beltz.

Chapman, O. (2014). Overall Commentary: Understanding and Changing Mathematics Teachers. In J.-J. Lo, K. R. Leatham, & L. R. van Zoest (Eds.), *Research in Mathematics Education. Research Trends in Mathematics Teacher Education* (pp. 295–309). Cham, s.l.: Springer International Publishing. https://doi.org/10.1007/978-3-319-02562-9_16

Corno, L. (2008). On Teaching Adaptively. *Educational Psychologist*, *43*(3), 161–173. https://doi.org/10.1080/00461520802178466

Dewolf, T., Van Dooren, W., Hermens, F., & Verschaffel, L. (Eds.) (2013). *Do students attend to and profit from representational illustrations of non-standard mathematical word problems?*: PME; Kiel, Germany.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*, 215.

Hardy, I., Decristan, J., & Klieme, E. (2019). Adaptive teaching in research on learning and instruction. *Journal for Educational Research Online*, *11*, 169–191.

Hattikudur, S., Prather, R. W., Asquith, P., Alibali, M. W., Knuth, E. J., & Nathan, M. (2012). Constructing graphical representations: Middle schoolers' intuitions and developing knowledge about slope and y-intercept. *School Science and Mathematics*, *112*, 230–240.

Helmke, A., & Schrader, F.-W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. *Teaching and Teacher Education*, *3*(2), 91–98. https://doi.org/10.1016/0742-051X(87)90010-2

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures* (First edition). Oxford, New York, Auckland: Oxford University Press.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*, 329.

Karst, K. (2012). *Kompetenzmodellierung des diagnostischen Urteils von Grundschullehrern*: Waxmann Verlag.

Klein, P., Viiri, J., & Kuhn, J. (2019). Visual cues improve students' understanding of divergence and curl: Evidence from eye movements during reading and problem solving. *Physical Review Physics Education Research*, *15*(1). https://doi.org/10.1103/PhysRevPhysEducRes.15.010126

Loibl, K., Leuders, T., & Dörfler, T. (2020). A Framework for Explaning Teachers' Diagnostic Judgements by Cognitive Modeling (DiaCoM). *Teaching and Teacher Education*.

Matin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin*, *81*(12), 899–917. https://doi.org/10.1037/h0037368

McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W., Ullrich, M. (2009). Diagnostische Fähigkeiten von Lehrkräften. *Zeitschrift Für Pädagogische Psychologie*, *23*, 223–235.

Nitsch, R. (2015). *Diagnose von Lernschwierigkeiten im Bereich funktionaler Zusammenhänge*. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-10157-2

Ostermann, A., Leuders, T., & Nückles, M. (2017). Improving the judgment of task difficulties: prospective teachers' diagnostic competence in the area of functions and graphs. *Journal of Mathematics Teacher Education*, *3*(5), 175. https://doi.org/10.1007/s10857-017-9369-z

Parsons, S., Vaughn, M., Scales, R., Gallagher, M., Parsons, A., Davis, S., Allen, M. (2018). Teachers' Instructional Adaptations: A Research Synthesis. *Review of Educational Research*, *88*(2), 205–242. https://doi.org/10.3102/0034654317743198

Philipp, K. (2018). Diagnostic Competence of Mathematics Teachers with a View to Processes and Knowledge Recources. In T. Leuders, K. Philipp, & J. Leuders (Eds.), *Mathematics Teacher Education: Vol. 11. Diagnostic Competence of Mathematics Teachers: Unpacking a Complex Construct in Teacher Education and Teacher Practice*. Cham: Springer International Publishing.

Rayner, K., Pollatsek, A., Ashby, J., & Clifton Jr, C. (2012). *Psychology of reading*: Psychology Press.

Reinhold, S. (2018). Revealing and Promoting Pre-service Teachers' Diagnostic Strategies in Mathematical Interviews with First-Graders. In T. Leuders, K. Philipp, & J. Leuders (Eds.), *Mathematics Teacher Education: Vol. 11. Diagnostic Competence of Mathematics Teachers: Unpacking a Complex Construct in Teacher Education and Teacher Practice*. Cham: Springer International Publishing.

Russell, M., O'Dwyer, L. M., & Miranda, H. (2009). Diagnosing students' misconceptions in algebra: Results from an experimental pilot study. *Behavior Research Methods*, *41*, 414–424.

Schindler, M., & Lilienthal, A. J. (2019a). Students' Creative Process in Mathematics: Insights from Eye-Tracking-Stimulated Recall Interview on Students' Work on Multiple Solution Tasks. *International Journal of Science and Mathematics Education*, 1–22.

Schindler, M., & Lilienthal, A. J. (2019b). 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*, *1*, 33.

Schrader, F.-W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts*: Lang.

Schreiter, S., Vogel, M., Rehm, M., & Dörfler, T. (accepted). Teachers' diagnostic judgment regarding the difficulty of fraction tasks: A reconstruction of perceived and processed task characteristics. *RISTAL*.

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building Student Capacity for Mathematical Thinking and Reasoning: An Analysis of Mathematical Tasks Used in Reform Classrooms. *American Educational Research Journal*, *33*(2), 455–488. https://doi.org/10.3102/00028312033002455

Strohmaier, A. R., MacKay, K. J., Obersteiner, A., & Reiss, K. (2020). Eye tracking methodology in mathematics education research: a systematic literature review. *Educational Studies in Mathematics*.

Strohmaier, A. R., Lehner, M. C., Beitlich, J. T., & Reiss, K. M. (2019). Eye Movements During Mathematical Word Problem Solving—Global Measures and Individual Differences. *Journal Für Mathematik-Didaktik*.

DOI: https://doi.org/10.23770/RT1845

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*(3), 743–762. https://doi.org/10.1037/a0027627

Sullivan, P., Clarke, D., & Clarke, B. (2012). *Teaching with tasks for effective mathematics learning* (Vol. 9): Springer Science & Business Media.

Unema, P., & Rötting, M. (1990). *Differences in Eye Movements and Mental Workload between Experienced and Inexperienced Drivers In: D. Brogan (Ed.), Visual Search*: London, Taylor & Francis.

Waern, Y. (1988). Thoughts on text in context: Applying the think-aloud method to text processing. *Text-Interdisciplinary Journal for the Study of Discourse*, *8*, 327–350.

Wildgans-Lang, A., Scheuerer, S., Obersteiner, A., Fischer, F., & Reiss, K. (2020). Analyzing prospective mathematics teachers' diagnostic processes in a simulated environment. *ZDM*, *52*(2), 241–254. https://doi.org/10.1007/s11858-020-01139-9

Kirsten Brunner

worked as a secondary school teacher for the subjects Mathematics, Technics and Physics. Since 2018 she is working on her doctoral thesis in the research training group "DiaKom" (Diagnostic Judgements by Cognitive Modeling) at the University of Education in Freiburg, Germany.

Andreas Obersteiner

is a professor for mathematics education at the Technical University of Munich, Germany. His research focuses on the cognitive processes of mathematical thinking and learning, and on teachers' diagnostic competencies.

Timo Leuders

is a professor for mathematics education at the University of Education in Freiburg, Germany with a research focus on teaching and learning in secondary education and teacher professionality. He is co-speaker of the research training group "DiaKom" within which this research was conducted.

DOI: https://doi.org/10.23770/RT1845