Educational Studies in Mathematics (2024) 115:177–196
https://doi.org/10.1007/s10649-023-10281-6

How pedagogical content knowledge sharpens prospective teachers' focus when judging mathematical tasks: an eye-tracking study

Kirsten Brunner[1] · **Andreas Obersteiner**[1,2] · **Timo Leuders**[1]

Accepted: 22 November 2023 / Published online: 19 December 2023

Abstract
Teachers' ability to accurately judge difficulties of mathematical tasks is an essential aspect of their diagnostic competencies. Although research has suggested that pedagogical content knowledge (PCK) is positively correlated with the accuracy of diagnostic judgments, experimental studies have not been conducted to investigate how PCK affects perception and interpretation of relevant task characteristics. In an intervention study with a control group, 49 prospective mathematics teachers judged the difficulty of 20 tasks involving functions and graphs while an eye tracker tracked their eye movements. Some of the tasks included characteristics well known to be difficult for students. Participants' domain-specific PCK of typical student errors was manipulated through a three-hour intervention, during which they learned about the most common student errors in function and graph problems. We found that the process of perception (relative fixation duration on the relevant area in the tasks) was related to judgment accuracy. Pre-post comparisons revealed an effect of the intervention not only on participants' domain-specific PCK of typical student errors but also on their perception and interpretation processes. This result suggests that domain-specific PCK of typical student errors allowed participants to focus more efficiently on relevant task characteristics when judging mathematical task difficulties. Our study contributes to our understanding of how professional knowledge makes teachers' judgment processes of mathematical tasks more efficient.

Keywords Task diagnosis · PCK · Diagnostic judgment · Eye tracking · Functions and graphs · Task difficulty

✉ Kirsten Brunner
Kirsten.Brunner@ph-freiburg.de

Andreas Obersteiner
Andreas.Obersteiner@tum.de

Timo Leuders
Leuders@ph-freiburg.de

[1] University of Education Freiburg, Kunzenweg 21, 79117 Freiburg, Germany

[2] Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany

1 Introduction

Accurate diagnostic judgments are an essential component of adaptive teaching (Hardy et al., 2019; Krolak-Schwerdt et al., 2014; Parsons et al., 2018). Diagnostic judgments include judgments of learner performance, task difficulty, and student solutions (Karst et al., 2017; Schrader, 1989). Previous studies have primarily focused on the quality of diagnostic judgments or teachers' activities associated with judgments (Herppich et al., 2018; Hoge & Coladarci, 1989; Karst, 2012; Leuders et al., 2022; McElvany et al., 2009; Südkamp et al., 2012). However, there is limited knowledge about the cognitive processes underlying diagnostic judgments and how pedagogical content knowledge (PCK) affects these processes (Herppich et al., 2018; Loibl et al., 2020; Schrader, 2009). Specifically, we lack an understanding of how PCK affects the perception of relevant task characteristics while judging task difficulty (Herppich et al., 2018; Loibl et al., 2020; McElvany et al., 2009; Rieu et al., 2022; Schrader, 2009). This knowledge gap is unfortunate because researchers agree that knowledge about the information processing of diagnosing is necessary to better understand judgments in pedagogical contexts (Herppich et al., 2018; Karst & Bonefeld, 2020; Loibl et al., 2020; Rieu et al., 2022; Schrader et al., 2018). In particular, knowledge about what information teachers perceive in a diagnostic situation and how they process this information is crucial for understanding how teachers come to their judgments (Carrillo-Yañez et al., 2018; Loibl et al., 2020). Understanding teachers' diagnostic processes is also a basis for providing adequate support during teacher training. In this study, we investigated prospective teachers' diagnostic processes when judging mathematical task difficulties. We specifically investigated how prospective teachers' domain-specific PCK of typical student errors affects their perception and interpretation of relevant task characteristics. We used eye-tracking methodology to measure visual perception processes.

1.1 Task judgment as a diagnostic process

Tasks are crucial tools for teaching mathematics (Sullivan et al., 2012). Teachers should be able to judge the overall difficulty of the tasks they select for their students. They should also be able to interpret how students will solve the tasks and what difficulties they will encounter while solving them (Chapman, 2014; Hardy et al., 2019; Kron et al., 2021; McElvany et al., 2009; Mellone et al., 2020; Philipp, 2018). Task judgments are considered a predictor of student solution processes and results (Leuders & Loibl, 2021).

During task judgment, teachers have to apply their knowledge to the information in the task (Brunner et al., 2021; Loibl et al., 2020; Rieu et al., 2022). Loibl et al. (2020) developed a framework describing teachers' diagnostic judgments by cognitive modeling (DiaCoM) to interpret the cognitive processes underlying task judgments. The DiaCoM framework conceptualizes diagnostic judgments in educational contexts as inferences made by individuals (e.g., teachers) about other individuals (e.g., students) or things (e.g., tasks) based on the information provided in a diagnostic situation (Loibl et al., 2020).

The framework distinguishes between invisible internal factors and externally observable factors. The internal factors are personal characteristics and diagnostic thinking (DT), while the external factors are the diagnostic situation (i.e., situation characteristics [SCs]) and diagnostic behavior (DB). We used the DiaCoM framework to describe the invisible internal diagnostic steps and relate them to externally observable indicators. Figure 1

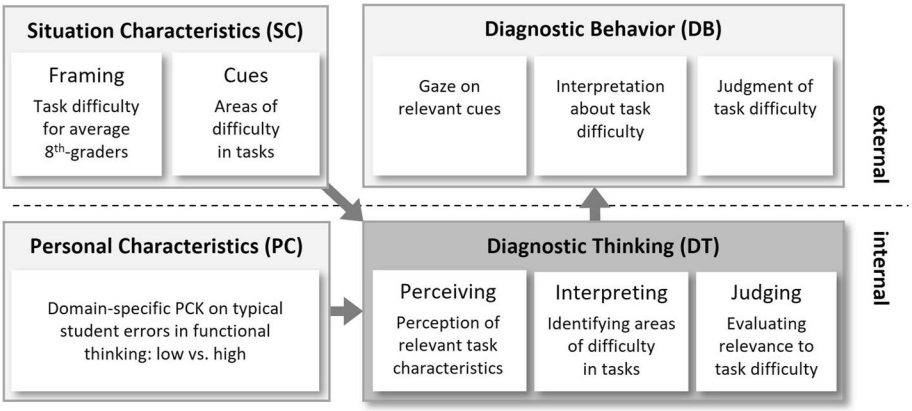

Fig. 1 DiaCoM framework (Loibl et al., 2020) specified for the situation of judging the difficulty of graphical tasks

shows the DiaCoM framework specified for the diagnostic situation we investigated in the present study: judging the difficulty of tasks about functions and graphs.

1.1.1 Situation characteristics and personal characteristics

As the figure shows, the situation characteristics (SCs; top left box in Fig. 1) are externally observable factors, including the information that initiates the judgment process. In our case, the situation characteristic was that participants were asked to judge how difficult specific tasks on functions and graphs were for average 8th-graders. The tasks contained specific cues, i.e., characteristics that are well known to affect task difficulty.

The personal characteristics (PC; lower left box in Fig. 1) describe the knowledge that underlies diagnostic thinking. This knowledge (e.g., domain-specific PCK) can be manipulated to experimentally test the influence of knowledge on diagnostic thinking. This experimental manipulation assumes that an intervention leads to higher knowledge. If an intervention group shows a significant improvement in task judgment compared to a control group, then the assumption that knowledge from the intervention was the reason for the improvement is validated (Loibl et al., 2020).

1.1.2 Diagnostic thinking and diagnostic behavior

According to the framework, diagnostic thinking (DT; lower right box in Fig. 1) includes three processes, namely perception, interpretation, and judging (Bromme, 1981; Loibl et al., 2020). These processes do not necessarily occur in a sequence and may partially overlap (e.g., interpretation may occur during perception). *Perception* involves observing the task at hand and the cues it contains. Regarding diagnostic behavior (DB; top right box in Fig. 1), perception can be externally observed, for example, by analyzing participants' eye movements (Brunner et al., 2021). *Interpretation* involves classifying perceived cues as relevant or irrelevant to task difficulty (Brunner et al., 2021; Mellone et al., 2020; Rieu et al., 2022). This is the basis for analyzing the specific difficulties in a task and is reflected in participants' interpretation of task difficulty. *Judging* refers to the final judgment of a task's difficulty in relation to another task or a group of students. This step is reflected

in the overall task judgment, such as whether the task is easy or difficult (Bless & Greifeneder, 2017; Südkamp et al., 2012).

Diagnostic behavior, an externally observable indicator, allows making inferences about diagnostic thinking (illustrated by the arrow from DT to DB). Note that eye-tracking methodology allows assessing perception processes in real-time, whereas interpretation and judgment can only be assessed after the processes have finished. In this sense, we consider participants' interpretation and overall judgment as products of interpretation and judging processes, respectively.

Recently, several researchers have used the DiaCoM framework to describe the steps of diagnostic thinking when judging the difficulty of mathematical tasks. Rieu et al. (2022) analyzed the influence of PCK on the accuracy of task judgments by asking participants to decide which of two tasks (word problems containing fractions) was more difficult. They showed that specific PCK (i.e., knowledge about problems in fractions) is necessary for making accurate task judgments (Rieu et al., 2022). However, the authors could not distinguish between the perception of relevant task characteristics and subsequent judging processes. In our previous eye-tracking study, we found that the perception of relevant task characteristics is related to the accuracy of judgments (Brunner et al., 2021). Participants who looked longer and more frequently at relevant task characteristics judged the tasks more accurately. However, it remained unclear whether different perceptions of relevant task characteristics were due to participants' eye-gaze patterns or to differences in their specific PCK.

Studies that have examined teachers' knowledge of the quality of diagnostic judgments (Südkamp et al., 2012) have mainly focused on whether teachers can rank task difficulty or decide which of two tasks is more difficult (Becker et al., 2020; Ostermann, 2018; Rieu et al., 2022). These studies assume that teachers reason about how their students solve tasks. However, they do not specify why teachers reason a task is challenging or *what difficulties* students may encounter when solving it. Therefore, a thorough description and empirical evaluation of these processes is required.

1.2 PCK as a prerequisite for diagnostic judgments

Judging task difficulties requires PCK. Specifically, one needs knowledge about relevant task characteristics and about students' potential solution processes, that is, about how students deal with tasks and what difficulties they might have (Brunner et al., 2021; Morris et al., 2009; Ostermann, 2018; Philipp, 2018). In this study, we examine participants' pedagogical content knowledge of typical student errors in tasks about functions and graphs, hereafter referred to as "domain-specific PCK of typical student errors." This type of knowledge falls into the subcategory of knowledge of content and students (KCS) described by Ball et al. (2008), which is a facet of PCK.

In the example shown in Fig. 2, the required specific knowledge of typical student errors is knowledge about the so-called "graph-as-picture error." This error means that students often interpret such graphs as a picture of the situation without considering the functional relationships (Brunner et al., 2021; Hattikudur et al., 2012; Nitsch, 2015).

The task shows a graph representing the speed of a roller coaster car during a ride. The question is, "At what time interval did the car take the steepest descent?" In this case, the relevant task characteristic is the section from $t = 8$ to $t = 12$, since this is a typical student error. Typical student errors are errors that occur frequently in a specific type of task

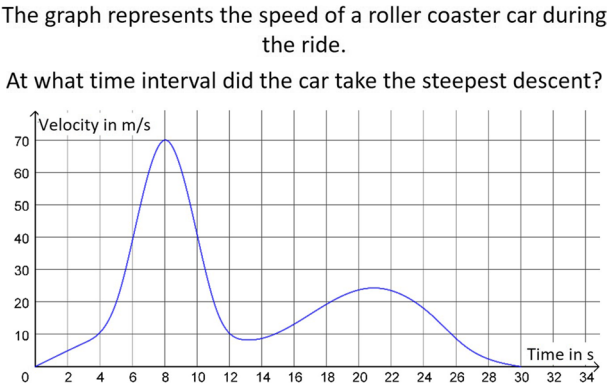

Fig. 2 An example of a task on functions and graphs

(Hattikudur et al., 2012; Nitsch, 2015). One needs to perceive and correctly interpret the relevant task characteristics to identify the possibility of such errors.

Several studies have documented that fostering teachers' PCK leads to more accurate judgments of task difficulty (Hammer, 2016; Kron et al., 2021; McElvany et al., 2009; Morris et al., 2009; Ostermann, 2018; Oudman et al., 2018). However, these studies have not examined *how* PCK leads to accurate judgments in detail. For example, although teachers' perceptions of relevant task characteristics are likely related to their interpretations of these characteristics, and their PCK likely affects perception and interpretation processes, empirical evidence is lacking. McElvany et al. (2009) investigated teachers' domain-specific PCK to perceive relevant task characteristics for text-picture integration. However, as they found only low correlations between domain-specific PCK and comparative judgments of task difficulty, it is still unclear to what extent teachers' perceptions of the characteristics relate to their judgments. Ostermann (2018) used an experimental design and discovered that a high level of PCK on task characteristics increased the accuracy of judging the relative difficulty of mathematical tasks. However, the authors could not infer why this judgment improved, that is, how the processes of perception and interpretation of relevant characteristics influenced the final judging of relative difficulty. Oudman et al. (2018) demonstrated the influence of knowledge of specific student characteristics (the student's name, the student's solution, or both) on teachers' judgment accuracy of student performance. However, it remained unclear whether and how teachers used their knowledge to make judgments.

In summary, although previous research has documented correlative and even causal connections between specific PCK and judgment accuracy, empirical evidence on how PCK affects the actual diagnostic process is scarce.

1.3 Assessing diagnostic processes with verbal reports and eye tracking

Detecting internal diagnostic processes requires externally measurable indicators that reflect these processes (Loibl et al., 2020). Verbal reports, such as thinking aloud, are one method of capturing auditory cognitive processes by reflecting on contexts or explanations (Ericsson & Simon, 1980). Verbalizing thoughts provides insight into participants' processing or interpretation processes (Dannecker, 2018). This verbalization can occur on three different occasions: during the situation (introspection), immediately after the situation (immediate retrospection), or a few days later (delayed retrospection). Although verbal

reports are an important method of capturing cognitive processes, they have limitations in that they cannot assess the diagnostic process in real-time, and the delay and verbalization may alter the actual processes. Nevertheless, this method, particularly immediate retrospection, is well suited to complement other methods such as eye tracking (Schindler & Lilienthal, 2019; Strohmaier et al., 2020).

Eye tracking is a more recent method than verbal reports, and it is increasingly being used in mathematics education research (for a review, see Strohmaier et al., 2020). Just and Carpenter (1980) made the following assumptions for interpreting eye-tracking data: visually perceived information is processed without temporal delay, and the eyes fixate on a specific location while the content of that location is being processed. These assumptions are based on reading research and may not apply to every situation. However, they seem reasonable for visually presented mathematical tasks (Brunner et al., 2021; Klein et al., 2018; Schindler & Lilienthal, 2019). Several studies have found that eye movements depend on the visual perceptual characteristics of mathematical tasks (Brunner et al., 2021; Lee & Wu, 2018; van der Schoot et al., 2009; Verschaffel et al., 1992) and on individual factors (Inglis & Alcock, 2012).

Fixations and saccades are important eye-tracking parameters. Fixations are periods of relative eye rest, usually lasting between 80 and 300 ms, during which the brain begins to process the visual information the eyes receive (Holmqvist et al., 2011). Saccades are rapid eye movements from one fixation to another. Saccades typically last 20 to 40 ms, and during this time, vision is extremely limited and detailed perception does not occur (Holmqvist et al., 2011; Matin, 1974). Saccades and fixations alternate in a continuous process. Fixations can be measured as fixation durations or as the number of fixations on both the entire item and the individual areas of the item, depending on the cognitive process being measured (Holmqvist et al., 2011; Strohmaier et al., 2019). To obtain detailed information about the perceptual processes in each area, a stimulus is divided into so-called areas of interest (AOIs), and eye-movement parameters are analyzed independently for each AOI (Holmqvist et al., 2011). Fixation duration on an AOI is the sum of the duration of all fixations on that AOI and is used as an indicator of the depth of processing on the AOI, with long fixation durations typically associated with deep cognitive processing (Brunner et al., 2021; Holmqvist et al., 2011; Ott et al., 2018).

In summary, eye tracking and verbal reports are suitable methods to assess the cognitive processes of perception and interpretation. In our study, we used these two methods to investigate whether participants perceive relevant task characteristics (measured with eye tracking) and interpret them as relevant (measured with verbal reports) when making judgments (again measured with verbal reports) about task difficulties.

1.4 The present study

This study investigates how prospective mathematics teachers process information when judging mathematical tasks. In particular, this study addresses the correlations between the three diagnostic processes (i.e., perception, interpretation, and judging), as well as the causal influence of domain-specific PCK on perception and interpretation. The mathematical content of the mathematical tasks is functions and graphs. We are interested in the role of teachers' domain-specific PCK of typical student errors in these tasks. These typical student errors are described in more detail below (see 2.3). We assess the process of perception by analyzing eye movements, and the processes of interpretation and judging by analyzing immediate retrospective protocols of judgments. We define *interpretation* as the

analysis of task difficulty (e.g., "*Why* is a task easy or difficult?") and *judging* as the overall judgment (e.g., "*Is* the task easy or difficult for a group of students?").

We hypothesize that high levels of domain-specific PCK of typical student errors promote effective focus on relevant task characteristics, which in turn promotes accurate interpretations of task difficulty. Specifically, our research questions are as follows:

RQ1a: Is the perception of relevant task characteristics related to the interpretation of task difficulty when judging mathematical tasks?
RQ1b: Is the interpretation of task difficulty related to judging?

We expect that participants with a more focused perception (longer fixation durations and more fixations on areas that provoke typical student errors) will analyze task difficulty (accurate interpretation corresponding to naming or describing typical student errors) more accurately than participants with shorter fixation durations and fewer fixations on these areas. Furthermore, we hypothesize that participants with more accurate interpretations of task difficulty will provide more accurate overall judgments of task difficulty than participants with less accurate interpretations.

RQ2: Does a high level of domain-specific PCK of typical student errors lead to a more focused perception and interpretation of relevant task characteristics and to a more accurate judgment of task difficulty?

We hypothesize that high levels of domain-specific PCK of typical student errors will result in more focused eye movements on relevant task characteristics (longer fixations on the areas that provoke typical student errors) and more accurate interpretation and overall judgment of task difficulty.

2 Methods

In an experimental design, we manipulated participants' domain-specific PCK of typical student errors in functional thinking to determine the effect of personal characteristics on diagnostic thinking and behavior.

2.1 Sample

The participants in this study were 49 prospective mathematics teachers (28 females and 21 males; age: $M = 23.43$ years; $SD = 2.07$). They were in their fifth or sixth semester of teacher training at a university in Germany. The study was conducted in a seminar that was part of the participants' study program. According to their study program and self-reports, none of the participants had received instruction of typical student errors in functional thinking prior to this study. We randomly assigned 30 participants to the intervention group and 19 to the control group (originally $n = 25$). Data from six participants in the control group had to be removed due to technical problems and data loss during the first eye-tracking measurement. All participants participated voluntarily and were informed about the procedure before the study began.

2.2 Materials

We designed 20 tasks on functions and graphs that varied in complexity according to visually perceptible characteristics. These tasks were previously utilized in our study about diagnostic processes (Brunner et al., 2021). We included 16 "error tasks" and four "error-free tasks." Each error task addressed one typical student error that is well documented in the mathematics education literature (Clement, 1989; Hattikudur et al., 2012; Nitsch, 2015; Russell et al., 2009). The error-free tasks were designed to be completely parallel in content, but they did not contain the characteristics relevant to a particular student error. Our focus was on three distinct errors: (1) graph-as-picture error, (2) slope–height confusion, and (3) confusion between slope parameter and x-axis intersection. We designed the tasks such that the difficulty of each task could be classified by visual perception and the correct interpretation of a relevant characteristic. We focused on comprehensibility and simple task structure to reduce the potential for varied eye movements. Each task comprised a short introduction to a situation, a question, and a functional graph. Each error task provoked one of the three typical student errors at a specific area of the function graph. This allowed us to ensure that the characteristics relevant to the task's difficulty were visually perceptible and that all other difficulties were kept constant across all tasks.

To illustrate the task characteristics of error tasks, Fig. 3 depicts a task that may provoke the slope-height confusion error. The graph shows the distance covered by three runners. The question concerns the graph that depicts the fastest runner at $t=5$ s. Students who exhibit the slope-height confusion error would interpret the maximum value (the height) at a specific point on the x-axis (Nitsch, 2015; Russell et al., 2009) with the slope at this point. They would respond to this task by stating that the top graph represents the fastest runner. This area is a relevant characteristic because it is where the typical student error is provoked. The correct answer would be the middle graph because it has the largest slope.

In a pilot study, the tasks were presented to 168 eighth-grade students from German middle schools to validate theoretical task difficulties. As expected, students were significantly less accurate on error tasks ($M=0.27$, $SD=0.13$) than on error-free tasks ($M=0.94$, $SD=0.06$), $t(18)=10.01$, $p<0.001$, $d=5.59$). This result corroborates the theoretically

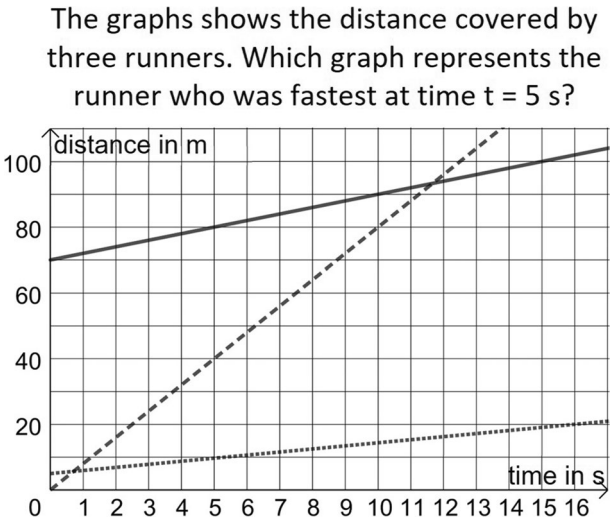

Fig. 3 Sample task for slope–height confusion

derived task difficulty, with error tasks being classified as difficult and error-free tasks being classified as easy.

2.3 Intervention

Since we were interested in how diagnostic processes are affected by domain-specific PCK of typical student errors, we manipulated participants' knowledge through an intervention. Participants in the intervention group received a domain-specific PCK intervention of typical student errors. After the posttest, participants in the control group received the same intervention with a delay. The intervention consisted of two 90-min sessions, during which participants received information about seven typical student errors in solving function tasks. The sessions were part of the regular seminar from which participants were recruited for the study. In the first session, participants received 14 task solutions that contained one out of seven typical student errors of functional thinking. Figure 4 shows a sample task. In the task, the student is asked which graph represents a car driving through a left bend at a constant speed. The bubble represents a student solution that is incorrect.

Participants were asked to identify typical errors in the student solutions. Subsequently, the instructor introduced seven typical student errors, using technical terms based on the literature. In addition, the participants received a description of and information about the possible causes of these seven most common typical student errors, which are well documented in the mathematics education literature (Hattikudur et al., 2012; Nitsch, 2015). Participants received a reading assignment (Nitsch, 2014) to deepen their knowledge and prepare themselves for the second session. They also received an assignment to complete an online test (http://codi-test.de). This test contained tasks that provoked typical student errors, and participants were asked to first complete the tasks and then assign the tasks to the specific student error they provoked. In the second session, the instructor collected the assignment results and checked for accuracy. Then, the characteristics and causes of the seven typical student errors were discussed in detail to consolidate participants' knowledge. The tasks created for the pre- and posttest of the study were not used in the intervention to avoid memory effects.

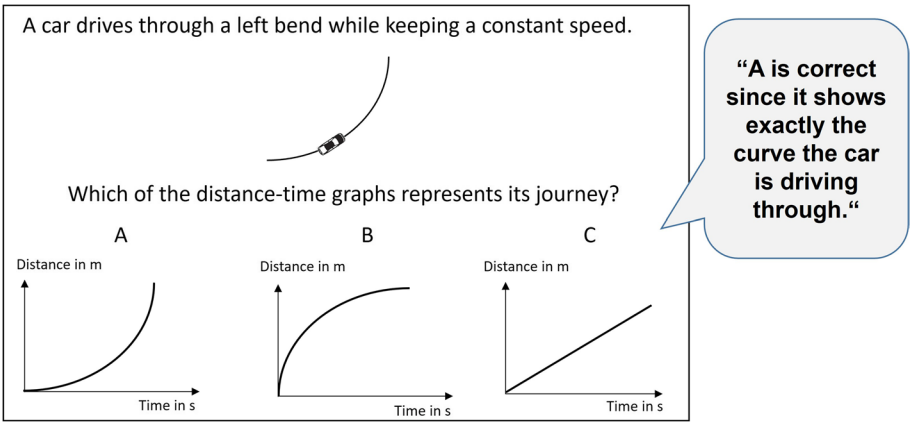

Fig. 4 Sample task from the intervention and one students' incorrect response

2.4 Equipment

We recorded participants' eye movements using a remote eye tracker (SMI 250) with a sampling rate of 250 Hz. Participants sat 65–70 cm away from a 24″ screen with a resolution of 1920 × 1080 px. We did not use a chin rest or other head fixations because the eye tracker was sufficiently robust to head movements. Nevertheless, we told the participants to avoid head movements whenever possible. We used SMI iView X and SMI Experiment Center software for data collection and stimuli presentation. Data were exported using the SMI BeGaze software (minimum fixation duration: 80 ms). Verbal responses were recorded with a digital audio recorder and then transcribed for analysis.

2.5 Procedure

During the pretest, participants completed a questionnaire about demographic data. Subsequently, they were informed about the eye-tracking procedure with a sample task. During the posttest, this sample task was shown again to remind them about the eye-tracking measurement. Further pretest and posttest procedures were identical. We controlled lighting conditions and participants' sitting positions before recording their eye movements. We conducted a five-point calibration before data collection. The maximum deviation in measurement accuracy required for precise analyses of fixation duration is 0.50° (Holmqvist et al., 2011). This was achieved in all calibrations. Thereafter, the participants viewed the 20 tasks (16 error tasks and four error-free tasks) one by one on a computer screen. They were instructed to verbally judge the difficulty of each task using the following scale: 1 (very easy), 2 (fairly easy), 3 (fairly difficult), and 4 (very difficult). There was no time pressure. On average, participation took 25 min to complete the eye-tracking experiment. After the participants finished making their judgment, the experimenter pressed a button to stop eye tracking. After each judgment, participants were asked to provide a reason for their judgment as an immediate retrospection (Konrad, 2010), with the task still visible on the screen. This interpretation was audio-recorded. After the eye-tracking session, the participants were asked to solve the 20 tasks themselves in a paper-and-pencil test. Since the participants solved 90% of the tasks correctly in the paper-and-pencil test, we can assume that they had sufficient knowledge of functional thinking. This suggests that participants' eye movements were related to diagnostic processes rather than to difficulty in solving the tasks. Nevertheless, we excluded eye-movement data from the 10% of tasks that were solved incorrectly.

2.6 Measures

Individual scales were used for each of the three diagnostic processes: perceiving, interpreting, and judging (see Fig. 1 "Diagnostic Thinking"). The operationalizations of the scales are described below. These refer to the "Diagnostic Behavior" in the DiaCom framework (see Fig. 1).

2.6.1 Perception

To measure the perception process, we analyzed local eye movements in the areas showing typical student errors (AOIs, according to Holmqvist et al., 2011). We used the fixation duration on the AOIs as an indicator of perceptual processing depth. Since the size of the

AOIs varied from task to task, and the fixation duration varied from participant to participant, we calculated the relative values as follows: We divided the size of each AOI by the total size of the item and divided the fixation duration on the AOI by the total fixation duration on the item. The relative sizes allowed us to create a scale across all the tasks and participants. For readability, we hereafter refer to the relative values as the "fixation duration on the AOI." The higher the value, the more the gaze is focused on the AOI. The pretest data showed sufficient internal consistency for fixation duration on the AOI, with a Cronbach's alpha of 0.63 ($M = 157.68$, $SD = 59.10$) (Brunner et al., 2021). The posttest data also showed sufficient internal consistency for fixation duration on the AOI, with Cronbach's alpha = 0.63 ($M = 158.47$, $SD = 60.07$).

2.6.2 Interpretation

To address questions about the interrelationships between diagnostic processes, we analyzed data from the posttest because it had a high variance due to the intervention. We measured interpretation by how accurately the participants reasoned why a task was easy or difficult. To operationalize the accuracy of interpretation, a four-level category system was developed using qualitative content analysis according to Mayring et al. (2008), which was inductively derived from the data material. Very high interrater reliability of 0.95 was achieved for two independent raters. Statements that were rated differently were agreed upon through subsequent discussion. Table 1 lists the categories and provides descriptions and examples of participants' statements. Participants received scores ranging from 0 to 3 for each task depending on how appropriate their interpretation was from a theoretical perspective (see Table 1), resulting in a maximum score of 60 across all 20 tasks. The internal consistency of this scale was acceptable in the pretest (Stadler et al., 2021), with a Cronbach's alpha of 0.59 ($M = 33.78$, $SD = 5.28$), and it was very high in the posttest, with a Cronbach's alpha of 0.93 ($M = 49.18$, $SD = 13.90$). This difference is not surprising because task difficulty is determined by a single characteristic that is hardly identifiable in a reliable way with little knowledge of typical student errors.

Table 1 Main categories of interpretation accuracy with descriptions and examples

Value	Label	Description	Example
3	Typical error mentioned or described	Technical terms mentioned or content-related descriptions of typical errors; mathematical and real situations were combined	Since the blue graph is above the black graph, but the black graph is steeper, it could lead to the error that the blue one is faster
2	Parts of the typical error mentioned	There is a reference to the real situation, but the central difficulty in the graph is not mentioned or is only described on a mathematical basis	The student must be able to read the gradient to discover the speed of the runners
1	Another difficulty than the relevant one mentioned	A general statement that the task is difficult without referencing to the real situation or mathematical difficulty	Since several graphs are shown, they have to be compared
0	No pedagogical reason mentioned	No relation to the difficulty of the task	One just has to read and understand

2.6.3 Judgment

We measured judging as the overall judgment accuracy. We asked participants to rate the difficulty of a task on the following scale: 1 (very easy), 2 (fairly easy), 3 (fairly difficult), and 4 (very difficult). We classified error-free tasks as very easy or fairly easy, and error tasks as fairly difficult or very difficult, based on our empirical data from a pilot study (see 2.2). Participants' judgments were then dichotomized as 1 (correct judgment) and 0 (incorrect judgment). Again, and not surprisingly, the internal consistency of this scale was acceptable in the pretest (Stadler et al., 2021), with a Cronbach's alpha of 0.60 ($M = 10.75$, $SD = 2.91$) (Brunner et al., 2021), and sufficiently high in the posttest, with a Cronbach's alpha of 0.75 ($M = 11.88$, $SD = 3.61$).

3 Results

As only the error tasks are relevant for answering the research questions, we excluded error-free tasks from data analysis. We also removed outliers from the eye-tracking data that deviated from more than 1.5 times the interquartile range of the distribution per task.

In the following, we analyze the data regarding the first research questions about the correlations between the diagnostic processes (RQ1a and RQ1b). Table 2 presents the descriptive posttest data, as well as the correlations between interpretation, perception and judgment.

There were significant correlations between perception (fixation duration on the AOI) and the accuracy of interpretation (RQ1a). Participants with long fixation durations on the areas that provoke typical student errors provided more accurate reasons for task difficulty. According to Cohen (2016), this was a medium effect. We also found significant and medium correlations between interpretation accuracy and overall judgment accuracy (RQ1b). Participants who provided relatively accurate interpretations also made relatively accurate overall judgments, and vice versa.

To assess the effects of our PCK intervention on typical student errors (RQ2), we performed a univariate analysis of variance (ANOVA) with repeated measures. The ANOVA included the factors *time* (pre-post; repeated) and *group* (intervention-control). The dependent measures were perception, interpretation, and judgment, respectively.

Table 2 Descriptive posttest statistics for perception, interpretation, judgment, and correlations with interpretation

Scales	*M*	*SD*	Correlations with interpretation
Perception	16.03	6.07	0.43*
Fixation duration on the AOI			
[Higher values correspond to a more focused gaze on the AOI]			
Interpretation	1.78	0.84	-
[value = 0–3; see Table 1]			
Judgment	0.51	0.23	0.43*
[1 = correct; 0 = incorrect]			

Note: M = mean; SD = standard deviation; * indicates $p < 0.05$

Table 3 Descriptive statistics for the perception, interpretation, and judgment

Scales		Intervention group		Control group	
		M	*SD*	*M*	*SD*
Perception	Pretest	9.22	3.72	9.51	3.42
	Posttest	17.23	5.52	14.12	6.55
Interpretation [scale: 0–3; see Table 1]	Pretest	0.95	0.28	0.75	0.28
	Posttest	2.32	0.57	0.92	0.34
Judgment [1 = correct; 0 = incorrect]	Pretest	0.49	0.19	0.36	0.14
	Posttest	0.60	0.21	0.37	0.17

Note: *M* = mean; *SD* = standard deviation

Table 3 presents the descriptive data. The mean values of the three scales show that the participants initially had similar values before the intervention and that both groups showed changes from pretest to posttest.

Regarding perception (fixation duration on the AOI), there was a significant main effect for *time* ($F(1, 47) = 60.72$, $p < .001$, $\eta^2 = .56$), suggesting that participants in the posttest gazed at the AOI with the typical student error for a significantly more extended period than in the pretest. The factor *group* had no significant main effect, implying that the intervention and control groups did not differ significantly overall ($F(1, 47) = 1.41$, $p = .241$, $\eta^2 = .029$). However, there was a significant interaction between *time* and *group* ($F(1, 47) = 4.41$, $p = .041$, $\eta^2 = .086$), suggesting that the intervention had an effect on participants' perception. Participants in the intervention group increased their fixation durations on the relevant task characteristics (i.e., the areas that provoke the typical student error) significantly more than participants in the control group.

Regarding interpretation, all participants had similar interpretation accuracy at the pretest (see Table 3). There were significant effects for *time* ($F(1, 47) = 116.60$, $p < .001$, $\eta^2 = .71$) and *group* ($F(1, 47) = 71.87$, $p < .001$, $\eta^2 = .94$), indicating an overall improvement of interpretation from pretest to posttest, and an overall difference between the intervention and the control group. The significant interaction between *time* and *group* ($F(1, 47) = 69.01$, $p < .001$, $\eta^2 = .595$) demonstrates that participants in the intervention group improved their interpretation accuracy significantly more strongly than participants in the control group. Specifically, this means that after the intervention, participants referred to typical student errors significantly more frequently in their interpretations.

Regarding participants' judgments, there was no significant main effect of *time* ($F(1, 47) = 2.69$, $p = .107$, $\eta^2 = .054$). There was a significant effect of *group* ($F(1, 47) = 19.88$, $p < .001$, $\eta^2 = .30$), but the interaction effect between *time* and *group* was also not significant ($F(1, 47) = 2.13$, $p = .151$, $\eta^2 = .043$). These results suggest that, overall, participants in the intervention group offered more accurate judgments than participants in the control group. However, although the increase from pretest to posttest was descriptively larger in the intervention group than in the control group (see Table 3), this difference was not statistically significant.

In sum, the results suggest that the intervention led to more focused perception and interpretation of relevant task features, although it did not increase participants' overall judgment accuracy. To illustrate the results, Figs. 5 and 6 show two eye-movement

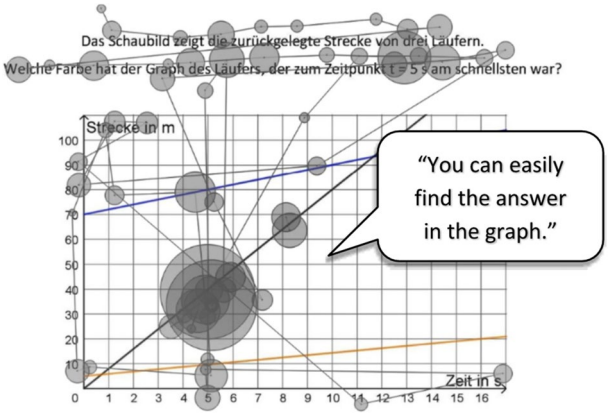

Fig. 5 Scanpath and reason pretest

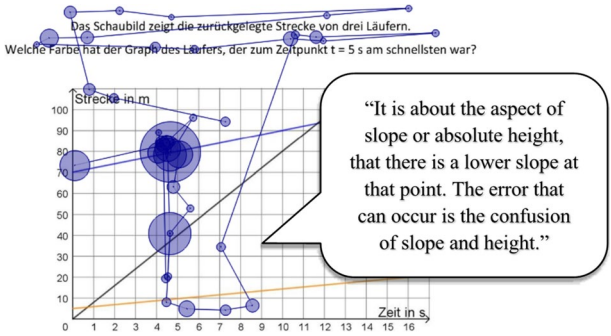

Fig. 6 Scanpath and reason posttest

patterns and verbal responses of one participant in the intervention group for the task displayed in Fig. 3 (i.e., the example task addressing the slope-height error).

Figure 5 shows the participant's eye movements during the pretest. Figure 6 shows the eye movements of the same participant during the posttest after the intervention. During the pretest, the participant fixated more on less relevant task characteristics and focused most of their visual perception on the area that provoked the correct answer. During the posttest, a significant portion of the participant's visual perception was focused on the relevant task characteristic, which is the area that provokes the typical student error. After that, the participant showed more efficient eye movements, indicating that less relevant task characteristics were perceived only briefly. This difference was also reflected in this participant's reasoning about task difficulty. In the pretest, the participant did not provide a good reason for task difficulty. However, in the posttest, this participant named the typical student error (confusing slope and height). The participant recognized the relevant task characteristics and interpreted them as potentially challenging for students.

4 Discussion

The purpose of this study was to increase our understanding of information processing in judging mathematical tasks. In particular, the aim was to explore the relationship between diagnostic processes and whether this relationship could be influenced by manipulating domain-specific PCK of typical student errors.

4.1 Theoretical and practical implications

Participants' eye movements reflected their ability to identify typical student errors in mathematical tasks. There was a significant correlation between participants' interpretation accuracy and their fixation duration on the area that provokes typical student errors. Participants who had longer fixation durations on the areas critical for typical student errors provided more accurate reasons for task difficulty (i.e., they mentioned the typical error more often in their interpretation) than participants who had shorter fixation durations. This demonstrates that the identification of characteristics relevant to task difficulty can be measured through eye movements and that this identification is associated with an in-depth cognitive process related to the areas that provoke typical student errors. Therefore, it seems plausible that participants' eye movements reflect the diagnostic processes of task judgment, and that eye tracking is suitable for recording diagnostic processes (Schindler & Lilienthal, 2019; Strohmaier et al., 2020).

There was also a significant correlation between the quality of interpretation of task difficulty and overall judgment of task difficulty. Participants who provided accurate interpretations also made accurate overall judgments. This finding is in line with recent findings by Rieu et al. (2022), which revealed that participants who identified and weighted characteristics relevant to task difficulty also made accurate judgments. It suggests that when the participants identified a typical student error in a task, they also interpreted it as relevant to task difficulty. We take this finding as support of our assumption that participants who correctly interpret the relevance of a typical student error will also be able to judge tasks more accurately and vice versa.

Regarding the causal influence of domain-specific PCK of typical student errors, we found that manipulating PCK (personal characteristics) influenced the diagnostic process toward a more accurate diagnosis. Participants who received the domain-specific PCK intervention perceived the relevant task characteristics for a significantly longer time than participants who did not receive the intervention. Thus, domain-specific PCK of typical student errors resulted in a focused perception of relevant task characteristics. It is worth emphasizing that the experimental design of our study allows us to draw causal conclusions. Specifically, intervention effects cannot be attributed solely to repetition effects or general experience. However, we can only make statements about fixation duration since the current tasks did not allow for reliable scales of other eye-tracking parameters that could potentially be relevant as well (Brunner et al., 2021).

While previous intervention studies (Dünnebier et al., 2009; McElvany et al., 2009; Ostermann, 2018; Oudman et al., 2018; Rieu et al., 2022) had already shown that task judgment improves with domain-specific PCK, our study suggests that the reason for improved task judgment is the application of participants' acquired domain-specific PCK.

Furthermore, we showed that manipulating domain-specific PCK of typical student errors significantly improved the interpretation of task difficulty. Participants in the intervention group were more able to identify perceived relevant task characteristics as potentially challenging for students than participants in the control group. This indicates that the interpretation of task difficulty can be positively influenced by modifying personal characteristics, meaning that participants are better able to correctly interpret perceived relevant task characteristics if they have acquired a sufficiently high level of domain-specific PCK. These results align with the findings of Mellone et al. (2020), who investigated the mathematical knowledge activated by interpreting student solutions (interpretive knowledge). Mellone et al. (2020) showed that interpretive knowledge changed significantly through discussions of student solutions.

4.2 Limitations

Our results suggest that perception mediated the effect of domain-specific PCK of typical student errors in analyzing task difficulty. However, the task design of our study made it impossible to calculate this mediation effect. The reason is that a typical student error can be found in each item; thus, we cannot exclude the unlikely possibility that participants mentioned the typical student error in their reasons for task difficulty even though they did not fixate on the actual relevant characteristic. Such behavior would lead to judgments that are not mediated by perception. This limitation can be addressed in another study by including similar tasks without typical student errors in the analyses.

Although we expected the intervention to lead to an improvement in overall judgment accuracy, we only discovered this improvement on the descriptive level, but it was not statistically significant. First, the lack of a significant effect could be due to the relatively small sample size. Another explanation could be that the scale we used to assess judgment accuracy was not sufficiently fine-grained to detect small effects. It is worth recalling that the participants had to indicate the problem difficulty on a four-level scale, which we then dichotomized to assess accuracy (see Sect. 2.6). The main focus of our study was not to assess judgment accuracy. Instead, we asked the participants for their judgments to initiate the cognitive processes of perception and interpretation of task difficulty.

4.3 Educational implications and summary

Our study has educational implications. Tasks are a central medium for designing mathematical instructions (Sullivan et al., 2012), and accurate diagnostic judgments are considered an essential component of adaptive instruction (Hardy et al., 2019; Krolak-Schwerdt et al., 2014; Parsons et al., 2018). Thus, the results of this study are an essential step toward an improved understanding of how prospective teachers use their domain-specific PCK of typical student errors to accurately judge the difficulty of mathematical tasks. Domain-specific PCK, in particular, appears to result in accurate task judgments that may enable prospective teachers to assess potential student errors in mathematical tasks. Our study provides the first causal evidence for the assumption that domain-specific PCK sharpens prospective teachers' focus when judging mathematical tasks. Knowledge of this relationship could be useful in designing digital learning environments for enhancing prospective teachers' diagnostic skills. Once eye

tracking can be applied in the ongoing process, such tools could provide adaptive support based on participants' eye-movement patterns.

We conclude that identifying relevant task characteristics can be assessed through local eye movements and that this identification is associated with deeper cognitive processing of the information provided by the areas that provoke typical student errors. Manipulating domain-specific PCK can improve both the perception and the interpretation of task diagnosis.

Funding Open Access funding enabled and organized by Projekt DEAL. This research is part of the graduate school "DiaKom", funded by the Ministry of Science, Research and the Arts in Baden-Wuerttemberg, Germany.

Data Availability The datasets generated and analyzed during the current study are not publicly available due the fact that they constitute an excerpt of research in progress but are available from the corresponding author on reasonable request. The interviews transcribed in this research are currently only available in German.

Declarations

Competing Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching. *Journal of Teacher Education, 59*(5), 389–407. https://doi.org/10.1177/0022487108324554

Becker, S., Spinath, B., Ditzen, B., & Dörfler, T. (2020). Der Einfluss von Stress auf Prozesse beim diagnostischen Urteilen–eine Eye Tracking-Studie mit mathematischen Textaufgaben. *Unterrichtswissenschaft, 48*(4), 531–550. https://doi.org/10.1007/s42010-020-00078-4

Bless, H., & Greifeneder, R. (2017). General framework of social cognitive processing. In R. Greifeneder, H. Bless, & K. Fiedler (Eds.), *Social cognition: How individuals construct social reality* (2nd ed., pp. 16–36). Psychology Press. https://doi.org/10.4324/9781315648156-2

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

Brunner, K., Obersteiner, A., & Leuders, T. (2021). How prospective teachers detect potential difficulties in mathematical tasks–an eye tracking study. *RISTAL, 4*(1), 109–126. https://doi.org/10.23770/rt1845

Carrillo-Yañez, J., Climent, N., Montes, M., Contreras, L. C., Flores-Medrano, E., Escudero-Ávila, D., Vasco, D., Rojas, N., Flores, P., Aguilar-González, Á., Ribeiro, M., & Muñoz-Catalán, M. C. (2018). The mathematics teacher's specialised knowledge (MTSK) model. *Research in Mathematics Education, 20*(3), 236–253. https://doi.org/10.1080/14794802.2018.1479981

Chapman, O. (2014). Overall commentary: Understanding and changing mathematics teachers. In J.-J. Lo, K. R. Leatham, & L. R. van Zoest (Eds.), *Research Trends in Mathematics Teacher Education* (pp. 295–309). Springer. https://doi.org/10.1007/978-3-319-02562-9_16

Clement, J. (1989). The concept of variation and misconceptions in cartesian graphing. *Focus on Learning Problems in Mathematics, 11*, 77–87.
Cohen, J. (2016). A power primer. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (pp. 279–284). American Psychological Association. https://doi.org/10.1037/14805-018
Dannecker, W. (2018). Lautes Denken. In J. M. Boelmann (Ed.), *Empirische Forschung in der Deutschdidaktik* (pp. 131–137). Schneider Verlag Hohengehren.
Dünnebier, K., Gräsel, C., & Krolak-Schwerdt, S. (2009). Urteilsverzerrungen in der schulischen Leistungsbeurteilung. *Zeitschrift Für Pädagogische Psychologie, 23*(34), 187–195. https://doi.org/10.1024/1010-0652.23.34.187
Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215–251. https://doi.org/10.1037/a0022388
Hammer, S. (2016). *Professionelle Kompetenz von Mathematiklehrkräften im Umgang mit Aufgaben in der Unterrichtsplanung* [Doctoral dissertation, University of Munich]. Electronic Theses of LMU Munichhttps://doi.org/10.5282/edoc.20439
Hardy, I., Decristan, J., & Klieme, E. (2019). Adaptive teaching in research on learning and instruction. *Journal for Educational Research Online, 11*(2), 169–191. https://doi.org/10.25656/01:18004
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*(4), 230–240. https://doi.org/10.1111/j.1949-8594.2012.00138.x
Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., Behrmann, L., Böhmer, M., Ufer, S., & Klug, J. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education, 76*, 181–193. https://doi.org/10.1016/j.tate.2017.12.001
Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*(3), 297–313. https://doi.org/10.3102/00346543059003297
Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures* (1st ed.). Oxford University Press.
Inglis, M., & Alcock, L. (2012). Expert and novice approaches to reading mathematical proofs. *Journal for Research in Mathematics Education, 43*(4), 358–390. https://doi.org/10.5951/jresematheduc.43.4.0358
Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329–354. https://doi.org/10.1037/0033-295X.87.4.329
Karst, K. (2012). *Kompetenzmodellierung des diagnostischen Urteils von Grundschullehrern*. Waxmann.
Karst, K., & Bonefeld, M. (2020). Judgment accuracy of preservice teachers regarding student performance: The influence of attention allocation. *Teaching and Teacher Education, 94*, 103099. https://doi.org/10.1016/j.tate.2020.103099
Karst, K., Klug, J., & Ufer, S. (2017). Strukturierung diagnostischer Situationen im inner- und außerunterrichtlichen Handeln von Lehrkräften. In A. Südkamp & A.-K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen* (pp. 102–114). Waxmann.
Klein, P., Viiri, J., Mozaffari, S., Dengel, A., & Kuhn, J. (2018). Instruction-based clinical eye-tracking study on the visual interpretation of divergence: How do students look at vector field plots? *Physical Review Physics Education Research, 14*(1), 010116. https://doi.org/10.1103/PhysRevPhysEducRes.14.010116
Konrad, K. (2010). Lautes Denken. In G. Mey & K. Mruck (Eds.), *Handbuch qualitative Forschung in der Psychologie* (pp. 476–490). Springer.
Krolak-Schwerdt, S., Glock, S., & Böhmer, M. (2014). Teachers' professional development: Assessment, training, and learning. *Sense Publishers*. https://doi.org/10.1007/978-94-6209-536-6
Kron, S., Sommerhoff, D., Achtner, M., & Ufer, S. (2021). Selecting mathematical tasks for assessing student's understanding: Pre-service teachers' sensitivity to and adaptive use of diagnostic task potential in simulated diagnostic one-to-one interviews. *Frontiers in Education, 6*, 604568. https://doi.org/10.3389/feduc.2021.604568
Lee, W.-K., & Wu, C.-J. (2018). Eye movements in integrating geometric text and figure: Scanpaths and given-new effects. *International Journal of Science and Mathematics Education, 16*(4), 699–714. https://doi.org/10.1007/s10763-016-9790-2
Leuders, T., & Loibl, K. (2021). Beyond subject specificity–student and teacher thinking as sources of specificity in teacher diagnostic judgments. *RISTAL, 4*(1), 60–70. https://doi.org/10.23770/rt1842
Leuders, T., Loibl, K., Sommerhoff, D., Herppich, S., & Praetorius, A.-K. (2022). Toward an overarching framework for systematizing research perspectives on diagnostic thinking and practice. *Journal Für Mathematik-Didaktik, 43*(1), 13–38. https://doi.org/10.1007/s13138-022-00199-6

Loibl, K., Leuders, T., & Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgements by cognitive modeling (DiaCoM). *Teaching and Teacher Education, 91*, 103059. https://doi.org/10.1016/j.tate.2020.103059

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

Mayring, P., Gläser-Zikuda, M., & Brunner, E. (2008). *Die Praxis der qualitativen Inhaltsanalyse* (2nd ed.). Beltz. https://doi.org/10.1007/978-3-8349-9258-1_42

McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W., Horz, H., & Ullrich, M. (2009). Diagnostische Fähigkeiten von Lehrkräften. *Zeitschrift Für Pädagogische Psychologie, 23*(34), 223–235. https://doi.org/10.1024/1010-0652.23.34.223

Mellone, M., Ribeiro, M., Jakobsen, A., Carotenuto, G., Romano, P., & Pacelli, T. (2020). Mathematics teachers' interpretative knowledge of students' errors and non-standard reasoning. *Research in Mathematics Education, 22*(2), 154–167. https://doi.org/10.1080/14794802.2019.1710557

Morris, A. K., Hiebert, J., & Spitzer, S. M. (2009). Mathematical knowledge for teaching in planning and evaluating instruction: What can preservice teachers learn? *Journal for Research in Mathematics Education, 40*(5), 491–529. https://doi.org/10.5951/jresematheduc.40.5.0491

Nitsch, R. (2014). Schülerfehler verstehen: Typische Fehlermuster im funktionalen Denken. *Mathematik Lehren, 187*, 8–11.

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

Ostermann, A. (2018). Factors influencing the accuracy of diagnostic judgments. In T. Leuders, K. Philipp, & J. Leuders (Eds.), *Diagnostic competence of mathematics teachers: Unpacking a complex construct in teacher education and teacher practice* (pp. 95–108). Springer. https://doi.org/10.1007/978-3-319-66327-2_1

Ott, N., Brünken, R., Vogel, M., & Malone, S. (2018). Multiple symbolic representations: The combination of formula and text supports problem solving in the mathematical field of propositional logic. *Learning and Instruction, 58*, 88–105. https://doi.org/10.1016/j.learninstruc.2018.04.010

Oudman, S., van de Pol, J., Bakker, A., Moerbeek, M., & van Gog, T. (2018). Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding. *Teaching and Teacher Education, 76*, 214–226. https://doi.org/10.1016/j.tate.2018.02.007

Parsons, S., Vaughn, M., Scales, R., Gallagher, M., Parsons, A., Davis, S., Pierczynski, M., & 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 resources. In T. Leuders, K. Philipp, & J. Leuders (Eds.), *Diagnostic competence of mathematics teachers: Unpacking a complex construct in teacher education and teacher practice* (pp. 109–127). Springer. https://doi.org/10.1007/978-3-319-66327-2_1

Rieu, A., Leuders, T., & Loibl, K. (2022). Teachers' diagnostic judgments on tasks as information processing–The role of pedagogical content knowledge for task diagnosis. *Teaching and Teacher Education, 111*, 103621. https://doi.org/10.1016/j.tate.2021.103621

Russell, M., O'Dwyer, L. M., & Miranda, H. (2009). Diagnosing students' misconceptions in algebra: Results from an experimental pilot study. *Behavior Research Methods, 41*(2), 414–424. https://doi.org/10.3758/brm.41.2.414

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

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

Schrader, F.-W. (2009). Anmerkungen zum Themenschwerpunkt Diagnostische Kompetenz von Lehrkräften. *Zeitschrift Für Pädagogische Psychologie, 23*(34), 237–245. https://doi.org/10.1024/1010-0652.23.34.237

Schrader, F.-W., Praetorius, A.-K., Rost, D., Sparfeldt, J., & Buch, S. (2018). Diagnostische Kompetenz von Eltern und Lehrern. In D. Rost, J. R. Sparfeldt, & S. Buch (Eds.), *Handwörterbuch Pädagogische Psychologie* (pp. 92–98). Beltz. https://www.zora.uzh.ch/id/eprint/161387/.

Stadler, M., Sailer, M., & Fischer, F. (2021). Knowledge as a formative construct: A good alpha is not always better. *New Ideas in Psychology, 60*, 100832. https://doi.org/10.1016/j.newideapsych.2020.100832

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, 2*(40), 255–287. https://doi.org/10.1007/s13138-019-00144-0

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, 104*, 147–200. https://doi.org/10.1007/s10649-020-09948-1

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. *Springer*. https://doi.org/10.1007/978-1-4614-4681-1_3

van der Schoot, M., Arkema, A. H. B., Horsley, T. M., & van Lieshout, E. C. (2009). The consistency effect depends on markedness in less successful but not successful problem solvers: An eye movement study in primary school children. *Contemporary Educational Psychology, 34*(1), 58–66. https://doi.org/10.1016/j.cedpsych.2008.07.002

Verschaffel, L., de Corte, E., & Pauwels, A. (1992). Solving compare problems: An eye movement test of Lewis and Mayer's consistency hypothesis. *Journal of Educational Psychology, 84*(1), 85–94. https://doi.org/10.1037/0022-0663.84.1.85