ZDM (2020) 52:241–254
https://doi.org/10.1007/s11858-020-01139-9

ORIGINAL ARTICLE

Analyzing prospective mathematics teachers' diagnostic processes in a simulated environment

Angelika Wildgans-Lang[1] · Sarah Scheuerer[1] · Andreas Obersteiner[2] · Frank Fischer[3] · Kristina Reiss[1]

Accepted: 29 January 2020 / Published online: 21 February 2020

Abstract
Diagnostic competences are an essential facet of teacher competence. Many studies have investigated the *quality* of teachers' judgments of students' competences. However, little is known about the *processes* that lead to these judgments and about the ways to promote these processes in the early phase of teacher training. The aim of the research project on which we report in this paper was to develop a simulated computer-based environment that allows assessing and promoting the diagnostic processes of prospective teachers. In the simulated environment, 'virtual third-graders' solve mathematical problems. Participants are asked to diagnose the students' competence levels according to a theoretical model, which has been empirically validated. Participants can repeatedly select mathematical problems of varying difficulty levels, assign them to a virtual student, and then receive the student's written solution. In this paper, we present the conceptualization of the simulated environment. We also report on the results of a pilot study with 91 prospective primary school mathematics teachers to analyze whether the environment allows an assessment of individual differences in diagnostic processes. The majority of participants rated the environment as authentic and as one in which they could become immersed. Overall, participants were fairly accurate in their diagnoses concerning the student's competence level. However, log data and participants' written notes indicated that there was large variability in their diagnostic processes. Participants varied greatly in the number of mathematical problems they assigned to a student during their diagnostic process, and in how strongly the difficulty of these problems deviated from the student's true competence level. Overall, the data suggest that the simulated environment has the potential to assess diagnostic processes in a valid way. We discuss open questions and issues for further development.

Keywords Diagnostic processes · Competence level model · Competence level · Diagnostic activities · Teacher education · Simulated environment

1 Introduction

Diagnostic competences[1] are an important facet of teacher competences (Corno 2008; Helmke and Schrader 1987; Karst et al. 2014; Krell et al. 2018). They include an ability to identify students' competences but also their potential misconceptions (i.e., misunderstandings that may cause student errors) or error patterns[2] (Klug et al. 2013). Accordingly, mathematics teachers need good knowledge about typical student misconceptions (Beitlich et al. 2015; Heinrichs 2015; Radatz 1980) and about student competences required to solve mathematical problems (Anders et al. 2010; Hoffmann 2017). Competence models that describe students' competences on different levels may be used as a frame of reference and may support prospective teachers in diagnosing students. Such a competence model is the basis of our current study.

Research on teachers' diagnostic competences has often focused on the *accuracy* of teachers' judgements (Artelt

✉ Angelika Wildgans-Lang
a.wildgans-lang@tum.de

[1] Heinz Nixdorf-Chair of Mathematics Education, TUM School of Education, Technical University of Munich, Arcisstr. 21, 80333 Munich, Germany

[2] Institute for Mathematics Education, Freiburg University of Education, Kunzenweg 21, 79117 Freiburg, Germany

[3] Chair of Education and Educational Psychology, Ludwig Maximilians University, Leopoldstr. 13, 80802 Munich, Germany

[1] Note that the term "assessment competences" is also common in the literature (e.g., Herppich et al. 2017a).

[2] For better readability, we only use the term "misconception" in the following, meaning also error patterns (see Sect. 1.3 for definitions).

and Rausch 2014). However, much less is known about the *processes* that lead to these judgements (Herppich et al. 2017a). Moreover, the question of how to support teachers effectively in their development of diagnostic competence is largely unclear, although it is receiving increased attention (Fischer and Opitz 2020; Leuders et al. 2018; Südkamp and Praetorius 2017). Because diagnosing in a real classroom is a complex situation (Stokking et al. 2003; Heitzmann et al. 2019), simulated environments that allow the reduction of the complexity may be beneficial for both the assessment of diagnostic processes and the support of diagnostic competences especially during the early phases of teacher training.

In this article, we present a computer-based simulated environment that we developed for assessing and enhancing primary school teachers' diagnostic competences in mathematics. In the following, we introduce a framework that we used as a basis for analyzing diagnostic processes. We then elaborate on a model of mathematical competence that is based on theory and empirical evidence, which was the basis for the development of our simulated environment. Finally, we present the first analyses of a pilot study that was intended to evaluate the simulated environment in a sample of 91 prospective primary school mathematics teachers.

1.1 Diagnostic competences

Diagnostic competences can be defined as "individual dispositions enabling people to apply their knowledge in diagnostic activities according to professional standards to collect and interpret data in order to make decisions of high quality" (Heitzmann et al. 2019, p. 9). Studies often emphasize the aspect of context specificity. For example, studies found that primary school teachers who are good at diagnosing students in mathematics are not necessarily good at diagnosing them in writing or reading (Klieme et al. 2008; Koeppen et al. 2008; Spinath 2005). Thus, to assess (and promote) the diagnostic competences of prospective teachers, we need to adopt a subject-specific approach. Research has often focused on the *accuracy* of teachers' judgements of students' abilities, with accuracy referring to the difference between an individual's judgement and a more objective assessment of performance (Artelt and Rausch 2014). However, in order to support prospective teachers in developing their diagnostic competences, we need to better understand the diagnostic *process* that leads to their judgements (Artelt and Rausch 2014). Moreover, the model of Blömeke et al. (2015) suggests that the relationship between teachers' knowledge and their performance in a specific situation is not straightforward, but that situation-specific processes in which teachers evaluate the available data (e.g., students' solutions) plays an important role as well. Therefore, analyzing (prospective) teachers' diagnostic processes can contribute to our understanding of how teachers make use of the available data to make a diagnostic decision.

1.2 The diagnostic process

According to the definition (see Sect. 1.1), diagnosing is a process that requires a person to draw inferences based on the evaluation of data. In that sense, diagnostic processes can be considered as a type of scientific reasoning. Although different conceptualizations of diagnostic processes exist, Fischer et al. (2014) define eight diagnostic activities for scientific reasoning processes that we consider relevant in diagnostic processes (Wildgans-Lang et al. 2019). The first activity is (1) *problem identification*: The teacher identifies which information is relevant in a given situation. For example, the teacher notices that a student fails in solving a mathematical problem. (2) *Asking questions*: The teacher starts asking how he/she could identify the problem. For example, the teacher may ask whether the student has understood the task, or whether he/she holds a misconception that needs to be addressed. (3) *Generating hypotheses*: The teacher generates a hypothesis about possible sources of the student's incorrect answers. For example, the teacher may think that a student does not understand the base-ten system, which could be the reason for the student's errors in an arithmetic problem. (4) *Construct artefacts*: Based on this hypothesis, the teacher may construct or redesign artefacts, that is, create or select mathematical problems that could be suitable for identifying the student's misconception. (5) *Generate* and (6) *evaluate evidence*: The teacher generates and evaluates evidence by presenting the created mathematical problem to the student, and then evaluating the student's written solution or verbal response. (7) *Drawing conclusions*: Based on this information, the teacher can draw conclusions. (8) *Communicate results*: Finally, the teacher can scrutinize and communicate the results to colleagues, the student, or the student's parents. It is worth noting that the assumption of this process model is that teachers can engage in each activity multiple times, and not necessarily in the order described here (Fischer et al. 2014).

Among the eight activities of this process model, four activities seem to be particularly important when teachers diagnose students' competences and difficulties based on their written work: generating hypotheses, generating evidence, evaluating evidence, and drawing conclusions. These activities are part of diagnostic process models proposed by other authors (e.g., Heinrichs and Kaiser 2018, p. 84; Herppich et al. 2017b, p. 82; Philipp 2018, p. 123), and they are also part of more general problem-solving and argumentation processes (e.g., Krell et al. 2018; Polya 1973; Rubinstein 1995; Schoenfeld 1985). The advantage of the diagnostic activities put forward by Fischer et al. (2014) is that they are clearly defined and delimited from each other.

Table 1 Excerpt from the competence level model published in Reiss and Winkelmann (2009) (see also Reiss et al. 2012; Reiss and Obersteiner 2019)

Competence level	Description	Examples in the domain "numbers and operations"
Level 1 (lowest level)	Basic technical knowledge (routine procedures based on elementary conceptual knowledge)	Structure of the decimal system Basic single-digit multiplication and addition Subtraction and addition of lower numbers Simple problems with addition and subtraction
Level 2	Basic use of elementary knowledge (routine procedures within a clearly defined context)	Dealing with various representations of numbers Simple problems related to the basic types of calculation Solution finding through systematic trial and error
Level 3	Recognition and utilization of relationships within a familiar context (both mathematical and factual)	Partially and fully written calculation procedures Using basic procedures of mental arithmetic in unfamiliar contexts Rounding numbers meaningfully
Level 4	Secure and flexible utilization of conceptual knowledge and procedures within the curricular scope	Problem solving with all types and variations of calculations Systematic use of the attributes of the decimal system and relationships between operations Use of solution strategies Modelling complex situations
Level 5	Modeling complex problems and independent development of adequate strategies	Solving difficult mathematical problems Comprehending and describing different solutions

However, it is yet not clear whether and how often (prospective) teachers actually engage in the different diagnostic activities during their diagnostic processes. It is also unclear whether engaging in these activities is actually related to better diagnostic results.

Diagnostic results can be the teachers' description of students' competences or students' misconceptions. However, diagnosing also includes an ability to rank students according to their competence levels, which is required, for example, in the classroom assessments or final exams (Artelt and Rausch 2014). Most empirical studies focused on the ranking of students according to (often ill-defined) competence levels, and considered only the accuracy of the diagnostic judgements (Artelt and Rausch 2014; Karst 2012; Spinath 2005; Südkamp et al. 2012).

1.3 Diagnosing with competence level models

Competence level models describe what students are able to do at different levels of competence. These models may assist teachers in their diagnostic activities because they offer a framework that can be used as a reference (Reiss and Obersteiner 2019). The competence level model developed by Reiss and Winkelmann (2009) for primary mathematics education (grade 4) describes mathematical competence on five levels. The model is based on theories of mathematical learning and has been empirically validated in large-scale assessments, e.g., in VERA (comparative assessment; in German: "Vergleichsarbeiten"; Stanat 2012). In VERA, data were scaled according to the Rasch model that assigns students' competences and task difficulties to one of the five levels of the competence model (Stanat 2012). A student is assigned to a specific level if he or she responds correctly to at least 60% of the tasks at the respective level of difficulty. The VERA test is based on the educational standard for school mathematics in Germany (Kultusministerkonferenz 2004) (similar to educational standards in the U.S.; Common Core State Standards Initiative 2012).

In the domain of numbers and operations, which is the focus of the current study, students at level 1 (which is the lowest level), only master tasks related to the basic structure of the decimal system. They may have problems dealing with various representations of numbers, which is a requirement at level 2. Students at level 3 are able to utilize conceptual knowledge and procedures, and at level 4 they are able to do so in a flexible way. Students at level 5 can use procedures and develop strategies for solving complex problems. Table 1 provides a brief overview of the five competence levels in the model. For a more detailed description of the model, see Reiss and Winkelmann (2009), Reiss et al. (2012) and Reiss and Obersteiner (2019).

The tasks used in the VERA assessment are designed for criteria-based assessment. Thus, they may also be beneficial for teachers when diagnosing the competence level of a student as defined in the model, because students who solve these tasks require skills described on the specific level of competence. When teachers diagnose students' competencies based on their written work on a mathematical problem, they need to make inferences about students' skills and compare them to the skills described at different levels of the model.

While competence models describe students' abilities, it is also possible to characterize student competency levels by their specific error patterns or misconceptions. Error

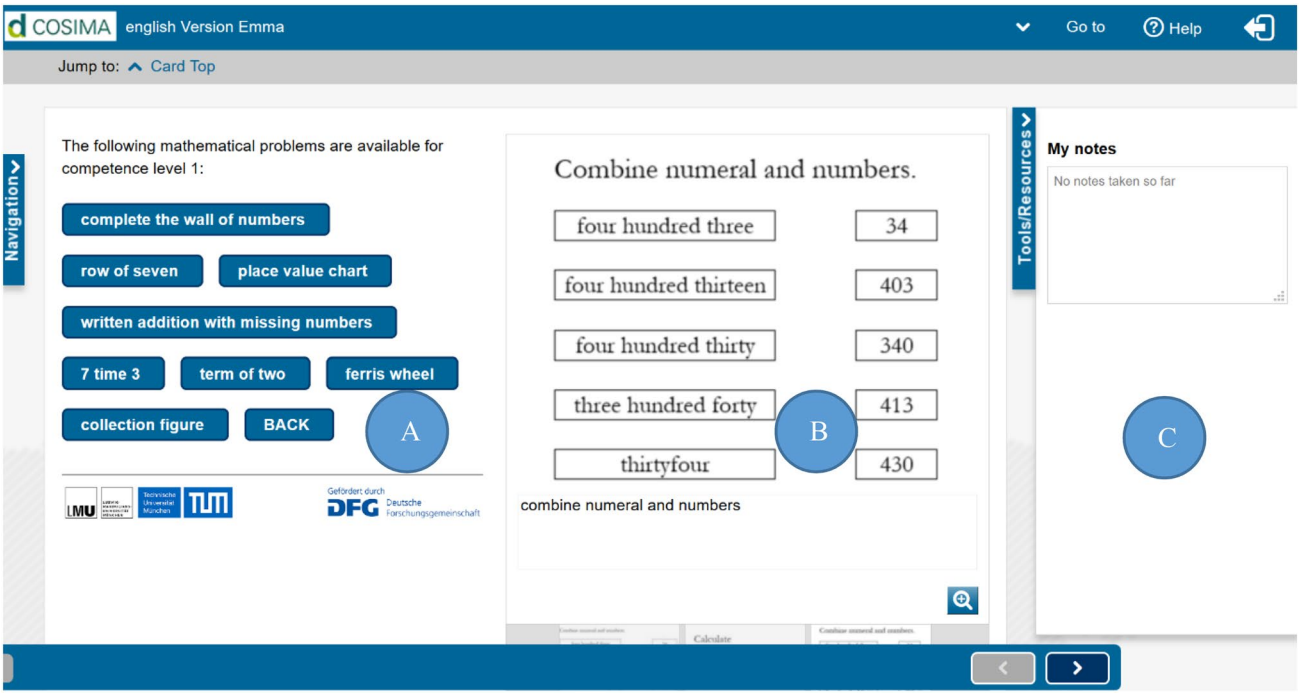

Fig. 1 A screenshot of the simulated environment. The environment is normally in German, here exemplary single views have been translated. After the choice of a competence level, a new task can be chosen (A). Every task has a preview in the second column (B). The display of the notes participants have taken earlier is on the right side (C)

patterns are "systematic procedures that students learn but which most often do not provide the correct answer" (Ashlock 2006, p. 9). Misconceptions are students' individual knowledge structures that are mathematically inaccurate and that cause incorrect answers in certain mathematical problems (see Holmes et al. 2013). In the following, we consider both, misconceptions and error patterns, although for better readability, we use the term misconceptions only. In contrast to the mathematical skills, students' misconceptions are not explicitly described in the model. Thus, describing students' misconceptions requires knowledge of potential misconceptions as well as an ability to identify the specific misconception in students' written work. In the simulated environment we develop in this project, the aim was to assess both types, namely, diagnosing competence levels and students' misconceptions.

1.4 The simulated environment

Diagnostic situations in the classroom are complex, so they are less suitable for assessing (and promoting) prospective teachers' diagnostic competences. Simulations, however, have the advantage that they allow diagnosing in controlled and yet authentic environments (Derry et al. 2014; Salen and Zimmerman 2006). Unlike a real classroom situation, a simulated environment offers pausing, reflecting and repeating the diagnostic process at any time without the constraints of a real classroom situation (Blomberg et al. 2013; Rich and Hannafin 2009; Santagata 2005).

We constructed a computer-based simulated environment to assess (and, in a later phase of our research project, to promote) teachers' diagnostic processes. Participants (i.e., prospective primary teachers) can use the simulated environment for presenting tasks to virtual-third-graders and diagnosing their misconceptions and competence levels. We used original test items and original students' written work from the large-scale German survey VERA, which took place between 2015 and 2017. All selected test items were related to the domain numbers and operations, as well as patterns and structures, as defined in the German educational standards for primary school mathematics. From the original data, we developed 15 virtual students who varied in their error patterns (and, therefore, in the proposed underlying misconceptions) and, consequently, in the assigned competence levels. The use of the original tasks and original students' solutions should make the simulated environment more authentic.

Figure 1 illustrates the structure of the user interface of the environment. On the left hand side (A), participants can first select a competence level and then choose a specific task that is assigned to that level (here: level 1). In the middle of the screen (B), participants can generate a preview of each task listed on the left side. On the right hand side (C),

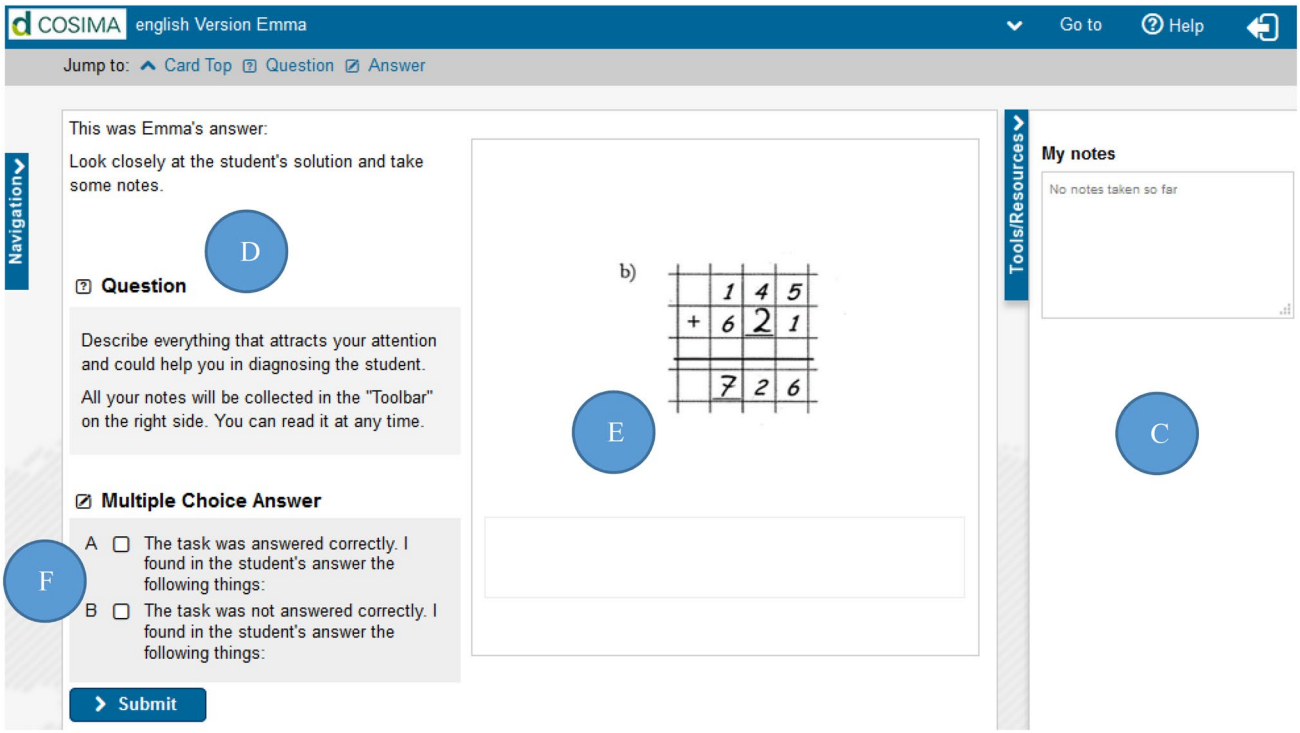

Fig. 2 Screenshot of the simulated environment. The student's solution is in the second column (E). The text on the left side explains the work order (D). The multiple-choice answer (F) asks whether the task was answered correctly or incorrectly by the student

there is the collection of the participant's notes that they are encouraged to make throughout their use of the simulated environment.

After choosing a task, the participants receive the student's answer (Fig. 2). On the left side of that screen, the participant finds the instruction (D), while the student's solution is displayed in the middle of the screen (E). Participants have to judge the correctness of the solution using the multiple-choice box (F). In addition, participants are asked to make notes to record and reflect their observations and thoughts. These notes appear on the right side (C).

The environment was created such that it is appropriate to assess (prospective) teachers' central diagnostic activities according to the model by Fischer et al. (2014). The activities of *problem identification* and *asking questions* are covered by introducing participants to the diagnostic situation, and by asking them to identify the virtual students' misconceptions and competence level according to the model. Accordingly, the simulated environment does not collect data on these two activities. Participants are supposed to *generate hypotheses*, which they can write in the textboxes. They *generate evidence* by selecting one task out of five to eight tasks per competence level and per virtual student. Thus, participants do not have the opportunity to *construct artefacts*.[3] However, they can select the tasks, *evaluate* the virtual students' responses (*evaluate evidence*), and *draw conclusions* from these responses. Lastly, participants are prompted to provide and *communicate* a diagnosis (i.e., write in the textbox). The log data allow us to assess *generating evidence* (i.e., mouse clicks required to select tasks), and the participants' written notes can be coded in order to assess the activities of *generating hypotheses*, *evaluating evidence*, and *drawing conclusions*.

Participants could repeat the activities any number of times. If participants indicate they wish to finish the process, they have to provide a final judgment, which includes (1) determining the student's competence level in a multiple-choice item, (2) specifying the virtual student's primary misconception in an open answer as well as in a multiple-choice item, and (3) providing reasons for their choices. After finishing the diagnosis of one virtual student, participants can choose the next student.

We created the virtual students such that their written work reflected the competences specified in the description of the appropriate level in the competence model. Although

[3] We considered this a valid situation because teachers often select mathematical tasks from instructional material (e.g., textbooks).

the model does not detail student misconceptions, we assigned one primary misconception to each virtual student, such that this misconception was in line with the virtual students' level of competence. The specific misconceptions were also based on those described in the educational literature for mathematics. For example, the misconceptions of the students on competence level 1 were misunderstandings of the place value system, lack of knowledge on how to multiply two numbers, or misconceptions about strategy use for solving word problems (Franke et al. 2010; Hasemann and Gasteiger 2014; Padberg and Benz 2011). On competence level 2, the students have problems with written addition and understanding calculation with zero. Students at the competence level 3–4 solve only simple word problems, but have difficulties with multistep word problems (Franke et al. 2010; Reiss and Winkelmann 2009).

As a concrete example, Jonas is a virtual student in the environment whose competence level is level 1, and whose primary misconception regards his understanding of word problems. He would always perform calculations with the numbers given in the text without considering the described situation. This is a typical misconception about word problems according to the German mathematics education literature ("Identifikationsfehler"; Franke et al. 2010, p. 114). Because Jonas is on competence level 1, he would theoretically solve about 60% of the tasks at this competence level (see Sect. 1.3). In the simulated environment, he actually provides the correct answer to one of two word problems at competence level 1, while solving the other incorrectly. At the higher competence levels he can solve arithmetical problems that are not embedded in word problems, but he would not be able to solve more complex word problems correctly, such as one on competence level 2: "A collector's album can hold 180 stickers. Paul already has 122. Paul is missing _ stickers". The example shows that during the diagnostic process, participants would have to recognize that Jonas' errors appear mainly in word problems. A closer look at his written work is necessary to identify Jonas' specific misconception, that he uses wrong operations and irrelevant information in his calculations, but has no problems in computing the numbers.

In the development of the simulated environment, we considered some principles of video game development such as the principle of "economy of information" (Salen 2004). This means that participants received a certain amount of information, namely the virtual student's solution to the chosen task. It was then up to the participant to decide which part of information they wanted to collect, to dismiss, and to use in order to draw their conclusions. For example, participants had to decide whether the presented virtual student's solution for the task did or did not provide a useful hint to make inferences about the student's misconception and his or her competence level. The participants were also free to decide how much information about a student they considered sufficient for diagnosing (i.e., how many tasks they wanted to present to the student). To assist participants in memorizing the information they received, they could write notes into a textbox. These notes could also support participants in distinguishing between more or less relevant information, in formulating concrete hypotheses about students' competences, and in formulating their conclusions. From a research perspective, we considered this feature as useful for the retracing of the diagnostic process. Together with the log data, the collected data should allow us to observe the occurrence of the diagnostic activities during diagnosing within the framework of the competence model.

2 Aims and research questions

The aim of this study was to analyze and describe the diagnostic processes of prospective teachers. A fundamental requirement for a valid measurement is that the simulated environment be accepted by the participants, that is, that they perceive the environment as authentic and can immerse themselves in the situation (Derry et al. 2014). Accordingly, the first question was whether participants rated the simulated environment as authentic and whether they could immerse themselves in the simulated environment. We expected that this would be the case because the simulated environment was based on original tasks and original students' solutions, and because participants were introduced to the setting ("imagine you are a teacher") at the beginning.

The second question was how frequently participants engaged in different diagnostic activities (according to the model by Fischer et al. 2014). We were also interested in whether the frequency of these different diagnostic activities was related to the accuracy of the diagnoses.

Our third question concerned how participants would make use of the competence level model during their diagnostic processes. Participants were supposed to use the competence model (see Sect. 1.3) as a reference for their diagnoses, and they could select tasks from folders labeled with the difficulty levels (see Sect. 1.4). This aspect allowed us to analyze different usages of this information during the diagnostic process. For example, one strategy would be to systematically select tasks based on whether or not the virtual student had solved the previous task correctly. In that case, the competence levels of the selected tasks would approach the virtual students' actual competence level step by step after each evidence evaluation.

3 Method

3.1 Sample

The participants in this study were 91 university students (77 female, 14 male, $M = 22.9$ years old, $SD = 2.96$, from 20

Table 2 Table of codes, descriptions, and sample responses of diagnostic activities

Diagnostic activity	Description	Sample responses
Generating hypothesis	Hypotheses, recognizable by signs or words of uncertainty that were or were not based on the analyzed evidences	"Problems with + and −?" "Perhaps he has problems with calculating, because 4 + 2 is not 2…"
Evaluating evidence	Evidence evaluation includes any references in the notes directly related to a specific information in the student's solution. It contains descriptions, explanations and evaluation of the analyzed evidence	"The student tried to complete a written calculation, but calculated 4 − 2 instead of 4 + 2" "The student made only one mistake in the second row. It would be interesting how the student calculated."
Drawing conclusion	Conclusions based on the analyzed evidences, evaluations and hypotheses. Every note that is universally valid, independent of the shown task	"The student will fail calculating written additions, because he demonstrated problems in calculating column by column several times." "The student will repeat written addition with smaller numbers and the teacher will show him calculating with Dienes."
Others	Further comments without any reference to the task, the student or a mathematical concept	"Girls are not as good at mathematics so they most likely need intensive support at all levels."

to 39) at two universities in Germany. They were prospective primary school teachers in their first to ninth semester ($M = 3.8$ semesters, SD = 1.44) who participated voluntarily. Thirty percent of the participants had already attended a mathematics education course in which diagnosing and/or students' typical errors and misconceptions in Grade 1–4 had been part of the curriculum. Accordingly, we expected to be able to identify individual variations in our participants' diagnostic competences. Participants worked on the simulated environment in two sessions of approximately 90 min each. Sessions were part of the diagnostic courses at the university.

3.2 Test instruments

Participants were asked to fill in questionnaires about perceived authenticity and presence, while they worked in the simulated environment. Authenticity and presence were assessed in the second session after participants had finished two diagnoses. Participants rated all items on a 5-point Likert-type scale from 1 (totally disagree) to 5 (totally agree). Perceived authenticity was assessed with three items that asked whether they rated the simulated environment as authentic, whether the simulated environment appeared like a real professional requirement situation, and whether the feeling was comparable to the feeling in a professional requirement situation (Seidel et al. 2011).

To assess presence we adapted four items from Seidel et al. (2011). The items asked whether the participants were fully concentrated on the situation, whether they forgot that they took part in a study for a period of time, whether they were mentally immersed into the situation, and whether they were entirely involved in the situation.

3.3 Coding scheme for diagnostic activities

While we assessed the diagnostic activity of *generating evidence* by analyzing the log files (participants' selection of tasks), we developed a coding scheme to assess the other three diagnostic activities of interest *generating hypothesis, evaluating evidence* and *drawing conclusions* (see Sect. 1.2). In a first step, the codes were defined based on our definitions of diagnostic activities (Fischer et al. 2014). The second step was an inductive process, in which we refined the codes using the available data. *Evaluating evidence* includes the descriptions, explanations and evaluations in the written notes (see Seidel and Stürmer 2014). If signs or words of uncertainty appeared in the written notes, we assigned them to the diagnostic activity *generating hypothesis*. If the written notes included predictions and support opportunities, we coded the diagnostic activity *drawing conclusion* (Table 2).

The coding was carried out by three independent raters[4] who, altogether, double-coded more than 65% of the written notes. The written notes were often very short, which made the coding process challenging for the raters. Yet, the inter-rater correlation for each two coders was, on average, sufficiently high (Cohen's Kappa: $\kappa = .73$). All disagreements were discussed among the coders until a full agreement was reached, and the coding manual was refined afterwards.

[4] Note that because one of the raters left the project for organizational reasons, the three raters double-coded, respectively, 21% (rater 1 and 2) 4% (rater 2 and 3), and 41% (rater 1 and 3) percent of participants' written notes. Cohen's Kappa was calculated as the weighted average of the three Kappas for each two raters.

4 Results

4.1 Acceptance of the simulated environment

Overall, participants rated the environment as fairly authentic ($M = 3.15$, SD = 0.91, $n = 82$; scale from 1 to 5). Only 19.5% disagreed (response 1 or 2) with the statement that the simulated environment was authentic. Participants showed a high presence ($M = 4.12$, SD = 0.81, $n = 82$). 79.3% fully agreed with the statement "I immersed myself in the situation".

4.2 Accuracy of the diagnostic results and occurrence of diagnostic activities

The number of diagnosed virtual students per participant varied between 1 and 11 ($M = 5.08$, SD = 1.91), indicating large variability between participants. Altogether, the 91 participants diagnosed the 15 virtual students in 485 cases, which is the number of times a virtual student was selected to be diagnosed by one participant. Each virtual student was diagnosed between 19 and 55 times ($M = 30.80$, SD = 11.66). Because the assessment time was limited, 20 of these cases were incomplete. The following analyses regarding competence levels are based on the 465 cases that were complete. The analyses regarding misconceptions are based on 462 cases, because three cases were terminated before participants responded to the misconceptions.

The number of tasks used to diagnose one virtual student varied between two and a maximum of 55 tasks ($M = 13.34$, SD = 8.22). There were hardly any differences in the number of tasks used between cases with accurate competence level and misconception diagnoses ($M = 13.6$, SD = 8.7, $n = 114$) or inaccurate competence level and misconception diagnoses ($M = 11.9$, SD = 7.3, $n = 50$), $t(162) = 1.21$, $p = .23$.

Overall, the diagnosed competence level of the virtual student was correct in 75.1% of the 462 cases. The two virtual students who were most often diagnosed correctly were a student at competence level 5 (94.3% correct) and a student at competence level 1 (91.9%). This result suggests that students with very high or very low competence were easier to diagnose than students with medium-level competences. For the current preliminary analysis, we included only cases on competence levels 2, 3 and 4 (216 cases), because cases on competence level 1 and 5 had high solution rates and less differences in the diagnosing processes. Recognizing the primary misconception of the virtual student was more difficult for our participants. Only 29.4% of the 462 cases were diagnosed correctly.

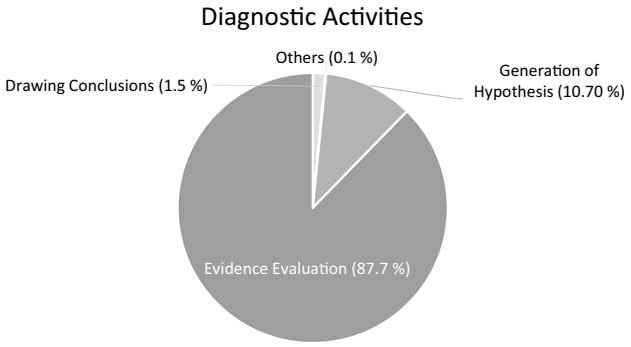

Fig. 3 Relative frequencies of diagnostic activities in 216 cases

Although participants had the opportunity to take notes of each student's work, they made use of this feature in only 69.0% of the tasks, resulting in a total of 2034 notes. Figure 3 shows the proportion of the diagnostic activities per case, i.e., per diagnosed virtual student and participant. As can be seen in the figure, the large majority of participants' notes (87.7%) were coded as *evidence evaluation* (for a detailed description see Table 2), which corresponds to an average of about $M = 8.33$ (SD = 4.40) evidence evaluations per case.

10.7% of the notes were hypotheses, and only 1.5% were conclusions. 0.1% of the notes could not be assigned to any diagnostic activity. There were some cases with up to twelve hypotheses, but there were 110 of 216 (51.0%) cases without any hypothesis. Sixty-eight (31.5%) cases had at least one hypothesis. For cases that contained at least one conclusion, the total average number of conclusions was $M = 1.29$ (SD = 0.55). Since the proportion of hypotheses and conclusions was so small, we did not analyze whether the occurrence of these diagnostic activities differed between accurate and inaccurate diagnoses. For evidence evaluation, we did not find a difference in frequency of this diagnostic activity for accurate diagnoses ($M = 8.63$; SD = 4.54) versus inaccurate diagnoses ($M = 8.46$; SD = 4.19), $t(54) = 0.14$, $p = 0.89$. This result suggests that the pure number of evidence evaluation is not a great predictor of correctness of diagnoses.

4.3 Analyzing diagnostic processes with competence level models

To analyze how participants made use of these competence levels, the cases were divided into correct or incorrect decisions. To illustrate the use of competence levels, Fig. 4 shows the mean task level of all participants who diagnosed both the competence level and the misconception *accurately* (Fig. 4a) or both *inaccurately* (Fig. 4b) for all virtual students at competence level 5.

As one may expect, the task level of the accurately-diagnosed virtual students at competence level 5 approaches

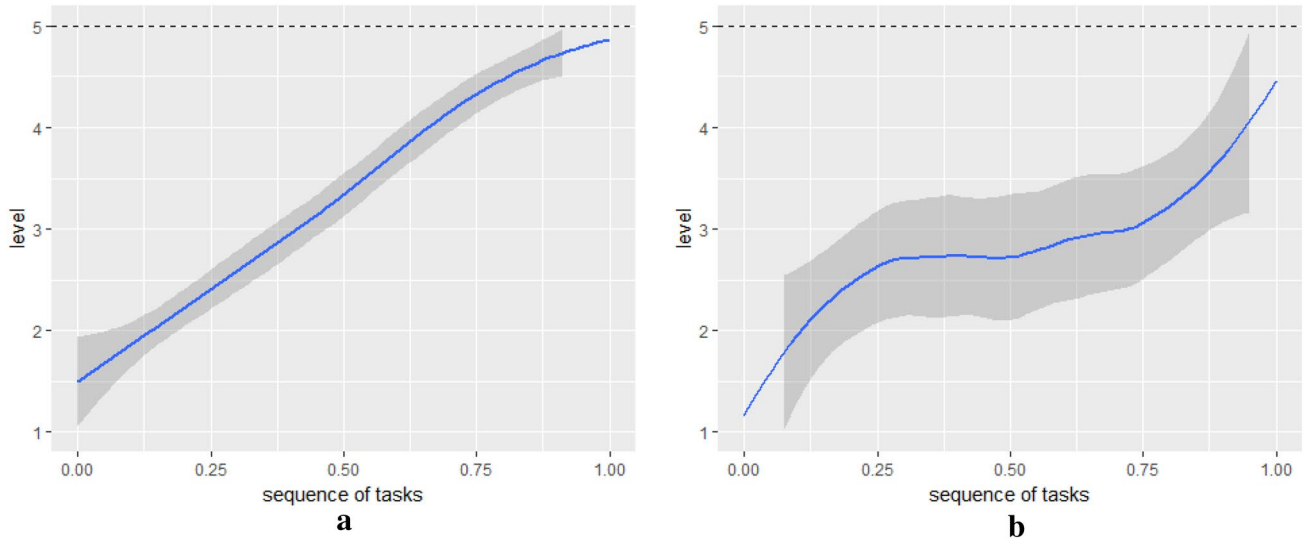

Fig. 4 The standardized processes show the mean task levels of the choice of tasks for **a** accurately-diagnosed virtual students at competence level 5 ($n=45$ cases) and **b** inaccurately-diagnosed virtual students at competence level 5 ($n=7$ cases)

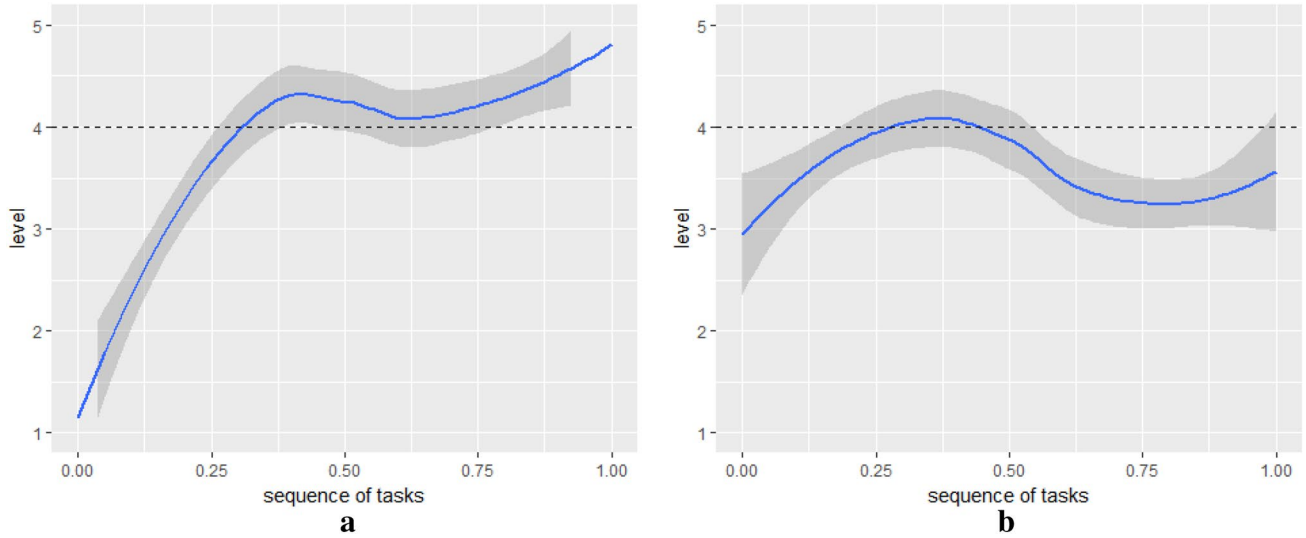

Fig. 5 The standardized processes show the task levels of the choice of tasks for accurately-diagnosed virtual students at competence level 4 ($n=4$ cases, **a**) and inaccurately-diagnosed virtual students at competence level 4 ($n=24$ cases, **b**)

the actual competence level 5 in a more straightforward way (Fig. 4a). Inaccurate diagnoses also approach level 5 (Fig. 4b), but most of the initial tasks are on much lower competence levels.

In contrast, the patterns of accurate and inaccurate diagnoses at competency level 4 are not as clear (Fig. 5a, b). The difference is that accurate diagnoses relied to a larger extent on tasks that were, on average, on levels 4–5 (last two-thirds of the sequence). Inaccurate diagnoses used tasks that were, on average, on levels 3–4 during the whole process.

Accurate and inaccurate diagnoses at competence levels 1, 2 or 3 did not show approximations to the competence level of the virtual student. Looking more closely at the four accurately-diagnosed cases of one virtual student at competence level 3 reveals different strategies, which are illustrated in Fig. 6. There are two participants who use almost every task per competence level increasing from 1 to 5 (A) or 2 to 4 (B). Two other participants seem to use tasks randomly on every competence level (C and D).

In the whole data set, 63.9% of the participants used tasks on at least three competence levels in purely ascending or

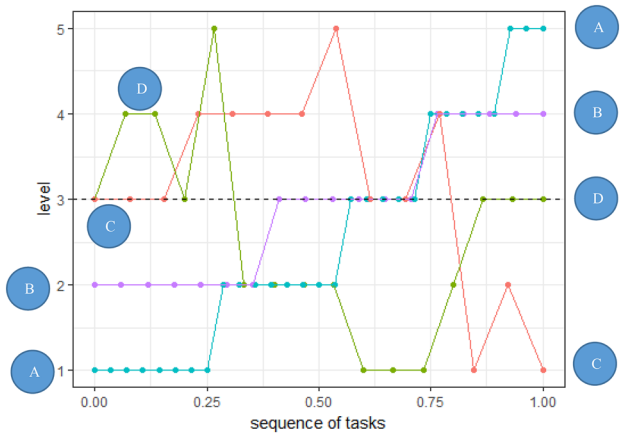

Fig. 6 Different strategies in the diagnostic processes of four participants diagnosing a virtual student at competence level 3

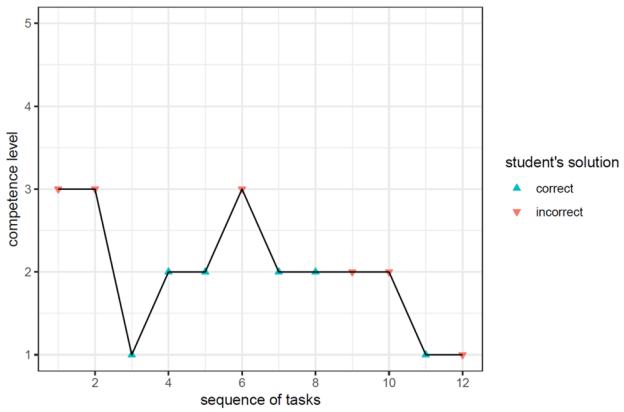

Fig. 7 Choice of the task's competence level depending on the sequence of tasks

descending order (similarly to A and B in Fig. 6). There were also strategies in which each competence level was tested with only one task (similarly to C in Fig. 6), and other strategies in which (almost) all tasks of each competence levels were used, similarly to the participant A in Fig. 6. In 58.1% of the participants, we could identify the strategy patterns A, B and D that tested at least three tasks per level at the minimum of two different competence levels. 42.1% combined the ascending or descending order with the selection of at least three tasks on at least two different competence levels (similarly to patterns A and B in Fig. 6).

Among the participants, 50.7% always chose the competence level of the next task based on the solution of the previous task. Figure 7 illustrates this strategy for one participant. The participant began with two tasks at competence level 3 that the virtual student (who was at competence level 2) answered incorrectly. The participant then chose a task at competence level 1 that the student solved correctly. After that, the participant presented two tasks at competence level 2, which the student solved correctly again. The participant then chose a task at competence level 3 and decided (as we could see in his notes not shown here) that the student must be on competence level 2. The participant used the subsequent tasks to identify the students' primary misconception.

5 Discussion

We presented a simulated computer-based environment to assess prospective teachers' diagnostic processes. The environment was based on a competence model of mathematics in primary schools and included real students' written work. The simulated environment offered the opportunity to measure the activities of diagnostic processes and assess the use of the competence level model. The results give rise to the assumption that the simulated environment can be used to measure aspects of diagnostic processes. We discuss the results in the order of the research questions.

5.1 Acceptance of the simulated environment

Participants rated the simulated environment as authentic and that they experienced presence. We assume that the authentic tasks and solutions in our simulated environment contributed to this result. Participants considered competencies and misconceptions of virtual students with real tasks and had to choose appropriate tasks like a real teacher. Arguably, our simulated environment may have the potential to be used not only for assessing but also for promoting prospective teachers' diagnostic competences, because simulated environments may help prospective teachers to better understand future requirements and support the transfer of their diagnostic competences into real classrooms (Seidel et al. 2010).

One may argue that prospective teachers may be less able to judge the authenticity of the simulated environment than in-service teachers would be. However, the environment is intended to be used primarily with prospective teachers, so that we consider their subjective assessment as most relevant.

5.2 Accuracy of diagnostic results

Participants turned out to be more accurate in diagnosing virtual students' competence levels than specifying students' primary misconceptions: three out of four diagnosed cases were correctly diagnosed in terms of competence levels, but only one out of four diagnosed cases was correctly diagnosed regarding misconceptions. The high accuracy in the diagnosed competence level—up to 94.3%—is not surprising because the mathematical problems that could be

selected in the environment were labeled and sorted according to their level of competence required (i.e., difficulty). Thus, this information was sufficient to evaluate virtual students' competence level. Our results suggest that the more challenging facet is assessment of students' specific errors and misconceptions.

Students who were at the lowest or highest competence level were especially easy to diagnose. This result suggests that prospective teachers are well able to identify students who can solve tasks particularly well or particularly poorly, but they are less able to make more fine-grained distinctions. This effect may have been particularly strong in our study, because we provided the information about mathematical task difficulty before participants needed to choose a specific task. Participants may have relied on the information about task difficulty without necessarily reasoning about students' specific competences and misconceptions. In sum, identifying and understanding students' specific errors could be another starting point for systematic support of preservice teachers' diagnostic competences (Brunner et al. 2011; Helmke et al. 2004).

As Artelt and Rausch (2014) point out, judgement accuracy is only one facet of diagnostic competences. The results measuring the diagnostic competences showed that the participants predominantly evaluated the evidence (i.e., virtual students' written work), while scarcely generating hypotheses or drawing conclusions during the diagnostic process. This seems unfortunate because according to current models of diagnostic processes, problem solving and scientific reasoning would be useful in order to provide accurate judgments (see Heinrichs and Kaiser 2018; Herppich et al. 2017b; Philipp 2018). Our result may provide a first hint that prospective teachers may particularly benefit from supporting these diagnostic activities, although further research is certainly needed.

5.3 Occurrence of diagnostic activities

Participants in our study could use as much information as they thought they might need for diagnosing. We found that many participants collected a substantial amount of information (up to 55 tasks for diagnosing a single virtual student) that was perhaps not actually necessary for reliable diagnoses. Additional analyses also showed that the number of tasks used for diagnosing was unrelated to the accuracy of diagnoses (see also Böhmer et al. 2015). In further studies, it could be interesting to test whether limiting the available number of tasks would force pre-service teachers to make more efficient use of the available data.

In order to encourage participants to reflect on their activities, we asked them to take written notes. We were particularly interested in which and how often diagnostic activities would appear in these written notes. Because we did not provide participants with specific instructions about the type of notes they were supposed to take (in order not to influence their natural diagnostic process), some of the written notes were very short and not very informative. This made coding more difficult, which is reflected in the interrater correlations that were not as high as expected. Further development of the coding scheme is certainly necessary to increase the reliability of assessment of diagnostic activities in future studies.

Results suggest that evaluating evidence was the most frequent diagnostic activity in our sample. This finding corresponds to the importance of collecting information highlighted in various studies (Fischer et al. 2014; Herppich et al. 2017b; Heinrichs and Kaiser 2018; Philipp 2018; Schoenfeld 1985; Rubinstein 1995; Polya 1973; Krell et al. 2018). Because the proportion of evidence evaluation was not related to diagnosis accuracy, we assume that the quality of the evidence evaluations may be more important than their frequency. Deeper analyses of the available evidence by prospective teachers would provide them with more useful hints than focusing only on superficial features. Deeper analyses, however, require theoretical knowledge (e.g., pedagogical content knowledge). The relation between diagnostic activities and professional knowledge is another important issue for further research.

Generating hypotheses rarely occurred in the participants' notes, and drawing conclusions occurred even less, although both activities are important in diagnostic processes (Heinrichs and Kaiser 2018; Herppich et al. 2017b; Philipp 2018) and in problem solving processes or argumentation (Krell et al. 2018; Polya 1973; Rubinstein 1995; Schoenfeld 1985). It would be interesting to test whether instructing prospective teachers more explicitly to making use of generating hypotheses and drawing conclusions would lead to increased use of these diagnostic activities, or whether more content-specific support is required.

5.4 Analyzing diagnostic processes with competence level models

The log data from our simulated environment allowed us to describe, in an explorative approach, patterns in participants' diagnostic processes regarding their use of competence levels. The patterns suggest that participants used a large variety of different strategies, and did not necessarily approach the virtual students' competence level in a systematic way, although we found that processes with accurate diagnoses were, on average, more systematic than processes with inaccurate diagnoses. In further analyses, one could look at the relation between these patterns and participants' diagnostic activities, and

test how the patterns change after training that fostered diagnostic competences.

5.5 Limitations and future perspectives

The simulated environment presented in this paper seems to be a suitable tool for assessing prospective teachers' diagnostic processes. However, several limitations and steps for further development and research remain. In addition to those already discussed in the previous sections, a limitation is that the analyses presented here focused on only two mathematical content domains, number and operations, as well as patterns and structures. The simulated environment would allow extending the content to other domains, although this extension would make it more challenging to provide detailed information about the competence levels of multiple domains.

As a more general challenge, we need further research to better understand diagnostic processes and the difficulties teachers encounter during these processes, in order eventually to develop effective learning environments in support of prospective teachers' diagnostic competences. Further analyses of participants' strategies and written notes as suggested in this paper may eventually provide better insights that could then be used to implement support elements in the simulated environment.

Acknowledgements Open Access funding provided by Projekt DEAL. This project is part of the COSIMA Research group funded by Deutsche Forschungsgemeinschaft (DFG) - German Research Foundation (FOR 2385) to Kristina Reiss, Andreas Obersteiner and Frank Fischer (Project: 5150171 DG RE 1247/12-1).

References

Anders, Y., Kunter, M., Brunner, M., Krauss, S., & Baumert, J. (2010). Diagnostische Fähigkeiten von Mathematiklehrkräften und ihre Auswirkungen auf die Leistungen ihrer Schülerinnen und Schüler. *Psychologie in Erziehung und Unterricht, 57,* 175–193. https://doi.org/10.2378/peu2010.art13d.

Artelt, C., & Rausch, T. (2014). Accuracy of teacher Judgments. In S. Krolak-Schwerdt, S. Glock, & M. Böhmer (Eds.), *Teachers' professional development: Assessment, training, and learning. The future of education research* (Vol. 3, pp. 27–43). Rotterdam: Sense Publishers.

Ashlock, R. B. (2006). *Error patterns in computation. Using error patterns to improve instruction* (9th ed.). Upper Saddle River, NJ: Pearson/Merrill Prentice Hall.

Beitlich, J., Moll, G., Nagel, K., & Reiss, K. (2015). Fehlvorstellungen zum Funktionsbegriff am Beginn des Mathematikstudiums. In M. Gartmeier, H. Gruber, & T. Hascher (Eds.), *Fehler: Ihre Funktionen im Kontext individueller und gesellschaftlicher Entwicklung* (pp. 211–223). Münster: Waxmann Verlag GmbH.

Blomberg, G., Renkl, A., Gamoran Sherin, M., Borko, H., & Seidel, T. (2013). Five research-based heuristics for using video in pre-service teacher education. *Journal for Educational Research Online/ Journal für Bildungsforschung Online, 5*(1), 90–114.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie, 223*(1), 3–13.

Böhmer, I., Hörstermann, T., Gräsel, C., Krolak-Schwerdt, S., & Glock, S. (2015). Eine Analyse der Informationssuche bei der Erstellung der Übergangsempfehlung: Welcher Urteilsregel folgen Lehrkräfte?: An analysis of information search in the process of making school tracking decisions: Which judgment rule do teachers apply? *Journal for Educational Research Online, 7*(2), 59–81.

Brunner, M., Anders, Y., Hachfeld, A., & Krauss, S. (2011). Diagnostische Fähigkeiten von Mathematiklehrkräften. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (pp. 215–234). Münster: Waxmann Verlag GmbH.

Common Core State Standards Initiative. (2012). *Mathematics standards*. Retrieved July 18, 2018, from http://www.corestandards.org/Math/.

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

Derry, S. J., Sherin, M. G., & Sherin, B. L. (2014). Multimedia learning with video. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning. Cambridge handbooks in psychology* (pp. 785–812). New York: Cambridge University Press.

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., et al. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research, 2,* 28–45. https://doi.org/10.14786/flr.v2i2.96.

Fischer, F., & Opitz, A. (Eds.). (2020). *Learning to diagnose with simulations: Examples from teacher education and medical education*. New York: Springer.

Franke, M., Padberg, F., & Ruwisch, S. (2010). *Didaktik des Sachrechnens in der Grundschule*. Heidelberg: Spektrum Akademischer Verlag.

Hasemann, K., & Gasteiger, H. (2014). *Anfangsunterricht Mathematik. Mathematik Primarstufe und Sekundarstufe I + II* (3rd ed.). Berlin: Springer Spektrum.

Heinrichs, H. (2015). *Diagnostische Kompetenz von Mathematik-Lehramtsstudierenden: Messung und Förderung. Perspektiven der Mathematikdidaktik*. Wiesbaden: Springer Spektrum.

Heinrichs, H., & Kaiser, G. (2018). Diagnostic competence for dealing with students' errors: Fostering diagnostic competence in error situations. In T. Leuders, K. Philipp, & J. Leuders (Eds.), *Diagnostic competence of mathematics teachers: Unpacking a complex construct in teacher education and teacher practice. Mathematics teacher education* (Vol. 11, pp. 79–94). Cham: Springer.

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M. R., et al. (2019). Facilitating diagnostic competences in simulations: A conceptual framework and a research agenda for medical and teacher education. *Frontline Learning Research, 7,* 1–24. https://doi.org/10.14786/flr.v7i4.384.

Helmke, A., Hosenfeld, I., & Schrader, F.-W. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz

von Lehrkräften. In R. Arnold & C. Griese (Eds.), *Schulleitung und Schulentwicklung* (pp. 119–144). Baltmannsweiler: Schneider Verlag Hohengehren.

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.

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., et al. (2017a). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education*. https://doi.org/10.1016/j.tate.2017.12.001.

Herppich, S., Praetorius, A.-K., Hetmanek, A., Glogger-Frey, I., Ufer, S., Leutner, D., et al. (2017b). Ein Arbeitsmodell für die empirische Erforschung der diagnostischen Kompetenz von Lehrkräften. In A. Südkamp & A.-K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen. Pädagogische Psychologie und Entwicklungspsychologie* (Vol. 94, pp. 75–94). Münster: Waxmann Verlag GmbH.

Hoffmann, L. (2017). *Lehrkräfte als Diagnostikerinnen und Diagnostiker: Untersuchungen zu ausgewählten Tätigkeiten von Grundschullehrerinnen und Grundschullehrern im Bereich der pädagogisch-psychologischen Diagnostik*. Doctoral dissertation, Universität Potsdam, Humanwissenschaftliche Fakultät.

Holmes, V.-L., Miedema, C., Nieuwkoop, L., & Haugen, N. (2013). Data-driven intervention: Correcting mathematics students' misconceptions, not mistakes. *The Mathematics Educator, 23*(1), 24–44.

Karst, K. (2012). *Kompetenzmodellierung des diagnostischen Urteils von Grundschullehrern. Empirische Erziehungswissenschaft* (Vol. 35). Münster: Waxmann Verlag GmbH.

Karst, K., Schoreit, E., & Lipowsky, F. (2014). Diagnostic competencies of math teachers and their impact on learning development of elementary school children. *Zeitschrift für Pädagogische Psychologie, 28*(4), 237–248.

Klieme, E., Hartig, J., & Rauch, D. (2008). The concept of competence in educational contexts. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (1st ed., pp. 3–22). Göttingen: Hogrefe Publishing.

Klug, J., Bruder, S., Kelava, A., Spiel, C., & Schmitz, B. (2013). Diagnostic competence of teachers: A process model that accounts for diagnosing learning behavior tested by means of a case scenario. *Teaching and Teacher Education, 30,* 38–46. https://doi.org/10.1016/j.tate.2012.10.004.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift für Psychologie/Journal of Psychology, 216,* 61–73. https://doi.org/10.1027/0044-3409.216.2.61.

Krell, M., Redman, C., Mathesius, S., Krüger, D., & van Driel, J. (2018). Assessing pre-service science teachers' scientific reasoning competencies. *Research in Science Education, 41,* 1–25. https://doi.org/10.1007/s11165-018-9780-1.

Kultusministerkonferenz. (2004). *Bildungsstandards im Fach Mathematik für den Primarbereich: Beschluss vom 15.10.2004*.

Leuders, T., Philipp, K., & Leuders, J. (Eds.). (2018). *Diagnostic competence of mathematics teachers: Unpacking a complex construct in teacher education and teacher practice. Mathematics teacher education* (Vol. 11). Cham: Springer.

Padberg, F., & Benz, C. (2011). *Didaktik der Arithmetik: Für Lehrerausbildung und Lehrerfortbildung. Mathematik Primar- und Sekundarstufe I + II: Didaktik der Mathematik* (4th ed.). Heidelberg: Spektrum Akademischer Verlag.

Philipp, K. (2018). Diagnostic competences 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. Mathematics teacher education* (Vol. 11, pp. 109–128). Cham: Springer.

Polya, G. (1973). *How to solve it: A new aspect of mathematical method* (Vol. 246). Princeton, NJ: Princeton University Press.

Radatz, H. (1980). *Fehleranalysen im Mathematikunterricht*. Braunschweig: Vieweg + Teubner Verlag.

Reiss, K., & Obersteiner, A. (2019). Competence models as a basis for defining, understanding, and diagnosing students' mathematical competences. In A. Fritz, V. G. Haase, & P. Räsänen (Eds.), *The international handbook of math learning difficulties: From the laboratory to the classroom*. Cham: Springer.

Reiss, K., & Winkelmann, H. (2009). Kompetenzstufenmodelle für das Fach Mathematik im Primarbereich. In D. Granzer (Ed.), *Bildungsstandards Deutsch und Mathematik: [Leistungsmessung in der Grundschule]. Beltz-Pädagogik* (pp. 120–141). Weinheim: Beltz.

Reiss, K., Roppelt, A., Haag, N., Pant, H.A., & Köller, O. (2012). Kompetenzstufenmodelle im Fach Mathematik. In P. Stanat, H.A. Pant, K. Böhme, & D. Richter (Eds.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (pp. 72–84). Münster: Waxmann.

Rich, P. J., & Hannafin, M. (2009). Video annotation tools: Technologies to scaffold, structure, and transform teacher reflection. *Journal of Teacher Education, 60*(1), 52–67.

Rubinstein, M. F. (1995). *Patterns of problem solving* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Salen, K. (2004). *Rules of play*. Cambridge, MA: MIT Press.

Salen, K., & Zimmerman, E. (2006). *The game design reader: A rules of play anthology*. Cambridge, MA: MIT Press.

Santagata, R. (2005). Practices and beliefs in mistake-handling activities: A video study of Italian and US mathematics lessons. *Teaching and Teacher Education, 21,* 491–508. https://doi.org/10.1016/j.tate.2005.03.004.

Schoenfeld, A. H. (1985). *Mathematical problem solving*. Orlando, FL: Academic Press.

Seidel, T., Blomberg, G., & Stürmer, K. (2010). "Observer"–Validierung eines videobasierten Instruments zur Erfassung der professionellen Wahrnehmung von Unterricht. *Zeitschrift für Pädagogik, 56*(Beiheft), 296–306.

Seidel, T., & Stürmer, K. (2014). Modeling and measuring the structure of professional vision in preservice teachers. *American Educational Research Journal, 51*(4), 739–771. https://doi.org/10.3102/0002831214531321.

Seidel, T., Stürmer, K., Blomberg, G., Kobarg, M., & Schwindt, K. (2011). Teacher learning from analysis of videotaped classroom situations: Does it make a difference whether teachers observe their own teaching or that of others? *Teaching and Teacher Education, 27,* 259–267. https://doi.org/10.1016/j.tate.2010.08.009.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz: Accuracy of teacher judgments on student characteristics and the construct of diagnostic competence. *Zeitschrift für Pädagogische Psychologie, 19*(1/2), 85–95.

Stanat, P. (Ed.). (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik: Ergebnisse des IQB-Ländervergleichs 2011*. Münster: Waxmann Verlag GmbH.

Stokking, K., Leenders, F., de Jong, J., & van Tartwijk, J. (2003). From student to teacher: Reducing practice shock and early dropout in the teaching profession. *European Journal of Teacher Education, 26,* 329–350. https://doi.org/10.1080/0261976032000128175.

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

Südkamp, A., & Praetorius, A.-K. (Eds.). (2017). *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen. Pädagogische Psychologie und Entwicklungspsychologie* (Vol. 94). Münster: Waxmann Verlag GmbH.

Wildgans-Lang, A., Obersteiner, A., & Reiss, K. (2019). Epistemisch-diagnostische Aktivitäten im Diagnoseprozess bei Lehrkräften im Mathematikunterricht. In Ehmke, T., Kuhl, P., Pietsch, M. Lehrer. Bildung. Gestalten. Weinheim Basel: BELTZ Juventa.