Educational Studies in Mathematics (2024) 115:125–149
https://doi.org/10.1007/s10649-023-10278-1

Developing a diagnostic framework for primary and secondary students' reasoning difficulties during mathematical problem solving

Anna Ida Säfström[1,2] · Johan Lithner[1,2] · Torulf Palm[1,2] · Björn Palmberg[1,2] · Johan Sidenvall[1,3] · Catarina Andersson[1,2] · Erika Boström[1,2] · Carina Granberg[1,4]

Accepted: 3 November 2023 / Published online: 19 December 2023

Abstract
It is well-known that a key to promoting students' mathematics learning is to provide opportunities for problem solving and reasoning, but also that maintaining such opportunities in student–teacher interaction is challenging for teachers. In particular, teachers need support for identifying students' specific difficulties, in order to select appropriate feedback that supports students' mathematically founded reasoning without reducing students' responsibility for solving the task. The aim of this study was to develop a diagnostic framework that is functional for identifying, characterising, and communicating about the difficulties students encounter when trying to solve a problem and needing help from the teacher to continue the construction of mathematically founded reasoning. We describe how we reached this aim by devising iterations of design experiments, including 285 examples of students' difficulties from grades 1–12, related to 110 tasks, successively increasing the empirical grounding and theoretical refinement of the framework. The resulting framework includes diagnostic questions, definitions, and indicators for each diagnosis and structures the diagnostic process in two simpler steps with guidelines for difficult cases. The framework therefore has the potential to support teachers both in eliciting evidence about students' reasoning during problem solving and in interpreting this evidence.

Keywords Problem solving · Mathematical reasoning · Diagnostic framework · Students' reasoning difficulties · Formative assessment · Design research

✉ Anna Ida Säfström
anna.ida.safstrom@umu.se

[1] Umeå Mathematics Education Research Centre, Umeå University, Umeå, Sweden

[2] Department of Science and Mathematics Education, Umeå University, Umeå, Sweden

[3] Municipality of Hudiksvall, Hudiksvall, Sweden

[4] Department of Applied Educational Science, Umeå University, Umeå, Sweden

1 Introduction

Problem solving has always been an important object of study within mathematics education research, and it has been known for at least 30 years that increasing opportunities for problem solving is beneficial for students' learning (Boaler, 2002; Lester & Cai, 2016). Problems are meant to require struggle (Hiebert & Grouws, 2007), but given the heterogeneity of mathematics students and teachers' limited time for selection and adaptation of tasks, some students will encounter difficulties that they cannot overcome on their own. To provide appropriate support, the teacher needs to understand the student's specific difficulty, but this can be challenging for teachers (Black & Wiliam, 2009).

There are several models for the complex processes problem solving entails (e.g., Pólya, 1945; Schoenfeld, 1985; Yimer & Ellerton, 2010), but these describe successful processes rather than difficulties. In addition, these frameworks are mostly based on studies of adult students or mathematicians and may not adequately describe what many school students do (Rott et al., 2021). In consequence, they may also be insufficient for characterising the difficulties school students' encounter during problem solving. In this study, we apply cycles of design experiments in collaboration with teachers, developing a diagnostic framework for the difficulties primary and secondary students face when constructing their own solutions to mathematical problems.

2 Background

2.1 Mathematical problems, reasoning, and learning

A mathematical problem is defined as a task for which the student does not have a given solution method, that is, a rule, template, or algorithm (Lithner, 2008; Schoenfeld, 1985). This means that the student must construct the significant parts of the solution themselves, by means of their own mathematical reasoning. Tasks that can be solved by applying given algorithms are denoted routine tasks (Schoenfeld, 1985). Problems and routine tasks elicit different types of reasoning (Boesen et al., 2010; Stein & Smith, 1998; Stein et al., 1996), which support learning to different extents (Brousseau, 1997; Jonsson et al., 2014; Stein & Lane, 1996). Problem solving can be seen as one type of productive struggle (Hiebert & Grouws, 2007) and even necessary for effective learning of mathematics (Schoenfeld, 1985).

We define mathematical reasoning as "the line of thought adopted to produce assertions and reach conclusions in task solving" (Lithner, 2008, p. 257). When solving any task, reasoning includes making a strategy choice and implementing the chosen strategy. Here, "strategy" ranges from local procedures to general approaches, and "choice" is seen in a wide sense, including also mere guesses. The choice can be supported by predictive argumentation—reasons why the strategy will solve the task—and the implementation can be supported by verificative argumentation—reasons why the strategy did solve the task—and a conclusion may be obtained. The properties of the predictive and verificative argumentation characterise different types of reasoning.

Solving routine tasks utilises a pool of algorithms and mental schemes that identify the task type and a suitable algorithm. Two cognitive faculties are required: the ability to identify similarities and the ability to imitate an algorithm (Vinner, 1997). The important point is that the strategy choice can be made through identification of similarities between

the task at hand and tasks with known solution algorithms based on superficial properties, without basing predictive and verificative argumentation on conceptual mathematical understanding (Lithner, 2008). In contrast, problem solving requires more complex cognitive faculties. In terms of mathematical reasoning, solving a problem entails one or more sub-tasks that the student has no given method for (Lithner, 2008). When encountering such sub-tasks, the student cannot successfully apply a solution algorithm based on superficial connections between the sub-task and previously solved tasks but need to construct a solution method on their own. To do so successfully, they have to support their strategy choice and implementation by predictive or verificative argumentation anchored in intrinsic mathematical properties of the mathematical components involved in the task. Thus, the student's reasoning needs to be mathematically founded and is therefore likely to develop mathematical understanding (Lithner, 2017).

2.2 Students' difficulties during problem solving

Whilst problems provide opportunities for students to engage in reasoning based on mathematical meaning, they do not guarantee that students do so. If students are unaccustomed to problem solving or the problem is too challenging, they might resort to imitative approaches—such as trying methods haphazardly or copying a friend's solution—or give up altogether (Boesen et al., 2010; Sidenvall et al., 2015). Students may also engage in mathematically founded reasoning without succeeding in solving the problem. In all these cases, the students have encountered difficulties where they need help to commence or continue their own construction of mathematical reasoning.

Students' difficulties can be characterised in different ways and for different purposes. Students' errors can be analysed to reveal their understanding of concepts (Radatz, 1980) or patterns in their failure to solve word problems (Clements, 1980) and students' solutions can be categorised as showing different levels of skill (Cai et al., 1996). General learning difficulties of struggling students can be studied to reveal cognitive, neurological, and motivational causes. In this study, we do not focus on such underlying causes or levels of difficulties but on the types of specific and in-the-moment difficulties students have whilst constructing their own line of reasoning for the purpose of solving a problem, hereafter denoted *specific reasoning difficulties*. Specific reasoning difficulties can arise at different points in the problem-solving process. This process has often been described in phase models (Rott et al., 2021). The models most widely used are Pólya's (1945) four-phase model and Schoenfeld's (1985) six-phase model, where the latter can be seen as an elaboration of the former.

Schoenfeld's first phase is reading and rereading the task text. This is not seen as a separate phase in any other model and is often disregarded when Schoenfeld's model is used (Rott et al., 2021). However, studies in error analysis of younger students' solutions to word problems find that reading is not trivial for school students (Clements, 1980; Whang, 1996). The second phase is analysis. If the student sees no apparent way to continue the problem solving, they may try to fully understand the problem, select an appropriate perspective, and reformulate the problem in those terms. The third phase is exploration. In this phase, the student explores the situation in the problem to get acquainted with the mathematical properties of concepts and relations and draw conclusions that may be useful for solving the problem, for example, by constructing representations from the information in the problem. In the fourth phase, the planning phase, the student uses conclusions drawn during exploration to plan for a solution method, and in the fifth phase, the solution method

is implemented. The sixth and last phase is verification. In this phase, the student reviews their solution and evaluates the validity of the result.

Phase models explicitly identify places in the problem-solving process where students may run into difficulties. However, most models are theoretical or based on data from adult students or professional mathematicians (Rott et al., 2021), which means that they may not accurately describe the problem-solving processes of younger students. For example, it has been repeatedly observed that students seldom engage in verification (Koichu et al., 2021), and that phases that are trivial for expert problem solvers, such as reading and performing standard calculations, may require considerable effort from school students (Clements, 1980; Wijaya et al., 2014).

2.3 Teacher support during problem solving

The quality of student–teacher interaction and its influence on student learning depends on the teacher's understanding of the student's thinking (Lee & Cross Francis, 2018; Teuscher et al., 2016). When supporting students' problem solving, the teacher needs to understand students' specific reasoning difficulties and be able to differentiate between them. If the teacher does not identify the specific difficulty, they risk either revealing a substantial part of the solution and depriving the student of productive struggle, the responsibility to solve the problem, and the opportunity to learn from it (Brousseau, 1997), or providing too vague feedback that does not help the student's construction of reasoning. Previous studies show that teachers' support often reduces problems to routine tasks during implementation in classrooms, especially for students unused to problem solving, and that this hinders students' engagement in mathematically founded reasoning (Stein & Smith, 1998; Stein et al., 1996). A long-term intention of developing tools for diagnosis of students' specific reasoning difficulties is therefore to help teachers provide feedback that is both sufficiently supportive for students to continue their own reasoning and sufficiently restricted that the responsibility to solve the problems remains with the students.

Identifying students' difficulties for the purpose of providing feedback that supports students' reasoning is a matter of assessing students' learning needs for formative purposes (Black & Wiliam, 2009). This is not an easy endeavour, especially when this assessment is carried out in the complex context of a classroom where other students are competing for the teacher's attention. This assessment involves eliciting evidence about the student's reasoning up to the point when they requested help and about what hinders them from continuing the construction of reasoning (Black & Wiliam, 2009). Such evidence is often not revealed by the student spontaneously and inspecting the student's written work may also be insufficient (Cai et al., 1996). Therefore, the teacher needs to elicit evidence from carefully crafted questions and prompts. The assessment also involves interpreting the available evidence and making inferences about the student's thinking and understanding (Black & Wiliam, 2009). These inferences may then be used for adapting feedback that meets the student's learning needs. Failure in this process of assessment weakens the basis for feedback that is tailored to the student's specific learning needs, and successful student reasoning and learning is therefore less likely to occur (Bennett, 2011; Gu, 2021).

2.4 Support for teachers

To master and successfully carry out the processes involved in formative assessment is difficult and requires a variety of teacher skills (Brookhart, 2011; Datnow & Hubbard,

2016; Gummer & Mandinach, 2015; Means et al., 2011). Consequently, there have been calls to develop support for teachers on both how to elicit evidence about students' thinking and how to interpret this evidence for instructional purposes (Mandinach & Gummer, 2016; Schneider & Gowan, 2013). Still, despite many attempts at professional development aimed at building teachers' capacity to use assessment data for instructional decisions, many teachers feel unprepared to do so (Datnow & Hubbard, 2016). Attempts have been frequently unsuccessful in supporting teachers to develop high-quality formative assessment practises to the extent that increased student achievement was obtained, regardless of school subject (e.g., Bell et al., 2008; Randel et al., 2016; Schneider & Randel, 2010).

Especially, mathematical problem solving is challenging for students (Verschaffel et al., 2020), and supporting students' problem solving is challenging for teachers (Liljedahl & Cai, 2021). Thus, teachers need tools for supporting students' reasoning during problem solving. Some suggestions have emerged about how to elicit student thinking (Shaughnessy et al., 2019) and how teacher knowledge about students' development of specific mathematical competencies may help them to interpret students' thinking (Dindyal et al., 2021). However, empirically grounded frameworks characterising students' specific reasoning difficulties and connecting the elicitation and interpretation of students' thinking to these difficulties are lacking. Such frameworks could be used in professional development initiatives to develop teachers' ability to identify students' reasoning difficulties and in turn help teachers select appropriate feedback in interaction with students.

3 Aim and research questions

We aim to develop a diagnostic framework that is functional for identifying, characterising, and communicating about the specific reasoning difficulties students' encounter when they try to solve a problem and need help from the teacher to continue the construction of their own line of reasoning. In the long term, the framework is intended to "do real design work in generating, selecting and validating design alternatives at the level at which they are consequential for students' learning" (diSessa & Cobb, 2004, p. 80), in our case alternative feedback that supports students' own reasoning. Therefore, the development is guided by assessing functionality based on three criteria:

1. *Range*: To aid generation of feedback, the framework includes a set of diagnoses that characterises the main specific reasoning difficulties students encounter when solving problems.
2. *Differentiation*: To aid selection of feedback, the framework separates difficulties that are theoretically and observably different in the sense that students need help with different aspects of their line of reasoning—theoretically, in the sense that definitions of diagnoses are mutually exclusive, and observably, in the sense that the indicators of each diagnosis are observable in the classroom.
3. *Operationalisability*: To aid validation of feedback, the structure of the framework and the definitions of diagnoses provide substantial support for diagnosing students' specific reasoning difficulties.

Our aim is therefore specified in three research questions related to the three criteria:

1. What are the main specific reasoning difficulties students encounter when solving problems, and how can they be incorporated into the framework?
2. How can difficulties that are theoretically and observably different, in the sense that students need help with different aspects of their line of reasoning, be differentiated?
3. How can the structure and definitions of the framework support the diagnostic process by relating it to indicators of diagnoses that are observable in classroom situations?

4 Methods

The study was conducted during three semesters of a longer design-research project where a research group collaborated with teacher teams at seven Swedish schools, spanning grades 1–12. During this phase of the project, teacher–researcher meetings focussed on diagnosis of students' specific reasoning difficulties. As we aimed to develop a framework that is functional for teachers and researchers, we devised a series of design experiments. A collaborative, iterative, and experimental approach was devised, as it can promote the grounding of the framework in real-world experiences by providing multiple exposures to empirical testing and theoretical refinement, and thereby develop a framework for diagnosing difficulties that is robust in its application across contexts (diSessa & Cobb, 2004).

4.1 Initial framework

Design experiments typically test and develop theoretical models for how learning occurs and can be supported and thus start out from conjectures about how means of support affect learning (Cobb et al., 2003). In this study, we use design experiments to test and develop a framework for students' difficulties during problem solving, and therefore, we start out from conjectures about what such a framework should look like, informed by previous research and pilot studies.

The initial framework builds on earlier research on characteristics of and causes behind learning difficulties related to mathematical reasoning, see Lithner (2017) for an overview. In order to study means for remedying these difficulties, we started in smaller scale with design experiments testing various frameworks for supporting students' problem solving (for an example, see Sidenvall et al., 2022). In the present study, we build on these experiments. A starting point for the study was that the main question in characterising a student's reasoning difficulty is where in the problem-solving process the student's difficulty emerged. The first dimension of the framework therefore concerned phases of problem solving (Rott et al., 2021). Various drafts for phases were considered and tested before the initial framework for this study was formulated, entailing four phases:

1. *Interpret* the information given in the problem, what is asked for and what type of answer is required. This does not include creating a solution, even if it is sometimes done in parallel with interpretation.
2. *Explore* the information and analyse mathematical properties of concepts and connections, construct useful representations, and try to draw conclusions that may be useful for solving the problem. This also includes trying to relate the problem to existing knowledge and previous experiences that could be useful.
3. *Create a solution idea* by analysing properties (perhaps found during the exploration) of the problem and formulating and evaluating hypotheses for a solution. The difference

between exploring and creating a solution idea is that the former concerns understanding the mathematics of the problem, whilst the latter means formulating basic ideas for the whole solution.

4. *Utilise the solution idea*. Sometimes, but not always, substantial work remains to utilise the solution idea and produce an answer in the expected manner, for example, complex calculations or interpretation of the result in a real-life context.

The second dimension concerned the strategy choices, implementations, and verifications that students' own constructions of reasoning require in each phase (Lithner, 2008). Theoretically, each phase could constitute a task where students have to apply reasoning, including making a strategy choice, implementing that strategy, and, if needed, constructing verificative arguments for the conclusions obtained. Pilot studies also indicated that students not only have difficulties implementing each phase but also selecting a suitable phase and verifying their conclusions. Therefore, we introduced three sub-phases:

A. *Initiate* or select a suitable phase.
B. *Implement* the phase.
C. *Evaluate* whether the initiation and implementation of the phase is correct and useful.

Our initial framework differed from Schoenfeld's six phases in three main ways. First, the planning phase (Pólya, 1945; Schoenfeld, 1985) characterises what expert problem solvers do when they meet a complex problem. Such planning requires active monitoring and control, mastering a battery of strategies and vast experiences of using these strategies in various problems. However desirable, this is rarely done by students in primary and secondary school. Instead, we proposed the phase "create solution idea," inspired by Lampert's (1990) emphasis on the importance for all students to make and test hypotheses in problem solving. Second, Schoenfeld's (1985) verification phase was replaced by a sub-phase, as evaluation of the progress can be needed for each phase, not only for the whole solution. Third, it was not seen as practically possible to distinguish between the analysis and exploration phases, which were therefore merged. Also, as the purpose of the framework was to differentiate between types of difficulties, the ambition was to keep overlaps as limited as possible. This led to partially different definitions compared to Schoenfeld's (1985) phases. In sum, our initial model of students' problem-solving processes consisted of 12 sequential steps: 1A–1B–1C–2A–2B–2C–3A–3B–3C–4A–4B–4C, where some steps can be omitted depending on the problem and the student's line of reasoning.

Thus, the framework included 12 hypothetical diagnoses, which all were possible answers to the main question: in what phase and in what sub-phase has the student's difficulty emerged? For example, diagnosis 2A meant that the difficulty had emerged in the initiation of exploration, whilst diagnosis 3B meant that the difficulty had emerged in the implementation of creating a solution idea. As an aid for answering the main question and making a diagnosis, three diagnostic questions (DQs) for eliciting evidence about students' thinking were also included in the framework.

Besides the structural overview presented in Fig. 1, the initial framework comprised 22 pages including descriptions of the purpose of the framework; principles of formative assessment; problem-solving phases, sub-phases, and competencies; and examples of diagnoses in relation to different problems with suggested feedback.

Main question:	In what phase and in what sub-phase has the student's difficulty emerged?			
Diagnostic questions:	DQ1: What does the task ask you to do? DQ2: What have you done so far? DQ3: Why did you do that?			
Phases / Sub-phases	**1 Interpret**	**2 Explore**	**3 Create solution idea**	**4 Utilise solution idea**
A Initiate	1A The student is not attempting to interpret the task, either by not even reading the task, or by reading superficially and giving up directly without really trying to understand the information and what is requested.	2A The student has made a good (correct and complete) interpretation of the task. The student is either stuck after the interpretation ("I don't know how to start") or has skipped exploration, even though it would be valuable, for example, by trying to recall a solution procedure.	3A The student has tried to explore the task but does not try to create a solution idea.	4A The student has a good solution idea, but does not try to utilise the idea and complete the solution.
B Implement	1B The student tries but gets stuck during interpretation. For example, because the task itself is difficult to interpret, the student has misunderstood information, the student has made a superficial/careless interpretation or because the student has made an incorrect interpretation that leads to a contradiction or an absurdity.	2B The student tries to explore but does not find a suitable strategy. The student may also find it difficult to see what an exploration will lead to, and/or to draw conclusions that are useful for creating a solution idea in a later phase.	3B The student tries to create a solution idea but does not find an appropriate strategy. It may also be that the student has tried to explore but has not been able to draw any useful conclusions, and the teacher believes that more task-specific feedback is needed.	4B The student has a good solution idea but is unable to utilise the idea and complete the solution.
C Evaluate	1C The student has started or completed the phase, but either does not see that the implementation is incorrect or inefficient; knows that something is wrong but cannot or does not try to understand why it is wrong, even though it would be valuable to know; or is unsure whether something is right or wrong but cannot or does not want to determine if this is the case.	2C	3C	4C

Fig. 1 The structure of the initial framework

4.2 Data collection and analysis

The study entailed five iterations of data collection, data analysis, and revisions of the framework. Iterations 1–2 focussed mainly on the development of the analysis method. Each subsequent iteration entailed six steps. Figure 2 gives an overview of how each step of the analysis used the result from the previous step as data, forming a recursive process that established a chain of inferences that stayed close to the data in each step (Simon, 2019). Steps 1–6 are described in Sects. 4.2.1–4.2.6.

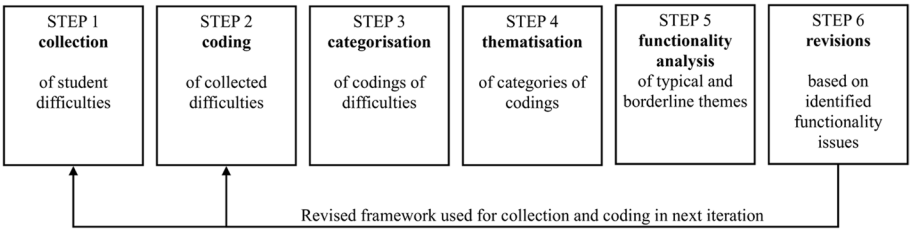


Fig. 2 An overview of the six steps of data collection, analysis, and revisions

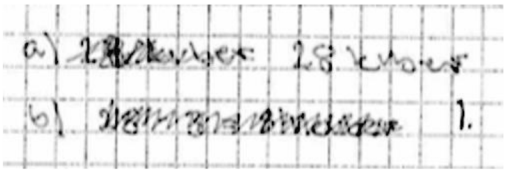
4.2.1 Collecting data on student difficulties

In Step 1, data on students' difficulties were collected from teachers' documentation of their own problem-solving lessons and consisted of the problem used and descriptions of 1–4 student difficulties selected by the teacher, often accompanied by photos of students' work (Fig. 3). Teachers were asked to select both examples of students' difficulties that they found easy to diagnose and examples that they found difficult. Between iterations, researchers reported on the distribution of diagnoses found in the latest iteration and asked teachers to look for difficulties that were scarce. During teacher–researcher meetings, attending researchers took notes on additional information given verbally by the teacher. One attending researcher then assembled the teacher documentation and notes from the meeting. Over all iterations, 285 difficulties (Grades 1–3: 73, Grades 4–6: 98, Grades 7–9: 64, Grades 10–12: 50) were collected from 152 lessons regarding 110 tasks.

Example 1 (from iteration 4, grade 10)

TASK:

a) How many cubes are needed to build the tower in the picture?
b) How many cubes are needed to build a similar tower that is 12 cubes high?

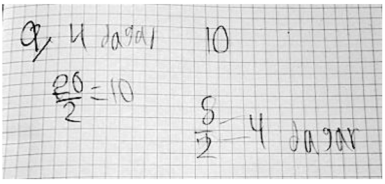


[The students' notes show the following:
a) (Crossed out text) 28 cubes.
b) (Crossed out text)]

Written teacher description: The students understand the task but didn't make the effort to explore. Easy [to diagnose].
Verbal teacher description: I asked the first [diagnostic] question. I think they had understood the task from what they said. But then nothing more happened on the second task. They couldn't make the effort; that's why I said [redacted]. I think they understood what they should find out, but they didn't start. Some other students I helped; there, I wasn't so sure if they had understood how the tower was constructed. But I think these students had.

Example 2 (from iteration 3, grade 4)

TASK: A snail crawled the same distance each day. After eight days it had crawled 20 cm.
a) How far had it crawled after 12 days?
b) How many days did it take before it had crawled 35 cm?
c) How many days did it take before it had crawled 47½ cm?'



[The student's notes show the following:
a) 4 days 10
20/2 = 10 8/2 = 4 days]

Written teacher description: These students I put in [redacted], since they had started [redacted] but got stuck when finishing the task.

Fig. 3 Two examples of student difficulties

Table 1 Number of discarded, typical, and borderline cases in Iterations 3–5

	Discarded step 2	Discarded step 3	Typical	Borderline	Total
Iteration 3	16	9	25	32	82
Iteration 4	0	12	47	19	78
Iteration 5	0	1	19	1	21

Typical and borderline cases are discussed in Sect. 4.2.3

In Iterations 1–3, a lot of data lacked teachers' arguments for diagnoses and sometimes even a diagnosis attempt, and about a fifth of the examples in Iteration 3 were discarded in Step 2 due to insufficient data (Table 1). Therefore, in Iterations 4–5, difficulties were only included if:

- the teacher had either diagnosed the difficulty, with one or more diagnoses, or explicitly stated that they could not determine the diagnosis,
- it was apparent what evidence the teacher had based their diagnosis on, and
- the teacher stated some kind of argument for their diagnosis or inability to diagnose, for example, "The student said … therefore I chose 2C".

4.2.2 Coding each difficulty

In Step 2, each difficulty was coded by teachers and researchers (referred to as *coders*). The teachers coded their own students' difficulties, using the latest version of the framework to diagnose the difficulty before the meetings. Teachers' codings were thus based on their documentation but may also have been informed by general knowledge about their students and memories of the classroom situation. Their codings included their suggested diagnosis, their written arguments, and researcher notes on any additional arguments given verbally by the teacher during meetings. The researchers used the teachers' documentation as data, with teachers' diagnoses obscured (see Fig. 3), and the latest version of the framework and a coding protocol asking the researchers to describe how they understood the situation as a whole and to answer the main question ("Where has the difficulty emerged?" in Iterations 1–2 and "What specific difficulty does the student need help with?" in Iterations 3–5) with either one of the diagnoses included in the framework; *other*, if no diagnosis in the framework fitted the difficulty; *no idea*, if no substantiation was found for any diagnosis; or *no difficulty*, if the student did not seem to have any difficulty. Coders could suggest more than one diagnosis if several fitted the difficulty. Researchers' codings also included whether they were sure or unsure regarding their diagnoses as well as justifications of diagnosis and level of certainty using explicit references to data and formulations in the framework (Table 2).

The result of Step 2 was a set of coded difficulties. Over all iterations, 38 teachers and 8 researchers participated in this step, making 697 codings (202 by teachers and 495 by researchers). The number of difficulties categorised as each diagnosis by at least one coder is displayed in Table 3.

Table 2 Codings of the difficulties presented in Fig. 3

Example 1	
Teacher	I asked the first [diagnostic] question. I think they had understood the task from what they said. But then nothing more happened on the second task. They couldn't make the effort; that's why I said 2A. I think they understood what they should find out, but they didn't start
Researcher 1	The student has solved the first sub-task, but then stopped and not made any effort on the next sub-task. The teacher's assessment is that the student does not even have the energy to try 2A (sure). Empirical basis: Mainly the teacher's assessment that they have not done anything, didn't have the energy to try, and the evidence that they did answer DQ1 good enough. Framework basis: "The student's answer to DQ1 shows that they have understood the explicit information in the task … and have not done anything else."
Researcher 2	The student has solved sub-task (a) and thus probably interpreted the task but does not get started with sub-task (b) 2A (sure). Empirical basis: That the teacher describes: "understands the task but does not make the effort to explore." Framework basis: "The student's answer to DQ1 shows that they have understood the explicit information in the task and/or the student has written the important information in the task and has not done anything else."
Example 2	
Teacher	These students I put in 2B since they had started exploring but got stuck when finishing the task
Researcher 1	The student has, on sub-task (a), drawn the conclusion that the snail crawls 10 cm in 4 days but then gotten stuck. I see this as an exploration, that the student has not been able to utilise a solution idea, but there is no evidence about the student's reasoning 3A (unsure). Empirical basis: See above. Framework basis: "3A: The student has implemented a good exploration of the task but does not try to create a solution idea." Empirical sources of uncertainty: Data on the student's reasoning is missing. I do not know if the student tries and fails or if they do not try
Researcher 2	The student has correctly calculated how far the snail crawls in 4 days, which can be used for solving sub-task a) but seems to make no progress after that 2B, 3A, 3B (unsure). Empirical basis: The student has done something that is correct and useful. Framework basis: 2B: "The student can also have a hard time drawing conclusions that in the next phase could be useful for starting to create a solution idea." 3A: "The student has implemented a good exploration but does not try to create a solution idea." 3B: "It can also be the case that the student has explored but failed to draw any sufficiently useful conclusions in order to create a solution idea." Empirical sources of uncertainty: No evidence regarding the student's description of what they have done or why. Framework sources of uncertainty: Overlap between 2 and 3B

Table 3 Number of difficulties categorised as different diagnoses by at least one coder

		1	2	3	4
Iterations 1–4	A	27	68	24	5
	B	77	88	51	35
	C	88	69	157	159
Iteration 5	C'	35			
	D	0	0	4	1
Iterations 1–5	Other	17			
	No idea	18			
	No difficulty	9			

Note. The total sum exceeds 285 since coders could give several codes and different coders could give different codes for the same difficulty

4.2.3 Comparing and categorising coded difficulties

In Step 3, the different codings of each difficulty were compared and categorised. In Iterations 3–5, two researchers did this step independently and then compared and discussed their categorisation to reach agreement. Some coded difficulties were also discarded in this step (Table 1), mainly because the student did not have a difficulty, or because the data were insufficient to draw conclusions regarding the cause of coders' disagreements. The remaining coded difficulties were summarised and categorised as either a typical example of a diagnosis (including the diagnosis other) if all coders with substantial justification agreed, or a borderline case between two or more diagnoses if coders with substantial justification disagreed (Tables 1 and 4). The result of this step was a categorisation of the diagnosed difficulties as either typical or borderline and summaries of the identified sources for agreement or disagreement for each diagnosed difficulty.

4.2.4 Thematising categorised codings

In Step 4, the categorised codings were thematised. The typical examples were thematised based on the observable indicators the coders referred to. The name of each theme consisted of the diagnosis and a description of common indicators of the difficulties within this theme (Table 5, Example 1). There could therefore be several themes for each diagnosis. The borderline cases were thematised based on the suggested diagnoses and the sources of disagreement, specifically whether the source was insufficient empirical data, formulations in the framework, or circumstances of the situation (Table 5, Example 2). The result of this step was a set of themes: some describing indicators for diagnoses and some describing obstacles to diagnosis.

Table 4 The categorisation of the coded difficulties presented in Table 2

Example 1	
Researcher 1	Typical 2A: All coders agree on 2A. The researchers base their diagnosis on the evidence that the student could answer DQ1 but has not tried anything
Researcher 2	Typical 2A: The researchers agree on 2A. Both refer to the teacher's description that the student has understood the task but has not made the effort to try anything, and to the formulation, "The student's answer to DQ1 shows that they have understood the explicit information given in the task … and has not done anything else." The teacher says 2A
Example 2	
Researcher 1	Borderline: One researcher says 2B, 3A, or 3B, referring to overlaps between the definitions of these diagnoses. One researcher says 3A. Both researchers are missing info on the student's reasoning (DQ2 and 3). The teacher says 2B but has given no arguments for this diagnosis
Researcher 2	Borderline: One researcher says 2B or 3AB based on overlaps in the diagnosis definitions. The other researcher says 3A. Both researchers state that there is a lack of empirical evidence

Table 5 The themes in which the difficulties presented in Fig. 3 were included

Type of theme	Example of theme
Typical example theme	2A: The student is able to answer DQ1 but shows that they have not tried anything to start solving the task (this theme included example 1, Fig. 3)
Borderline case theme	2B or 3AB (framework): The definitions of 2B, 3A, and 3B produce overlaps between these diagnoses, as they all currently include some, but not sufficient, exploration by the student (this theme included example 2, Fig. 3)

4.2.5 Analysing functionality

In Step 5, the themes were analysed to identify shortfalls of functionality of the framework. The analysis was guided by a series of questions related to the three functionality criteria described in Sect. 3 (Table 6).

In Iterations 3–5, Steps 4–5 (Fig. 2) were made independently by two researchers, who then compared their results and solved disagreements by discussion. The result of Step 5 was a set of issues regarding the functionality of the framework. For example, in the case of the theme including Example 2 (Table 5), the issue was that the current definitions of diagnoses 2B, 3A, and 3B overlapped, constituting one answer to Analytic question 3 (Table 6).

4.2.6 Revising the framework, reflecting with teachers, and decisions for the next iteration

In Step 6, revisions that addressed the issues identified in Step 5 were formulated. The rationale for each revision was based on the identified themes. In the case of revisions regarding borderline issues, the potential effect of the revision was evaluated by describing whether and how each borderline case would be remedied by the revision. Revisions were also informed by teachers' suggestions during teacher–researcher meetings. The revisions, their rationale, and potential effects were then critically assessed by another researcher, who also considered adverse side effects. Only revisions that were based on several difficulties and could be seen to remedy several borderline examples without affecting typical examples were incorporated into the framework.

When revisions of a particular diagnosis stopped or became marginal, theoretical saturation (Eisenhardt, 1989) was considered reached, and cases of this diagnosis were no longer selected for analysis. Theoretical saturation for initiate and implement diagnoses (Fig. 1) was reached in iteration 4. Therefore, only evaluate diagnoses (Fig. 1) were included in iteration 5, after which theoretical saturation was reached for them as well, and the study ended.

Table 6 Questions guiding the analysis in Step 5

	Analytic question	Rationale	Potential issues
Range	1. What characterises (any) difficulties that are not covered by the framework?	Based on how coders describe these difficulties, range should be increased by incorporating difficulties not covered in the framework through adjusting definitions or structure	Too narrow definitions or missing diagnoses
Differentiation	2. Are there qualitative differences between themes of typical examples within one diagnosis that motivate partitioning of the diagnosis, in the sense that students in fact need help with different things?	A diagnosis should not cover difficulties where students need help with different things, since the framework then would not help teachers select a suitable way of supporting the student and would thus lack differentiation	Theoretically and/or observably different difficulties are not separated
	3. What formulations in the framework are sources of uncertainty and disagreements?	Formulations that produce theoretical overlaps between two or more diagnoses hinder differentiation. Inconsistent formulations can point in opposite directions and produce multiple diagnoses for the same difficulty	Theoretically inseparable diagnoses, need for clarifications, and redefining boundaries
	4. What characteristics of the situation are sources of uncertainty and disagreements?	If diagnoses are only theoretically different, whilst always or often practically co-occurring, the framework does not sufficiently differentiate between difficulties	Observably inseparable diagnoses and/or lack of guidelines for prioritising
	5. What irretrievable evidence about the students' difficulties are sources of uncertainty and disagreements?	If the evidence required to decide between two or more diagnoses is unlikely or impossible to be accessed by a teacher in the classroom, the situation is essentially the same as above (Analytic question 4)	Observably inseparable diagnoses and/or lack of guidelines for prioritising
Operationalisability	6. What formulations in the framework and what types of empirical data do coders refer to regarding difficulties belonging to the typical themes of each diagnosis?	The answer to this question is a set of indicators and examples for each diagnosis. Observable indicators and examples increase operationalisability by supporting diagnosis, since they focus attention on salient details and support validation by linking data to definitions	Missing indicators and examples
	7. What retrievable evidence about the students' difficulties are sources of uncertainty and disagreements?	Answers to this question could also contribute to the set of indicators and increase operationalisability as described above (Analytic question 6)	Missing indicators and examples

The first column indicates what functionality criterion (and thus which research questions) is in focus. The second column contains the corresponding analytic questions. The third column explains the rationale for the question, and the fourth column states the types of issues that can be revealed as a result

5 Results

In this section, we first present the final framework that was the result of Iteration 5. The final framework consisted of the structural overview presented in Fig. 4 and 24 pages including elaborated descriptions and examples. The research questions are mainly answered by the structure and elements of this framework. The rationale for the revisions is then described in terms of how range, differentiation, and operationalisability were assessed and developed through the iterations, further elaborating the answer to each research question.

5.1 The final framework and summary of results

Through the iterative analyses, some of the aspects of the initial framework proved to be functional, whilst other aspects were developed or altered due to identified issues with range, differentiation, or operationalisability. The four phases were found to adequately capture students' problem-solving processes, and no phases were added or removed. Teachers in lower grades did, however, note that their students seldom got stuck in phases 3 and 4, since exploration of their problems often led directly to an answer. The most substantial finding was the increased functionality obtained by restructuring the initial sub-phases A, B, and C (Fig. 1), into two main types of difficulties: stuck and needing evaluation (Fig. 4). This led to a reconceptualisation of the sub-phases as difficulty types, a revision of the main question and the introduction of three sub-questions with three auxiliary guidelines. These elements provided additional support for the diagnostic process by structuring it in two simpler steps and reflecting the different logics of specifying the diagnosis within each difficulty type (Fig. 4, Sects. 5.3.1). Within the needing-evaluation type, the need for a non-phase-specific diagnosis was also discovered (Sect. 5.3.2). In addition to these more substantial revisions, the analyses resulted in removal of one diagnosis (Sect. 5.2), clarifications of diagnoses including both phases and types (Sect. 5.3.1), and introduction of connections to diagnostic questions and observable indicators for all diagnoses (Sect. 5.4). These revisions and their empirical bases are further elaborated below.

5.2 Range

The analyses revealed no need to include additional phases or difficulty types but led to widening of diagnosis C and the removal of diagnosis 4A (Fig. 4). Seventeen difficulties (6%) were coded as other by one or more coders, but only three difficulties were categorised as typical examples of other. The first such difficulty concerned communication of the solution, which could be argued to not be a reasoning difficulty, as it is not necessary to solve a problem. However, as the two additional examples were collected and analysed, it became clear that these difficulties did not mainly concern students' inability or unwillingness to communicate their solution, but students obtaining an answer without constructing a full line of reasoning to serve as a foundation for it. One student had obtained an answer using a method that they did not understand, and another had very quickly obtained an answer but could not explain how. Therefore, cases where the student could not explain or argue for their solution, but nonetheless asked the teacher to verify their answer, were included in diagnosis C (Fig. 4).

Diagnostic questions:	DQ1: What does the task ask you to do?	DQ2: What have you done so far? DQ3: Why did you do that?		
Main question:	What specific difficulty does the student need help with? **Sub-question 1:** Is the student stuck or in need of evaluation? **Guideline 1:** When the student is both stuck and needing evaluation, choose needing evaluation (C or D).			
Types \ **Phases**	**1 Interpret**	**2 Explore**	**3 Create solution idea**	**4 Utilise solution idea**
The student is **stuck**	**Sub-question 2a:** Where is the student stuck? **Guideline 2a:** If there are several potential stuck diagnoses, choose the earliest diagnosis.			
A before the phase	1A *Student does not initiate interpretation* The student is not attempting to interpret the task. Either by not even reading the task, or by reading superficially and giving up directly without really trying to understand the information and what is requested. The student cannot answer DQ1 at all.	2A *Student does not initiate exploration* The student's answer to DQ1 (or equivalent information) shows that they have understood the explicit information in the task and/or the student has written down the important information in the task, and has done nothing more (answers, e.g., "I do not know how to start" or "nothing" to DQ2). The student has thus made a good interpretation of the task but is stuck after the interpretation.	3A *Student does not initiate creating solution idea* The student's answers to DQ2&3 (or equivalent information) shows that they have conducted a reasonable exploration but have not tried to create a solution idea. That is, the student has formulated relationships and characteristics, tried examples, and/or created representations that are sufficient to create a solution idea, but have not performed an overview and analysis of what they have done to see if it can provide the basis for a solution idea.	[redacted]
B in the phase	1B *Student tries, but fails to interpret the task* The student tries but gets stuck in the interpretation of the task. This is indicated by the student only partially answering DQ1 or having a question about some part of the task formulation.	2B *Student tries, but fails to explore* The student's answer to DQ1 (or equivalent information) shows that the student has understood the explicit information in the task and/or the student has written down the important information in the task. The student's answer to DQ2 (or equivalent information) shows that they have tried to explore but have not found an appropriate strategy, or an appropriate way to apply a strategy. The student may also have drawn some conclusions about the mathematical characteristics and relationships involved in the task, but not enough to create a solution idea.	3B *Student tries, but fails to create a solution idea* The student's answers to DQ2&3 (or equivalent information) shows that they have conducted a reasonable exploration and started trying to create a solution idea. That is, the student has formulated relationships and characteristics, tried examples, and/or created representations that are sufficient to create a solution idea, and started trying to perform an overview and analysis of what they have done to see if it can provide the basis for a solution idea. It may also be that the student has evaluated a previous solution idea, identified an error, and understood the reason for the error, and now fails to revise the incorrect solution idea to a correct one.	4B *Student tries, but fails to utilise the solution idea* The student's answer to DQ3 (or equivalent information) shows that they have a good solution idea, and the answer to DQ2 (or equivalent information) shows that the student has tried to start utilising the idea and completing the solution but has not succeeded.
The student is **needing evaluation**	**Sub-question 2b:** What does the student need help to evaluate? **Guideline 2b:** If there are several potential needing-evaluation diagnoses, select diagnosis C.			
C the whole solution	C *The student needs help evaluating everything they have done so far* There are several different situations where the student needs to evaluate their entire solution, or the entire part of a solution that the student has come up with so far. On the one hand, it can be situations where the student does not see any need to evaluate themselves, for example: • There is an error in the solution, but the student believes that everything is correct. • The student has solved the task (perhaps correctly) without justifying or explaining his or her solution and wants the teacher to confirm that it is correct. On the other hand, there can be situations where the student wants to evaluate the entire solution themselves, but does not know how, for example: • The student is unsure whether the solution is right or wrong. • The student sees that something in the solution is wrong, but not what. This diagnosis also includes cases where the student has made a specific error that you as a teacher have identified, but which the student has not yet discovered. This is included in this diagnosis, as phase-specific feedback risks taking over responsibility unnecessarily from the student.			
D a specific phase	1D *The student needs help evaluating their interpretation* The student expresses uncertainty about having interpreted the task correctly or wants confirmation from the teacher of having interpreted it correctly.	2D *The student needs help evaluating their exploration* The student expresses uncertainty about whether all or part of the exploration is right or wrong, wants confirmation from the teacher that the exploration is right, or knows that something is wrong but not why.	3D *The student needs help evaluating their solution idea* The student expresses uncertainty about whether all or part of the solution idea is right or wrong, wants confirmation from the teacher that the solution idea is right, or knows that something is wrong but not why.	4D *The student needs help evaluating their utilisation of their solution idea* The student expresses uncertainty as to whether all or part of the utilisation of the solution idea is right or wrong, wants confirmation from the teacher that the utilisation is right, or knows that something is wrong but not why.

Fig. 4 An overview of the structure, main and sub-questions, guidelines, and diagnosis definitions of the final framework

Diagnosis 4A was removed from the framework since no typical examples of this diagnosis were found. The absence of such difficulties is theoretically reasonable, since it is unlikely that someone who has a good solution idea will not try to utilise it. Whenever a student had a viable solution idea but did not utilise it, the student was uncertain regarding the correctness of the idea, which falls under diagnosis 3D.

5.3 Differentiation

The most commonly found issues regarded differentiation between diagnoses. Solving these issues often required several subsequent iterations, either because initial attempts to revise the framework were found insufficient or caused new issues, or because the results of one iteration did not provide sufficient guidance for revisions to be made. However, the only cause for separation of diagnoses (Analytic question 2, Table 6) was found in Iteration 2, where the evaluated diagnoses were divided into three sub-diagnoses for each phase: the student has made an error or mistake that they have not identified, the student knows something is wrong but not what or why, and the student is unsure whether something is right or wrong. However, this intermediate structure for evaluation diagnoses caused new issues and was further revised (Sect. 5.3.2). We therefore focus on Analytic questions 3–5 below.

5.3.1 Framework formulations producing uncertainty or disagreement

Some of the initial definitions and the initial main question gave rise to uncertainty regarding how diagnoses were to be separated. For example, the diagnosis initiate exploration (2A, Fig. 1) was defined as students not trying to explore. Such situations often entailed students having jumped ahead to an unsubstantiated solution idea, which caused an overlap with evaluate solution idea (3D). Issues with differentiation due to unclear or overlapping definitions of diagnoses were mainly found and remedied in Iteration 2, though the definitions were further developed as additional examples gave rise to clearer links between diagnostic questions and diagnoses in Iterations 3–4.

The evaluate diagnoses (1–4C, Fig. 1) were initially defined as the student needing to evaluate a specific phase. In contrast, the initial main question (where did the student's difficulty emerge?) and the term "sub-phase" indicated that the evaluate diagnoses concerned difficulties emerging whilst evaluating. This resulted in disagreements regarding the common situation where a student had neither identified an error nor attempted evaluation. Researchers diagnosed this as evaluate, but teachers did not. This issue indicated the existence of two main types of difficulties, for which the process of diagnosis followed two different logics: stuck and needing evaluation. When being stuck, everything the student has done so far is correct or already evaluated, but the student does not know how to proceed, and diagnosis is a question of where the student is stuck in the sequence 1A–1B–2A–2B–3A–3B–4B. When needing evaluation, there is an error or uncertainty regarding the correctness of (some part of) the solution, and diagnosis is a question of what the student needs to evaluate, not how far they have come in the solution process. These two difficulty types, as well as a new main question and three sub-questions (Fig. 4), were introduced and further developed in successive iterations.

5.3.2 Differences between diagnoses merely theoretical or based on irretrievable evidence

The intermediate restructuring of the needing-evaluation difficulties (Sect. 5.3) was theoretically clear but proved to require irretrievable evidence regarding the student's state and thinking. First, it was not always possible to identify the need for evaluating a specific phase. If the student had made an undetected error, determining where that error had occurred was often impossible without engaging the student in evaluation of the whole solution, resulting in a catch-22 between eliciting evidence and providing feedback. For example, students often presented solution ideas that were incompatible with the information in the task (Fig. 5). In those cases, the error could be a misinterpretation of the task (Phase 1), an incorrect conclusion drawn during exploration (Phase 2), or an incorrect solution idea (Phase 3). Also, when the student was uncertain regarding the correctness of the solution, they often expressed a general uncertainty rather than uncertainty regarding a specific phase. These two issues were solved by introducing a general diagnosis for when the student needed help evaluating the whole solution or what they had done so far (C, Fig. 4). Second, the difference between diagnoses in the intermediate framework (Sect. 5.3) was not empirically valid, since students often expressed a mixture of uncertainty and knowing a certain phase was wrong but not why, but still had the same difficulty: needing evaluation of the phase. This issue was remedied by redefining the phase-specific needing-evaluation diagnoses (1–4D, Fig. 4) to include both uncertainty regarding a phase and knowing a phase was wrong but not why.

One situation where coders disagreed was when the student had made an incorrect solution attempt that could be interpreted as haphazard and perceived themselves as stuck. In those cases, coders disagreed on whether the student needed help evaluating the incorrect solution idea or starting over with another approach. This issue was resolved by introducing Guideline 1: "When the student is both stuck and needing evaluation, choose needing evaluation" (Fig. 4). This guideline was justified by our focus on reasoning difficulties and the long-term aim of the framework: to help teachers generate, select, and validate feedback that support students in continuing their own reasoning. Helping a student start over without first evaluating their mistake or uncertainty would not support the student's own construction of reasoning but rather make them adopt a new line of reasoning, often proposed by the teacher.

Task

Elin picks 7 toy animals from a box. She picks dogs and roosters. The animals have 18 legs altogether.

How many dogs does she pick?
How many roosters does she pick?

Note: there were not enough dogs and roosters in our box of animals, so the students were told to use four-legged and two-legged animals and pretend they were dogs and roosters.

Solution

There are seven animals, but the number of legs is not correct. They have not checked that both conditions are fulfilled. They thought they had found the solution.

Fig. 5 An example of an incorrect solution idea, which could be caused by misinterpretation (e.g., not reading the whole text), an incorrect conclusion drawn in exploration, or an incorrect solution idea

In Iteration 4, after the elaborations of stuck diagnoses, the borderline cases for these diagnoses were few (5 of 78) and mostly concerned difficulties in separating adjacent diagnoses, such as 2A and 2B or 2B and 3A. Disagreements or uncertainties concerned whether the student had tried to initiate the phase or not, or where the line between explore and create a solution idea should be for a specific solution to a specific problem. As these difficulties were due to characteristics of specific tasks and students, we concluded that it was unlikely that further elaboration would solve this issue and chose instead to introduce Guideline 2a: "If there are several potential stuck diagnoses, choose the earliest diagnosis [in the sequence 1A–1B–2A–2B–3A–3B–4B]." This guideline was justified by the long-term intentions of the framework: to help teachers support students' own construction of mathematical reasoning. A too early diagnosis would, at worst, be useless and could then be followed by a later diagnosis, whilst a too late diagnosis could be detrimental by leading the teacher to help the student with too much of the solution.

Since guidelines were now in place for determining the difficulty type and amongst stuck diagnoses, we also introduced a guideline for needing-evaluation diagnoses in line with the long-term intentions of the framework. Since the phase-specific diagnoses (1–4D, Fig. 4) were defined as needing help evaluating a specific part of the solution, support designed for these diagnoses would be more specific than for the general diagnosis (C). Also, as described above, starting evaluation of the whole solution could result in a more specific diagnosis. Therefore, Guideline 2b was formulated: "If there are several potential needing-evaluation diagnoses, choose diagnosis C." The last iteration tested this guideline and revealed no need for revisions.

5.4 Operationalisability

A potential for aiding validation by linking answers to the diagnostic questions to phases was identified already in Iteration 2, but the analysis did not provide sufficient basis for revisions, except for the link between Phase 1 and DQ1, which asks for the student's interpretation of the task. This link was visualised in the overview of the framework (Fig. 4). In iteration 3, such links were also made explicit within the definitions of diagnoses, for example, by adding the formulation "The student's answer to DQ1 shows that the student understood the explicit information in the task and/or the student has written down the important information in the task" to initiate exploration (2A).

The analysis of both typical and borderline themes revealed that evidence of what the student had done so far (answering DQ2) could aid in determining how far the student had come within phases, especially Phases 2 and 4, whilst evidence about why the student had done what they had done (answering DQ3) could aid in determining whether the student had created a solution idea or not, that is, whether they had passed 3B. These results led to revisions of all stuck diagnoses in Phases 2–4, for example, adding the formulation "The student's answer to DQ2 shows that they have tried to explore but have not found a suitable strategy" to implement exploration (2B).

The typical cases were used to formulate indicators for the diagnoses. For example, for diagnosis 3B, the formulation that the student should have "conducted a good exploration" was elaborated as "The student has formulated relationships or properties, tried examples, and/or created representations that are sufficient to create a solution idea." Each diagnosis also included examples related to four different problems. These examples were revised in all iterations based on the typical examples, successively making the examples more authentic and more in line with the definitions.

During teacher-researcher meetings, we also noted that the sub-questions and guidelines increased operationalisability. The sub-questions structured the diagnostic process in two simpler steps: first determining whether the student is stuck or needing evaluation (Sub-question 1, Fig. 4) and then where the student is stuck (Sub-question 2a, Fig. 4) or what they need to evaluate (Sub-question 2b, Fig. 4). Sub-question 1 was usually answered quickly, whilst Sub-questions 2a and 2b could require lengthy discussions and thus seemed to constitute the main obstacle for using the framework in the classroom. Guidelines 2a and 2b provided a way of settling such discussions and the teachers reported that they also helped in-the-moment diagnosis in the classroom.

6 Discussion

In this study, we have developed a framework for characterising students' specific reasoning difficulties by devising iterations of design experiments that successively increased the empirical grounding and theoretical refinement of the framework. We will first reflect on the extent to which the final framework answers our three research questions and thus fulfils the three functionality criteria: range, differentiation, and operationalisability, and then discuss the contribution and limitations of our study as well as suggestions for future research.

Regarding range, the diagnoses of the final framework capture the 285 specific reasoning difficulties included in our data. The revisions of the definitions and structure of the framework made it possible to incorporate difficulties that were not covered by the initial framework, such as a student not being able to explain how an answer was obtained. The framework also includes difficulties regarding interpretation of the problem, a phase in the problem-solving process previously often omitted (Rott et al., 2021). This phase was found to be non-trivial for students in our study, just as in studies in error analysis regarding word problems (Clements, 1980; Whang, 1996). This finding indicates that models developed from studies of adult students and professional mathematicians may not be directly generalisable to students in Grades 1–12. Within these grades, however, we did not find any differences that warranted a need for separate frameworks for different grade levels. Whilst not specifically studied, we have indications that the difficulties emerging depend more on the type of problem than the grade level. For example, some problems require little time for interpretation but extensive time for exploration, whilst some problems take effort to interpret but are then easily solved. The phases may thus require different amounts of effort for different problems, which in turn may increase or decrease the likelihood of certain diagnoses to occur for different problems. Surprisingly, we did not see large differences in types and complexity of problems teachers used for different grades, which may also explain the lack of large differences in students' difficulties. Future research studying other types of tasks may thus reveal difficulties that are not captured by the framework. We do nonetheless argue that the diversity of our data—spanning 38 teachers' classrooms, 152 lessons, and 110 mathematical problems—constitutes a sound empirical base for answering Research Question 1: "What are the main specific reasoning difficulties students encounter when solving problems, and how can they be incorporated in the framework?"

The framework presents one answer to Research Question 2: "How can difficulties that are theoretically and observably different, in the sense that students need help with different aspects of their line of reasoning, be differentiated?" Revisions based on assessment of the differentiation of the framework led to sharper definitions and borders between diagnoses, leading to decreasing numbers of borderline cases over iterations (Table 1). The introduction of

the three guidelines (Fig. 4) provided support for deciding on a diagnosis in uncertain cases. We do, however, believe that the information provided in Fig. 4 is not sufficient for differentiating between diagnoses but must be accompanied by an understanding of the underlying model for students' problem-solving processes described in the full framework, comprising 25 pages. In particular, the character of students' needing-evaluation difficulties implied that evaluation should neither be seen as a final phase, as in Schoenfeld's (1985) model, nor as sub-phases of each main phase, as in our initial framework. Instead, when the need for evaluation emerges, it may concern the relationship of any two previous or current phases, as well as the solution as a whole. Evaluation may thus not be easily fitted into an otherwise sequential phase model, but rather be seen as a different mode in the problem-solving process. When seen in this way, teachers were more likely to identify needing-evaluation difficulties, even though students seldom engaged in evaluation spontaneously, as seen in previous studies (Koichu et al., 2021).

Research Question 3, "How can the structure and definitions of the framework support the diagnostic process by relating it to indicators of diagnoses that are observable in classroom situations?", was answered in two key ways. First, the three sub-questions with auxiliary guidelines (Fig. 4) supported the diagnostic process by structuring it in two simpler steps. This increased operationalisability as the teachers were able to decide on a diagnosis quicker and with more certainty. Second, the indicators of diagnoses in the final framework were expressed in terms of students' answers to the diagnostic questions. Such indicators stressed the importance of asking diagnostic questions and supported diagnosing based on students' description of their reasoning rather than inferences from students' written work. This may increase the accuracy of diagnoses, as it is long known that students' written communication is often insufficient for characterising students' difficulties (Cai et al., 1996; Clements, 1980).

6.1 Contribution

Teachers and researchers have several purposes for analysing and characterising students' mathematical work and have developed different tools for supporting such analyses. Patterns in students' errors were an early focus of analysis, in particular, for revealing students' understandings of specific content (Radatz, 1980) and causes for incorrect solutions to arithmetic word problems (Clements, 1980). There are also several rubrics for assessing levels of skills and competences expressed in students' work (e.g., Cai et al., 1996). Information obtained from using these tools can guide teachers' didactical choices in the longer perspective—for an upcoming period or the next time they give a course. Our framework contributes another perspective on students' difficulties by focussing what hinders students' own construction of reasoning in-the-moment—both in terms of errors and being stuck—regardless of mathematical content.

In doing so, the framework contributes to specifying the elicitation and interpretation components of formative assessment for mathematical problem-solving processes. It is well-known that providing opportunities for problem solving and reasoning is key for promoting students' mathematics learning (Hiebert & Grouws, 2007; Lester & Cai, 2016), but also that maintaining such opportunities in student–teacher interaction is challenging for teachers (Black & Wiliam, 2009; Liljedahl & Cai, 2021; Stein et al., 1996). The diagnostic questions can support teachers' elicitation of information about students' reasoning difficulties, and the diagnoses and guidelines of the framework can help teachers interpret this information in terms of precise needs that students have when solving problems. Identifying students' precise needs is a prerequisite for generating and selecting feedback that supports students' own construction of mathematical reasoning and thereby their learning (Lee & Cross Francis, 2018; Teuscher et al., 2016).

This study also contributes a description of a systematic method for how design-based theorising can aid the generation, selection, and validation of design alternatives that are consequential for students' learning (diSessa & Cobb, 2004). We have explicated how two key characteristics of design research (McKenney & Reeves, 2018)—multiple iterations and teacher–researcher collaboration—contributed to the increased functionality of the framework. Multiple iterations allowed us to test and revise the framework until theoretical saturation was reached, and collaboration with teachers granted access to classroom data on a wide range of student difficulties—in terms of both age of students and types of problems—and allowed us to assess and increase the framework's functionality for diagnosing students' specific reasoning difficulties for both teachers and researchers.

6.2 Limitations and future research

Our framework is focussed on, and thus limited to, difficulties that arise in the moment when students try to construct their own solutions to mathematical problem. This means that other important aspects of student difficulties are disregarded. For example, the framework does not identify more general mathematical difficulties that develop over time, that can be captured in competence models (Niss & Højgaard, 2019). Our study neither investigates difficulties that are specific for certain types of problems, such as proving tasks, word problems or modelling (Stylianides et al., 2017; Verschaffel et al., 2020; Whang, 1996), or certain aspects of problem solving, such as metacognition and heuristic strategies (Mevarech et al., 2018; Schoenfeld, 1985). In addition, the framework focusses on difficulties with managing mathematics itself and ignores the many types of cognitive, linguistic, affective, and social difficulties that students may have. However, including such aspects would increase the complexity of the framework and potentially decrease its functionality.

Our framework builds on a sequential process model developed from Schoenfeld's (1985) six phases and Lithner's (2008) conceptualisation of reasoning. A different starting point, for example, Jeannotte and Kieran's (2017) non-sequential characterisation of processes involved in reasoning, would have affected the initial framework. However, our analyses of empirical data on students' difficulties and teachers' diagnosis processes had a substantial impact on the final framework. It is a question for future research to determine whether a different initial framework would converge to a similar final framework if subjected to similar iterations of systematic testing and revision.

Whilst this study was based on a broad sample of student difficulties and no major differences over grades 1–12 were identified, our analyses do not reveal how diagnoses are distributed for different types of problems, different student groups, or different cultural contexts. Neither does the framework consider different levels of difficulties. Further analyses could reveal systematic differences in the relations between, for example, student age and the degree of difficulty, which could lead to more specific and less complex frameworks for specific contexts.

Furthermore, our study does not examine what challenges might arise for teachers when beginning to use this framework. As previous studies show, describing students' problem-solving processes require complex models with multiple phases (Clements, 1980; Rott et al., 2021). We developed the framework over 1.5 years. Whilst we do not expect learning to use the framework to take as much time as developing it, it is still likely that it requires substantial professional development efforts. One possibility for organising such efforts is to start using parts of the framework and adding elements over time. A first step can be to implement the diagnostic questions (DQ1–3, Fig. 4). The majority of our teachers did this and found that it was both manageable and had positive impact in the classroom. A second step can be to focus

only on the two main difficulty types. We have successfully tried this in one-hour workshops with teachers previously unfamiliar with the framework. Future research is needed to study these suggestions systematically and to investigate the effects of the framework on teachers' attention to and interpretation of students' reasoning difficulties, the feedback they provide on the basis of such interpretations and its effect on students' reasoning.

Acknowledgements This study is part of a project supported by grants from the Swedish Research Council (2017-03663) and the Swedish Institute for Educational Research (2019-00038).

Author contribution The methodology of this study was developed by A.I.S., B.P., and J.L. All authors contributed to steps 1–2 of the analysis. Subsequent steps of analysis were conducted by A.I.S., B.P., J.L., and J.S. The first draft of the manuscript was written by A.I.S., T.P., B.P., and J.L., and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by Umea University.

Declarations

Ethics approval Informed consent for participation in this study was gathered in accordance with applicable laws and guidelines. The Swedish Ethical Review Agency has determined that the research conducted within this project do not need Ethical Review according to Swedish law.

Competing interests The authors declare no competing interests.

References

Bell, C., Steinberg, J., Wiliam, D., & Wylie, C. (2008). *Formative assessment and student achievement: Two years of implementation of the Keeping Learning on Track program*. https://www.dylanwiliam.org/Dylan_Wiliams_website/Papers.html

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25. https://doi.org/10.1080/0969594x.2010.513678

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Boaler, J. (2002). *Experiencing school mathematics: Traditional and reform approaches to teaching and their impact on student thinking*. Lawrence Erlbaum. https://doi.org/10.4324/9781410606365

Boesen, J., Lithner, J., & Palm, T. (2010). The relation between types of assessment tasks and the mathematical reasoning students use. *Educational Studies in Mathematics*, *75*(1), 89–105. https://doi.org/10.1007/s10649-010-9242-9

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, *30*(1), 3–12. https://doi.org/10.1111/j.1745-3992.2010.00195.x

Brousseau, G. (1997). *Theory of didactical situations in mathematics*. Kluwer. https://doi.org/10.1007/0-306-47211-2

Cai, J., Jakabcsin, M. S., & Lane, S. (1996). Assessing students' mathematical communication. *School Science and Mathematics*, *96*(5), 238–246. https://doi.org/10.1111/j.1949-8594.1996.tb10235.x

Clements, M. A. (1980). Analyzing children's errors on written mathematical tasks. *Educational Studies in Mathematics*, *11*(1), 1–21. https://doi.org/10.1007/BF00369157

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, *32*(1), 9–13. https://doi.org/10.3102/0013189X032001009

Datnow, A., & Hubbard, L. (2016). Teacher capacity for and beliefs about data-driven decision making: A literature review of international research. *Journal of Educational Change*, *17*(1), 7–28. https://doi.org/10.1007/s10833-015-9264-2

Dindyal, J., Schack, E. O., Choy, B. H., & Sherin, M. G. (2021). Exploring the terrains of mathematics teacher noticing. *ZDM-Mathematics Education, 53*(1), 1–16. https://doi.org/10.1007/s11858-021-01249-y

diSessa, A. A., & Cobb, P. (2004). Ontological innovation and the role of theory in design experiments. *Journal of the Learning Sciences*, *13*(1), 77–103. https://doi.org/10.1207/s15327809jls1301_4

Eisenhardt, K. (1989). Building theories from case study research. *Academy of Management Review*, *14*(4), 532–550. https://doi.org/10.2307/258557

Gu, P. Y. (2021). An argument-based framework for validating formative assessment in the classroom. *Frontiers in Education*, *6*, 605999. https://doi.org/10.3389/feduc.2021.605999

Gummer, E., & Mandinach, E. (2015). Building a conceptual framework for data literacy. *Teachers College Record: The Voice of Scholarship in Education*, *117*(4), 1–22. https://doi.org/10.1177/016146811511700401

Hiebert, J., & Grouws, D. (2007). The effects of classroom mathematics teaching on students' learning. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1293–1312). Information Age Publishing.

Jeannotte, D., & Kieran, C. (2017). A conceptual model of mathematical reasoning for school mathematics. *Educational Studies in Mathematics*, *96*(1), 1–16. https://doi.org/10.1007/s10649-017-9761-8

Jonsson, B., Norqvist, M., Liljekvist, Y., & Lithner, J. (2014). Learning mathematics through algorithmic and creative reasoning. *Journal of Mathematical Behavior*, *36*, 20–32. https://doi.org/10.1016/j.jmathb.2014.08.003

Koichu, B., Parasha, R., & Tabach, M. (2021). Who-is-right tasks as a means for supporting collective looking-back practices. *ZDM-Mathematics Education*, *53*(4), 831–846. https://doi.org/10.1007/s11858-021-01264-z

Lampert, M. (1990). When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching. *American Educational Research Journal*, *27*(1), 29–63. https://doi.org/10.3102/00028312027001029

Lee, M. Y., & Cross Francis, D. (2018). Investigating the relationships among elementary teachers' perceptions of the use of students' thinking, their professional noticing skills, and their teaching practices. *Journal of Mathematical Behavior*, *51*, 118–128. https://doi.org/10.1016/j.jmathb.2017.11.007

Lester Jr., F. K., & Cai, J. (2016). Can mathematical problem solving be taught? Preliminary answers from 30 years of research. In P. Felmer et al. (Eds.), *Posing and solving mathematical problems* (pp. 117–135). Springer. https://doi.org/10.1007/978-3-319-28023-3_8

Liljedahl, P., & Cai, J. (2021). Empirical research on problem solving and problem posing: A look at the state of the art. *ZDM-Mathematics Education, 53*(4), 723–735. https://doi.org/10.1007/s11858-021-01291-w

Lithner, J. (2008). A research framework for creative and imitative reasoning. *Educational Studies in Mathematics*, *67*(3), 255–276. https://doi.org/10.1007/s10649-007-9104-2

Lithner, J. (2017). Principles for designing mathematical tasks that enhance imitative and creative reasoning. *ZDM-Mathematics Education, 49*(6), 937–949. https://doi.org/10.1007/s11858-017-0867-3

Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education*, *60*, 366–376. https://doi.org/10.1016/j.tate.2016.07.011

McKenney, S., & Reeves, T. C. (2018). *Conducting educational design research* (2nd ed.). https://doi.org/10.4324/9781315105642

Means, B., Chen, E., DeBarger, A., & Padilla, C. (2011). *Teachers' ability to use data to inform instruction: Challenges and supports*. US Department of Education, Office of Planning, Evaluation, and Policy Development. https://files.eric.ed.gov/fulltext/ED516494.pdf

Mevarech, Z., Verschaffel, L., & De Corte, E. (2018). Metacognitive pedagogies in mathematics classrooms: From kindergarten to college and beyond. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 109–123). Routledge.

Niss, M., & Højgaard, T. (2019). Mathematical competencies revisited. *Educational Studies in Mathematics*, *102*(1), 9–28. https://doi.org/10.1007/s10649-019-09903-9

Pólya, G. (1945). *How to solve it*. Princeton University Press.

Radatz, H. (1980). Students' errors in the mathematical learning process: a survey. *For the Learning of Mathematics*, *1*(1), 16–20. https://www.jstor.org/stable/40247696

Randel, B., Apthorp, H., Beesley, A., Clark, T., & Wang, X. (2016). Impacts of professional development in classroom assessment on teacher and student outcomes. *The Journal of Educational Research*, *109*(5), 491–502. https://doi.org/10.1080/00220671.2014.992581

Rott, B., Specht, B., & Knipping, C. (2021). A descriptive phase model of problem-solving processes. *ZDM-Mathematics Education, 53*(4), 737–752. https://doi.org/10.1007/s11858-021-01244-3

Schneider, M. C., & Gowan, P. (2013). Investigating teachers' skills in interpreting evidence of student learning. *Applied Measurement in Education*, *26*(3), 191–204. https://doi.org/10.1080/08957347.2013.793185

Schneider, M. C., & Randel, B. (2010). Research on characteristics of effective professional development programs for enhancing educators' skills in formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 251–276). Routledge.

Schoenfeld, A. (1985). *Mathematical problem solving*. Academic Press. https://doi.org/10.1016/C2013-0-05012-8

Shaughnessy, M., Boerst, T. A., & Farmer, S. O. (2019). Complementary assessments of prospective teachers' skill with eliciting student thinking. *Journal of Mathematics Teacher Education*, *22*, 607–638. https://doi.org/10.1007/s10857-018-9402-x

Sidenvall, J., Granberg, C., Lithner, J., & Palmberg, B. (2022). Supporting teachers in supporting students' mathematical problem solving. *International Journal of Mathematical Education in Science and Technology*. https://doi.org/10.1080/0020739X.2022.2151067

Sidenvall, J., Lithner, J., & Jäder, J. (2015). Students' reasoning in mathematics textbook task-solving. *International Journal of Mathematical Education in Science and Technology*, *46*(4), 533–552. https://doi.org/10.1080/0020739X.2014.992986

Simon, M. A. (2019). Analyzing qualitative data in mathematics education. In K. R. Leatham (Ed.), *Designing, conducting, and publishing quality research in mathematics education* (pp. 111–122). Springer. https://doi.org/10.1007/978-3-030-23505-5_8

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal*, *33*(2), 455–488. https://doi.org/10.3102/00028312033002455

Stein, M. K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation*, *2*(1), 50–80. https://doi.org/10.1080/1380361960020103

Stein, M. K., & Smith, M. S. (1998). Mathematical tasks as a framework for reflection: From research to practice. *Mathematics Teaching in the Middle School*, *3*(4), 268–275. https://doi.org/10.5951/MTMS.3.4.0268

Stylianides, G. J., Stylianides, A. J., & Weber, K. (2017). Research on the teaching and learning of proof: Taking stock and moving forward. In J. Cai (Ed.), *Compendium for research in mathematics education* (pp. 237–266). National Council of Teachers of Mathematics.

Teuscher, D., Moore, K. C., & Carlson, M. P. (2016). Decentering: A construct to analyze and explain teacher actions as they relate to student thinking. *Journal of Mathematics Teacher Education*, *19*(5), 433–456. https://doi.org/10.1007/s10857-015-9304-0

Verschaffel, L., Schukajlow, S., Star, J., & Van Dooren, W. (2020). Word problems in mathematics education: A survey. *ZDM-Mathematics Education, 52*(1), 1–16. https://doi.org/10.1007/s11858-020-01130-4

Vinner, S. (1997). The pseudo-conceptual and the pseudo-analytical thought processes in mathematics learning. *Educational Studies in Mathematics*, *34*, 97–129. https://doi.org/10.1023/A:1002998529016

Whang, W.-H. (1996). The influence of English-Korean Bilingualism in solving mathematics word problems. *Educational Studies in Mathematics*, *30*(3), 289–312. https://doi.org/10.1007/BF00304569

Wijaya, A., van den Heuvel-Panhuizen, M., Doorman, M., & Robitzsch, A. (2014). Difficulties in solving context-based PISA mathematics tasks: An analysis of students' errors. *The Mathematics Enthusiast*, *11*(3), 8. https://doi.org/10.54870/1551-3440.1317

Yimer, A., & Ellerton, N. F. (2010). A five-phase model for mathematical problem solving: Identifying synergies in pre-service-teachers' metacognitive and cognitive actions. *ZDM-Mathematics Education*, *42*, 245–261. https://doi.org/10.1007/s11858-009-0223-3