Analyzing problem solving in secondary physics education: A rubric-guided approach to explore individual learning needs

André Meyer,* Stanley Fischer, and Gunnar Friege†
Institute for Mathematics and Physics Education—Physics Education Group, Leibniz University Hannover, Hannover, Germany

(Received 13 July 2025; accepted 14 November 2025; published 28 January 2026)

This study examines how the Minnesota Assessment of Problem-Solving rubric can be used to diagnose learning needs in secondary physics education. In total, 51 students (grades 9 and 11) from 5 German secondary schools solved a physics problem-solving task and completed an energy concept test. Rubric-guided analysis of their written problem solutions revealed seven characteristic performance patterns, ranging from expertlike reasoning to incomplete or failed approaches. While students generally managed descriptions and basic calculations, many struggled to apply the conservation principle and to structure their reasoning coherently. The combination of rubric scores with conceptual test results highlighted whether difficulties stemmed from missing knowledge or from challenges in applying it. These findings demonstrate that this combination of assessment tools, adapted to the secondary level, offers a practical approach for formative assessment and targeted support in physics problem solving. However, difficulties arise when students submit incomplete problem solutions. Future developments could focus on atomized approaches or the integration of AI technologies for automatic assessment.

DOI: 10.1103/3bs7-fnrd

I. INTRODUCTION

Secondary school students often struggle to apply their conceptual understanding of physics to unfamiliar contexts and problem-solving situations [1,2]. In addition to these conceptual difficulties, a lack of mathematical fluency frequently hinders students' ability to solve physics problems effectively [3,4]. One particularly demanding aspect is the mathematization of physical phenomena [5], that is, translating real-world situations into mathematical models. These challenges may contribute to the decline in students' interest and motivation toward science subjects, that is observable in secondary education [6–8].

In German secondary schools, students begin engaging with quantitative reasoning in physics, especially in the domains of mechanics, thermodynamics, and electricity. At this stage, energy is typically the first concept introduced with explicit quantitative considerations. This transition marks a critical phase in students' physics education, where foundational conceptual knowledge must be integrated with emerging mathematical competencies. To support students during this pivotal stage, it is essential for instructors to better understand the specific challenges they face in solving quantitative physics problems. The present study aims to investigate the assessment of students' use of conceptual knowledge and their ability to model and solve physics problems quantitatively, with a focus on the principle of energy conservation. The study seeks to identify individual learning needs and difficulties using a rubric-guided approach. In this context, rubrics could be a practical and time-efficient tool for teachers, offering insights that may inform instructional strategies to strengthen students' problem-solving skills early in their physics learning trajectory [9,10].

II. THEORETICAL BACKGROUND

A. Problem solving

Problem solving is considered to be one of the most important skills that students need to achieve in science, technology, engineering, and mathematics (STEM) education [11]. Therefore, it has been subject to PER for many decades [12–14]. The definition of what constitutes a problem varies slightly across the literature, but a common distinction is made between problems and routine exercises. A problem is typically characterized as a task with a defined goal, but without an immediately apparent path to its solution [13,15]. In contrast, a routine exercise is a task for which the solver knows from the outset how to proceed. Importantly, this distinction does not imply that problems are inherently difficult or that routine tasks are necessarily easy [16]. Rather, the key difference lies in the cognitive

*Contact author: a.meyer@idmp.uni-hannover.de
†Contact author: friege@idmp.uni-hannover.de

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

demands: solving a problem requires active decision making about which concepts and principles to apply [17–19]. Whether a given task is perceived as a problem or a routine exercise depends on the individual's prior knowledge and experience. Thus, the classification is not inherent to the task itself but relative to the solver; what constitutes a problem for one learner may be a routine exercise for another [13].

In this study, we use well-defined problems similar to those in physics textbooks [20]. Well-defined problems are characterized by a clearly described initial state and goal state, whereas ill-defined or complex problems have no clearly defined goal [21].

Problem-solving processes are described in various models [22–24]. Despite some differences, all of these models have four phases in common. At first, there is a representation phase, where the solver is supposed to understand the problem situation, collect information about given and missing quantities, and, for example, draw a picture of the situation. This representation phase is often done implicitly by expert problem solvers [23]. In the second phase, the solver devises a plan for the problem solution. This phase might be characterized by trying out different approaches to solving the problem. In the third phase, this plan is then executed. In the execution phase, the problem solver needs to perform mathematical and physical operations toward the goal that are predefined after the planning phase [22]. The formulation of a solution is also included in the execution phase. After the solution is formulated, there is a phase of evaluation where the solver "looks back" on the process and reflects on the solution. In school contexts, the evaluation phase is often left out because of the well-defined nature of the problems [17,23].

For solving problems, the students need to identify and apply conceptual knowledge and knowledge about problem schemes [25–27]. The application of conceptual knowledge about physics to physics problems poses a significant challenge for students [16,26,28]. For that reason, it can be a good predictor for success in solving domain-specific problems [27]. For domain-general problems, students with lower conceptual knowledge can also succeed in problem solving if they are able to comprehend new contexts quickly [24].

Understanding problem schemes can highly influence the problem-solving success [25,27]. Students who know problem schemes can adapt solution processes to new problem situations and get to a correct solution quickly [23]. Expertlike problem solvers recognize the conceptual structure of problems and are likely to choose correct problem schemes. Novice problem solvers have limited scheme knowledge and must rely solely on a few examples and their conceptual knowledge, which is more error-prone. These students also tend to be distracted by superficial aspects of the problem situation and might therefore choose inappropriate schemes [29]. Extensive training of certain problem types can help to improve the problem-scheme knowledge, but it can also lead to an unreflected use of problem schemes, preventing students from improving their conceptual knowledge through problem solving [30,31]. This also has the effect of preventing the transfer from well-defined textbook problems to ill-defined real-world problems, meaning that students may become good problem solvers at school but are unable to transfer these skills to the real world. This is because real-world problem solving requires different skills than textbook problem solving [21].

Nonetheless, well-defined problems play an important role, especially for beginning to achieve problem-solving skills. They can help to bridge the gap between routine exercises and complex problem solving by showing the students how to apply conceptual knowledge to problem situations. Also, these simple problems provide students with problem scheme knowledge. This progression from routine exercises over well-defined problems to complex problems is called progressive problem solving [32].

B. Problem-solving assessment

Like problem solving in general, its assessment specifically has been researched for decades. This extensive research has produced different approaches for assessing problem-solving skills. For example, there are assessment approaches focusing on metacognition [33], knowledge [30], or problem-solving processes [34]. Various constructs are being measured for problem-solving assessment like the time used for task completion [35], or the percentage of correct solutions [36]. A well-known largescale assessment of problem-solving is conducted by the OECD's Programme for International Student Assessment (PISA). PISA assesses 15-year-olds' skills in mathematics, reading, and science across more than 80 countries around the world. In the PISA assessments, students have to answer items including multiple-choice, drag-and-drop, and short-text responses using a computer [37]. They use a model of four subprocesses of problem-solving similar to those explained in Sec. II A [24]. This model was also used for qualitative analysis of students' written problem solutions [38].

Qualitative assessment of problem-solving processes is frequently supplemented by interviews or think-aloud protocols. In interview studies, the participants solve problems and are interviewed afterward regarding their thoughts and the decisions they made [39]. This allows deeper insights into the cognitive processes of problem solvers in comparison to only analyzing the written solutions. One disadvantage of interview studies is that they can only examine the processes retrospectively based on what the participant remembers.

In think-aloud studies, the participants' thoughts are covered as well, but for this approach they are prompted to verbalize their thoughts during their solution process [39,40].

This allows mental processes to be recorded directly and with less distortion than through recall. However, participants are potentially exposed to a high cognitive load because they need to verbalize their thoughts simultaneously. This verbalization is different from usual social interaction, and students might need "practice problems before the actual experiment" [39] (p. xiv) in order to meet the correct mode of communicating their thoughts. Nonetheless, the insights of think-aloud studies are the "gold standard" of problem-solving assessment [17].

If no additional data such as interviews or think-aloud protocols are available, the qualitative analysis can solely focus on the written problem solution. In studies like this, rubrics can be used to quantify qualitative analysis of written solutions using multiple dimensions of performance. These scores can be used to compare different solutions more easily, because they provide clear criteria for the assessment. This also increases validity and objectivity [41]. They are more detailed than a single-dimensional scoring of problem solutions as correct or incorrect. However, like always, when summarizing qualitative analyses into quantitative numbers, some information is lost. For example, rubric scores do not contain any information about the sequential structure of a problem solution [42]. Like other methods relying on written problem solutions, rubric-guided analyses reach their limits for incomplete solutions that do not allow detailed insights. And it can be hard to rate problem solutions using innovative solution paths that are fundamentally different from the expected solution path.

Which method to choose relies on the objectives of the assessment. Studies focusing on cognitive processes in detail use additional data in order to get the most detailed information possible [27]. In PISA, for example, the objective is to analyze a large population of students in regard to precisely defined subprocesses. For that reason, the assessment design needs to allow as much automatic evaluation as possible in order to reduce the effort that is necessary. Furthermore, this goal makes it advantageous to assess the subprocesses separately, because students might not execute certain steps if they are not explicitly prompted to do so. For example, expertlike problem solvers are known to implicitly create a representation of the problem situation [23]. In contrast, items that explicitly prompt the participants to execute subprocesses are not suitable for studies that aim to qualitatively analyze problem-solving processes, because step-by-step instructions strongly influence the process under examination [34]. For these objectives, a study design with open items, leaving room for the creativity of the participants, is preferable.

C. The Minnesota Assessment of Problem Solving

A scientifically well-known rubric for problem-solving assessment is the Minnesota Assessment of Problem Solving (MAPS) [17,34]. This rubric is relatively easy to use, which is important for applications involving multiple raters. The categories of the MAPS rubric focus on the subdimensions of *Useful Description*, *Physics Approach*, *Specific Application of Physics*, *Mathematical Procedures*, and *Logical Progression*. They are scored using an ordinal scale from 0 (worst) to 5 (best). Additionally, there are NA scores for categories that are not applicable to a problem or a solution [34]. A more detailed explanation of the rubric categories and their application to physics problem solutions can be found in Sec. IV C.

The MAPS rubric was originally designed and validated as an assessment tool for problem-solving skills for university-level students. The goal was to create a measure that is more detailed than a simple correct or incorrect distinction, but still easy to use in educational settings [34]. Since its publication in 2009, MAPS has been used as a scientific assessment tool in several studies involving both university and high school students. The majority of studies using the MAPS rubric were conducted at the university level across different countries [43–48]. Only a few studies used the MAPS rubric at the secondary level with high school students [31,49]. These studies use the rubric as a scientific measure, e.g., as a baseline to compare and evaluate newly developed measures for problem-solving skills [44]. Other use cases are analyses of correlations between different dimensions of problem-solving literacy [47,48] and studies that analyze the effects of various interventions for training problem-solving skills, [31,45,46,49]. Besides its scientific use, the developers of MAPS claim its usability for formative assessment by instructors. They provide a few brief examples of conclusions from the rubric scores for targeted feedback [17,34]. However, there is currently a lack of studies that address the application to secondary school contexts in detail, especially concerning the utility of MAPS for formative assessment.

III. RESEARCH QUESTIONS

Our project aims to enable individualized training of problem-solving skills for secondary students who are at the beginning stage of quantitative physics considerations. While intensive qualitative analyses of students' written problem solutions, especially supplemented by think-aloud data, would allow detailed insights into the individual cognitive processes of students, they are also overly complex and time consuming for everyday school teaching. A rubric-guided approach shows promise of being applicable, but it is unclear whether this approach allows for effectively analyzing students' needs. We therefore conducted a study aiming to analyze the MAPS rubric's potential as a tool for formative assessment in secondary school science teaching. Since domain-specific problem solving relies on conceptual knowledge as well, we combined the rubric-guided assessment with an assessment of the conceptual knowledge about energy.

The study is guided by the following research questions:

(1) What learning needs can be analyzed using the MAPS rubric scores on secondary physics problem solutions?
(2) What challenges emerge when applying the MAPS rubric to problem solutions from secondary school students with limited problem-solving experience?
(3) Can the combination of rubric scores and conceptual knowledge about energy improve these insights for problem solving about energy?

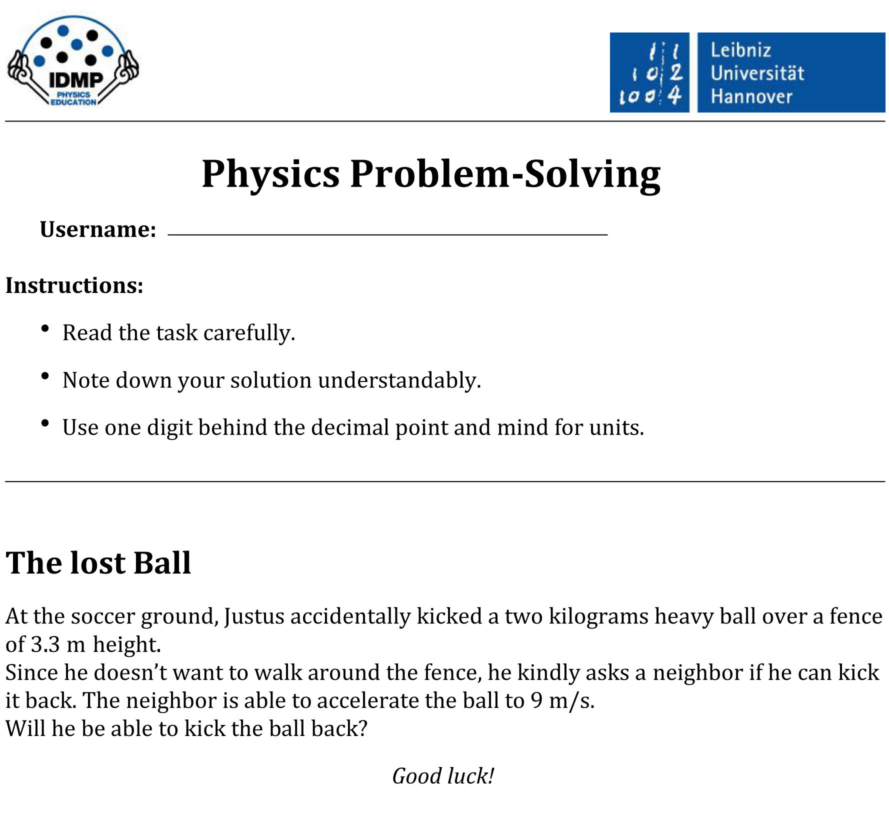
IV. MATERIALS AND METHODS

The study was conducted in physics classes, where students were taught the principle of conservation of energy. The participants were given a worksheet (see Fig. 1) with the problem-solving task and some instructions. They solved the task using pen and paper. Additionally, they were allowed to use a calculator and a formula sheet. There was no time limit set for the task. Every student participated voluntarily during their regular physics lesson and the data collection was anonymized using randomly assigned usernames, which the students wrote down on their solutions. In addition, every participant filled out a short personal questionnaire and a test of their conceptual knowledge about energy concepts using tablets.

A. Participants

The study was conducted at five different secondary schools in and around Hannover, Germany. Four schools were a “Gymnasium”—the highest form of German secondary schools. One school was a “Kooperative Gesamtschule (KGS),” which is a combination of the three German secondary school types. Students visiting a KGS can graduate after year 9, 10, or 13 with different degrees. Therefore, a KGS usually has more heterogeneous classes than a Gymnasium. Seven classes from these schools were part of the study (6x Gymnasium and 1x KGS). Schooling is under federal jurisdiction in Germany, and we had only permission to include schools from the state of Lower Saxony (“Niedersachsen”). For practical reasons, we focused on schools near Hannover.

A total of 51 students participated, of which 23 were males, and 28 were females. The mean age of all participants was 15.6 (SD 1.1) years. Of these, 30 students are from two ninth-grade classes with a mean age of 14.8 (SD 0.6) years, and 21 students are from five 11th-grade classes with a mean age of 16.6 (SD 0.5) years.


IDMP PHYSICS EDUCATION

Leibniz Universität Hannover

Physics Problem-Solving

Username: ____________________

Instructions:

- Read the task carefully.
- Note down your solution understandably.
- Use one digit behind the decimal point and mind for units.

The lost Ball

At the soccer ground, Justus accidentally kicked a two kilograms heavy ball over a fence of 3.3 m height.
Since he doesn’t want to walk around the fence, he kindly asks a neighbor if he can kick it back. The neighbor is able to accelerate the ball to 9 m/s.
Will he be able to kick the ball back?

Good luck!

FIG. 1. Problem-solving task and instructions that were given to the participants. The original worksheet was in German. For this figure, it was translated into English by the authors.

B. Problem-solving task

Seven different problem tasks were developed for this study. They are based on the applicable curriculum and tasks from commonly used physics textbooks. Following the curriculum, students are introduced to qualitative considerations about energy in year 7 or 8. Building upon this knowledge, they focus on quantitative energy considerations between years 9 and 11. Within the unit on quantitative energy considerations, the students learn to use the formulas of thermal energy, gravitational potential energy, and kinetic energy. Therefore, the designed problem tasks only rely on these three formulas. For example, in one task, the students were asked to calculate the velocity of a plastic projectile shot by an airsoft pistol. In another task, they were asked to calculate the energy of a hot cup of tea.

In order to get scoring results enabling comparisons between students, we decided to use the same problem task for every student. The tasks were analyzed regarding various criteria for complexity of a problem [50], including the number of potential distractions, the length and readability of the problem situation, the level of required algebraic transformations, and the logical reasoning steps necessary for a solution.

In the selected problem task (see Fig. 1), a juvenile kicked his ball over a fence and does not want to walk around the fence to get his ball back. Therefore, he asks a neighbor to kick it back. The study participants are supposed to determine whether the neighbor can kick the ball over the fence. The mass of the football, the height of the fence, and the maximum velocity that the neighbor can accelerate the ball to are given within the text describing the problem situation. This task was considered the most suitable, because it requires active decision making about which approach to use and interpretation of results at the end. This makes it more "problem-alike" than classical textbook tasks directly asking students to calculate a certain quantity in a given situation, as in the above-mentioned tasks with the airsoft pistol and the cup of tea. Especially, the task with the cup of tea can be considered a routine exercise since it is clear from the beginning that it can be solved using the formula for thermal energy. The airsoft task uses only the kinetic energy; therefore, students do not need the principle of energy conservation for its solution. With two forms of energy being relevant, the football problem is more complex and necessitates the concept of energy conservation. The problem situation is adapted to the everyday experiences of young people and is intended to represent a relatable real-life scenario. Therefore, the situation is not too complex.

In order to show that the football task is indeed a problem task and not a routine exercise for our participants, we discuss students' solutions in Sec. VI A.

The anticipated approach for solving this problem is to use the principle of energy conservation. Following this principle, the amount of kinetic energy of the ball after it was kicked needs to be at least as high as the gravitational potential energy of the ball at the top of the fence. This approach can be represented by Eq. (1):

$$m \cdot g \cdot h = \frac{1}{2} \cdot m \cdot v^2. \tag{1}$$

This equation, together with the information from the task text, can be used to either calculate the minimum velocity of the ball required for it to reach the top of the fence,

$$v = \sqrt{2 \cdot g \cdot h}, \tag{2}$$

or to calculate the maximum height that the ball can reach with the given velocity

$$h = \frac{v^2}{2 \cdot g}. \tag{3}$$

After using these formulas, the students can argue that the neighbor can kick the ball faster than the minimum needed speed or higher than the fence. Therefore, the neighbor can kick the ball back.

A mathematically easier way to get to this conclusion would be to separately calculate the two forms of energy and compare their values. Since the amount of kinetic energy is larger than the amount of potential energy needed, the neighbor can kick the ball with enough energy to surpass the fence. The approach for this strategy stays the same, because the students have to argue using the principle of energy conservation. However, the algebraic operations are easier, especially using a calculator. Another way to get to Eq. (3) is using the dynamics of a vertical throw. A function for the height of the ball at a given time is defined by

$$h(t) = v_0 \cdot t - \frac{1}{2} \cdot g \cdot t^2. \tag{4}$$

This function reaches its maximum for $t_{\max} = \frac{v_0}{g}$ with v_0 being the start velocity. The maximum height of the ball at that time is:

$$h(t_{\max}) = v_0 \cdot \frac{v_0}{g} \cdot \frac{1}{2} \cdot g \cdot \left(\frac{v_0}{g}\right)^2 = \frac{v_0^2}{2 \cdot g}. \tag{5}$$

This would be an approach without explicitly using the conservation of energy. Since the students were not necessarily taught on dynamics or Newton's laws before, they were not expected to use this approach.

C. Scoring criteria for applying the MAPS rubric

The written problem solutions were analyzed by two human coders using the MAPS rubric. At first, we randomly selected five problem solutions. They were rated separately and discussed afterward to align the criteria for the different scores with our problem.

The MAPS rubric has five categories: *Useful Description*, *Physics Approach*, *Specific Application of Physics*, *Mathematical Procedures*, and *Logical Progression*. The students' solutions are assigned a score between 0 and 5 or an NA (not applicable) for every category according to the scoring instructions by Docktor *et al.* [17,34]: A score of 5 indicates that the student shows complete and correct processing of the category. A score of 4 indicates a useful category with minor omissions or errors. Lower scores are assigned to solutions that are partly (3) or mostly (2) incorrect, incomplete, or inappropriate. If a solution is completely incorrect regarding a category, a score of 1 is assigned. If a solution does not contain evidence of one category, a score of 0 is assigned, with the exception that NA is given if the student did not have to explicitly perform the category and still continued to work correctly.

A student's description is considered useful, appropriate, and complete and therefore scored as 5 if the problem solution contains a drawing that correctly shows the problem situation and the given quantities. Alternatively, the description is also rated as 5 if there is a complete and correct list of the given quantities and the variables used for them.

In the *Physics Approach* category, the score reflects whether the student chose appropriate physics concepts and principles to solve the problem. In our case, the problem can be solved using the concept of energy and the law of conservation of energy as an appropriate principle, as explained in Sec. IV B. Since most of the students did not explicitly verbalize their approach, the human coders had to interpret what the student tried to do. A score of 5 was assigned to problem solutions that applied the law of conservation of energy to the potential energy of the ball at the top of the fence and the kinetic energy of the ball directly after being kicked. If the student calculated these two energies without comparing or equating them, a score of 3 was assigned, because the concept of energy is being used and the principle of energy transformation can be assumed. A score of 2 was assigned to problem solutions that, for example, calculated only one of the involved energies, since it correctly contains the energy concept but no principle of transformation or conservation. If neither of the two energies was calculated, a score of 1 or 0 was assigned.

In the *Specific Application of Physics* category, we rated the students' skills in following their approach. Deductions in scoring this category were made, for example, if the students used wrong formulas, used wrong values within formulas, or used incorrect units for values. It is to be noted that no deductions in this category were made if the problem solution did not use units or only used units for the results and not for the other values in a formula. Instead, this leads to deductions in the *Logical Progression* category, because we assume the student might know the correct units but did not write them down. This leads to a lack of transparency in the student's reasoning, since the steps leading to the correct unit in the final result are not evident from the written solution. If incorrect units were used, we interpret that the student made a physical mistake because they are supposed to know the correct units for physical quantities. Therefore, that is taken into account in the *Specific Application of Physics* category.

In the *Mathematical Procedures* category, the human coders analyzed the problem solution for any errors following the rules of mathematics. For example, deductions were made if students did not use brackets appropriately or if they made algebraic errors when transforming an equation or calculating a result.

In the *Logical Progression* category, we rated how focused the problem solution seems and how well it is communicated. Deductions were made if the problem solution is not organized logically or if we were not able to understand certain steps that the student took.

In addition to these five categories from the MAPS rubric, we coded every problem solution as either solved correctly or not solved (correctly). The problem solution was only considered correct and complete if the student answered the question whether the neighbor can kick the ball over the fence using their calculated results as reasoning. Problem solutions without a clear answer are coded as not solved because, in this problem, interpretation of the results is an important part of the solution. A problem solution is also coded as not solved if it contains an answer without any understandable reasoning, since we do not want to score guessing rather than problem solving.

D. Methods for analyses

After aligning the scoring criteria, all of the problem solutions were rated again by the two human coders. In addition to applying the MAPS rubric, the human coders rated whether the student's result shows a correct solution to the problem. Subsequently, every problem solution was discussed, and a consensus rating was defined. Median scores for the different categories are used for insights into the skills of the group of participants. The individual scores of every student were analyzed for interesting combinations of scores and similar scoring patterns across different participants.

E. Energy test

The conceptual knowledge about energy was assessed using a multiple choice energy test. This energy test was used in other studies of our institution before [51]. The items of this test are mainly based on the Energy Concept Assessment (ECA) by Neumann *et al.* [10]. The ECA is a validated assessment tool regarding the four

TABLE I. Number of students per score and category.

	Description	Approach	Physics	Math	Logic
0	4	10	12	12	9
1	0	8	2	2	5
2	2	10	5	3	15
3	3	5	8	2	8
4	11	0	9	16	11
5	24	14	15	16	3
Not applicable	7	4	0	0	0

conceptions: *Forms of Energy*, *Transformation*, *Degradation*, and *Conservation* that are crucial for learning the energy concept [52]. A learning progression for conceptual knowledge about energy in secondary physics education is known to be based on these conceptions. Students start to learn about forms and sources of energy in years 5 and 6 and move on with learning about energy transformation in years 7 and 8. From year 9 on, students start to achieve a deeper understanding of energy conservation and degradation [10]. Our participants are therefore currently learning these aspects of the energy concept and they do need the principle of energy conservation for solving the problem task under examination. For that reason, the energy test provides us with detailed insights into the individual state of the learning progression of energy for our participants.

For our study, we used 20 items ranging from relatively easy to relatively difficult items. We used five items for each of the four conceptions of energy. The item difficulties were obtained from Ref. [10] and from the above-mentioned project of our institution [51]. These difficulties are based on Rasch analyses of more than 2000 students answering different sets of the items. The students in the presented study took the test within a digital learning environment using tablets. Each item was rated as correct or incorrect, leading to a score between 0 and 20 for the conceptual knowledge in general and scores between 0 and 5 for the different conceptions of energy.

V. RESULTS

A. Scoring results

The human coders scored every student's problem solution using the described scoring criteria. Thereafter, the two coders compared their results and discussed every case of differing scores. Finally, a consensus score was defined, and the number of students per score and category was counted. The results can be seen in Table I. More detailed information, like the individual scoring of every student, can be found in Table III and Fig. 6 in Appendix A. The human raters achieved an interrater agreement of $\alpha > 0.8$ for every category. This is considered a satisfactory level of agreement, and it indicates reliable ratings [53]. Krippendorff's α was used, because the rubric scores are mainly ordinally scaled, except for the difference between 0 and 1 (see Sec. VI B).

In total, 14 students were able to solve the problem correctly. The highest scores were assigned to the *Useful description* category, where the students reached a median score of 5. The lowest scores were assigned to the *Physics Approach* and *Logical Progression* categories with a median score of 2. In the *Useful Description* category, 42 students reached high scores of 4, 5, or NA (solver), which indicates ceiling effects. Especially the *Mathematical Procedures*, and to a lesser extent, the *Specific Application of Physics* category show a strong separation of high-performing and low-performing students with only a few students achieving medium scores.

We compared the results of students from year 9 with the results of students from year 11. Since the MAPS scores are ordinally scaled and the results are not normally distributed, we used the Wilcoxon rank-sum test (WRS). Classes number 1 and number 7 (see Table III) are the two classes from year 9, the other classes are from year 11. The WRS revealed a significant difference for the *Useful Description* category ($p < 0.05$), but no significant difference for the other four categories ($p > 0.05$). Eight students from year 9 solved the task correctly, and six students from year 11 did so.

We divided the MAPS scores into high scores (4, 5, NA) and low scores (0, 1, 2, 3). Students with high scores show overall correct results regarding the category with at most minor deductions. Low scores indicate overall erroneous results of a category. We analyzed patterns of high- and low-scoring categories within the dataset. Problem solutions showing the same scoring pattern were compared to each other in order to find similarities. With this procedure, we were able to describe seven different groups of students with respect to their problem-solving skills (see Table II).

The *experts* show high scores in every MAPS category. A very similar group is called *unstructured experts*. These students show high scores in all of the categories except the *Logical Progression* category. Together, the experts are the two most successful groups of problem solvers in our study indicated by the large amount of correct solutions. The user "teal snake" is an example of an expertlike problem solver. His solution can be seen in Fig. 2. In total, 15 students are considered to be (unstructured) experts.

The *Docktor's Pattern* group shows high scores for the *Physics Approach* and the *Mathematical Procedures* categories but low scores for the *Specific Application of Physics* category. In our study, three students created this pattern in their solutions. An example solution can be seen in Fig. 3. The student "salmon bat" used the approach of equating kinetic and gravitational potential energy and was able to correctly calculate the energy values, but he failed in interpreting his results.

The *routine-exercised* group of students is characterized by high scores for the *Useful Description*, the *Specific*

TABLE II. Patterns in the MAPS scores. Scores are identified as "high" (4, 5, NA) or "low" (0, 1, 2, 3).

No.	Group	Description	Approach	Physics	Mathematics	Logic	Quantity	Correct
1	"Experts"	High	High	High	High	High	9	7
2	"Unstructured experts"	High	High	high	High	Low	6	5
3	"Docktor's pattern"		High	Low	High		3	1
4	"Routine-exercised"	High	Low	High	High		8	1
5	"Novices"	High	Low	Low	High	Low	4	0
6	"Approach failed"	High	Low	Low	Low	Low	12	0
7	"Completely failed"	Low	Low	Low	Low	Low	5	0
8	Miscellaneous						4	0

Application of Physics, and the *Mathematical Procedures*, but low scores for the *Physics Approach*. In this group, there are students with low or high scores for the *Logical Progression*. An example solution out of this group can be seen in Fig. 4. The students of this group identified at least one correct form of energy and used the corresponding formulas correctly. However, they were unable to connect the forms of energy using the principle of energy conservation appropriately. They therefore fail to solve the problem most of the time.

Problem solutions from the *Novices* group of students show high scores for the *Useful Description* and the *Mathematical Procedures* categories, but low scores for the other categories. Similar to the *routine-exercised* group, these students considered at least one of the two forms of energy separately, but they made mistakes while doing so. An example of a solution from this group can be seen in Fig. 5.

The largest group of students shows the *approach failed* pattern. These students achieved a high score for the *Useful Description*, but low scores for all other categories. Many of these solutions are incomplete. The students, for example, made a drawing of the problem situation, but canceled their solution process afterward—meaning there is no

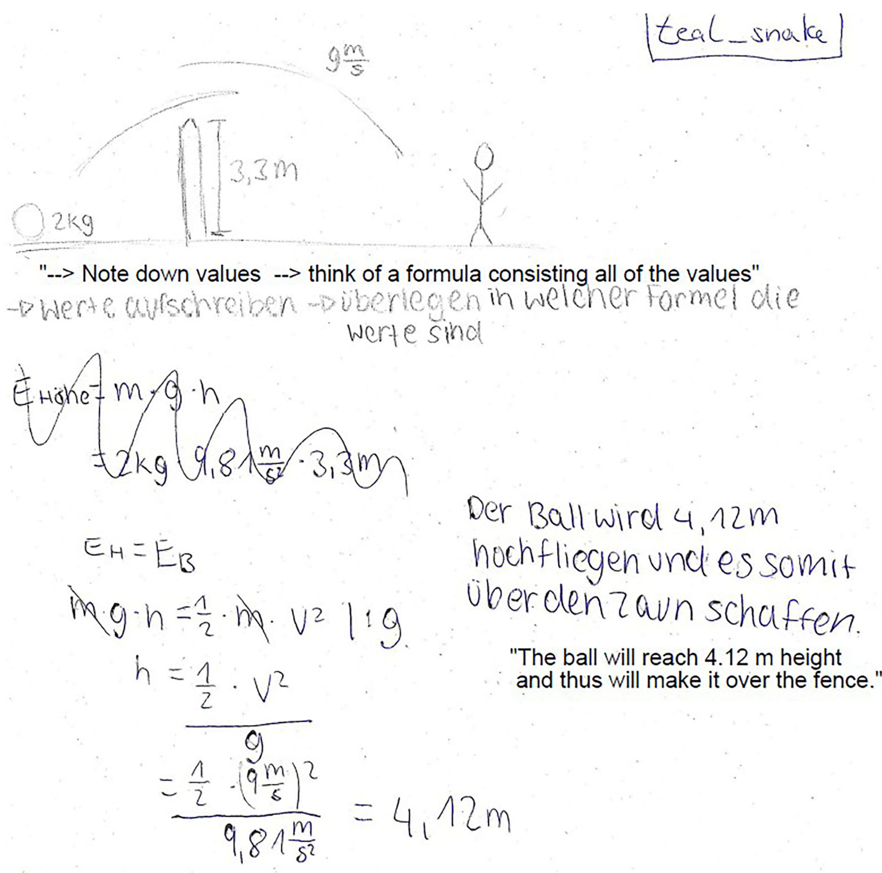


FIG. 2. Problem solution of the user "teal snake." This solution was assigned a score of 5 in every category and is considered an expertlike solution. It shows a correct solution to the problem task.

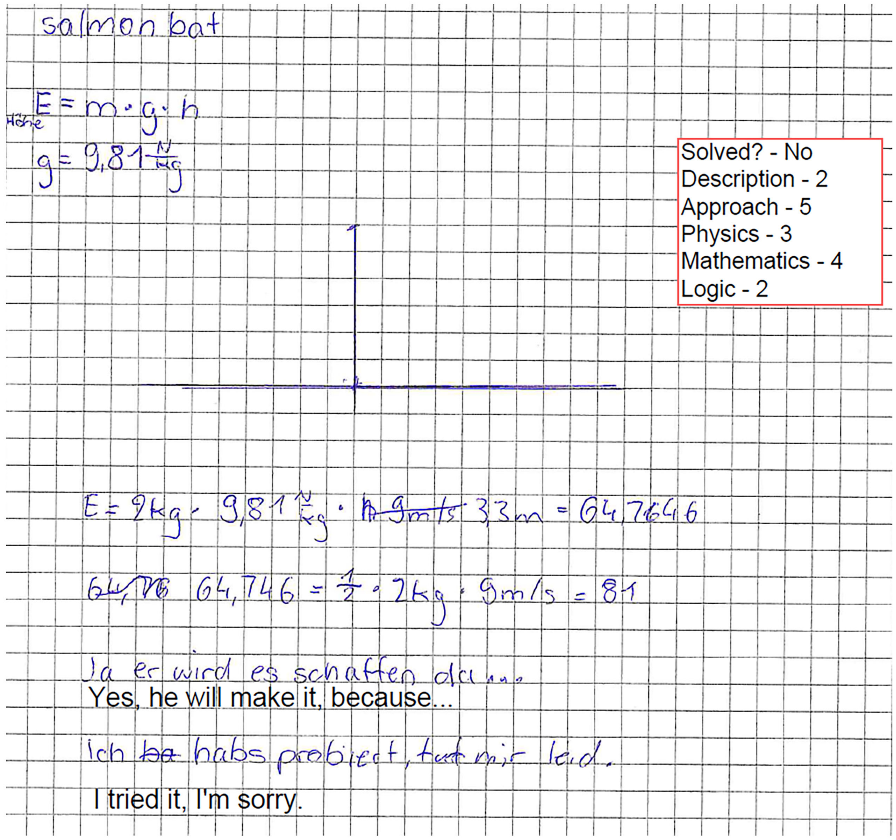


FIG. 3. Problem solution of the user "salmon bat." This solution was assigned to the *Docktor's Pattern* group. It shows a solution with a correct approach, but the students failed to interpret the results.

evidence of the other subprocesses of problem solving in these solutions.

The students' solutions of the *completely failed* group are even more incomplete. These students achieved low scores in every category. Some of these solutions are only blank sheets of paper; others are written explanations stating that the student had absolutely no idea what they were supposed to do in the task.

B. Conceptual knowledge about energy

Overall, the conceptual knowledge of the students is normally distributed (Shapiro-Wilk test: $p \approx 0.34$) across the entire scale with a mean total score of 10 (SD 4) correct items. The mean scores for the different conceptions of energy vary slightly. The students performed best for the transformation of energy, with a mean of 3.0 (SD 1.4) correct items. For the forms of energy conception, the students achieved a mean of 2.7 (SD 1.2) correct items. For the dissipation of energy, the students answered a mean of 2.2 (SD 1.3) items correctly. And the worst performing conception is the conservation of energy, with a mean of 1.7 (SD 1.3) correct items. More detailed information, like the individual scoring of every student, can be found in Table IV and Fig. 7 in Appendix B.

The energy test results are significantly different between students from year 9 and year 11. The difference in the transformation conception was analyzed using a t test and revealed no significant difference ($p > 0.05$). The results for the other conceptions did not allow a t test due to a lack of normality within the subsamples. They were therefore analyzed using a Wilcoxon rank-sum test. The WRS revealed significant differences for the conceptions: *Forms* ($p < 0.01$), *Conservation* ($p < 0.05$), and *Degradation* ($p < 0.05$). For all conceptions with significant differences, the year 11 students performed better than the year 9 students, with approximately one more correct item per conception.

The *expert* group concerning the MAPS score patterns shows varying results in the energy test. Most of the students of this group achieved relatively high scores between 11 and 17 correct items. However, there are also expertlike problem solvers who had poor results for the energy test. For example, "plum raccoon" answered only one item correctly in the energy test. The mean score is 9.8 (SD 4.8). The same applies to the *unstructured experts* group with slightly lower results in general (Mean 6.3, SD 3.4). In total, eight students achieved overall high scores for their problem solution but low scores for their conceptual knowledge.

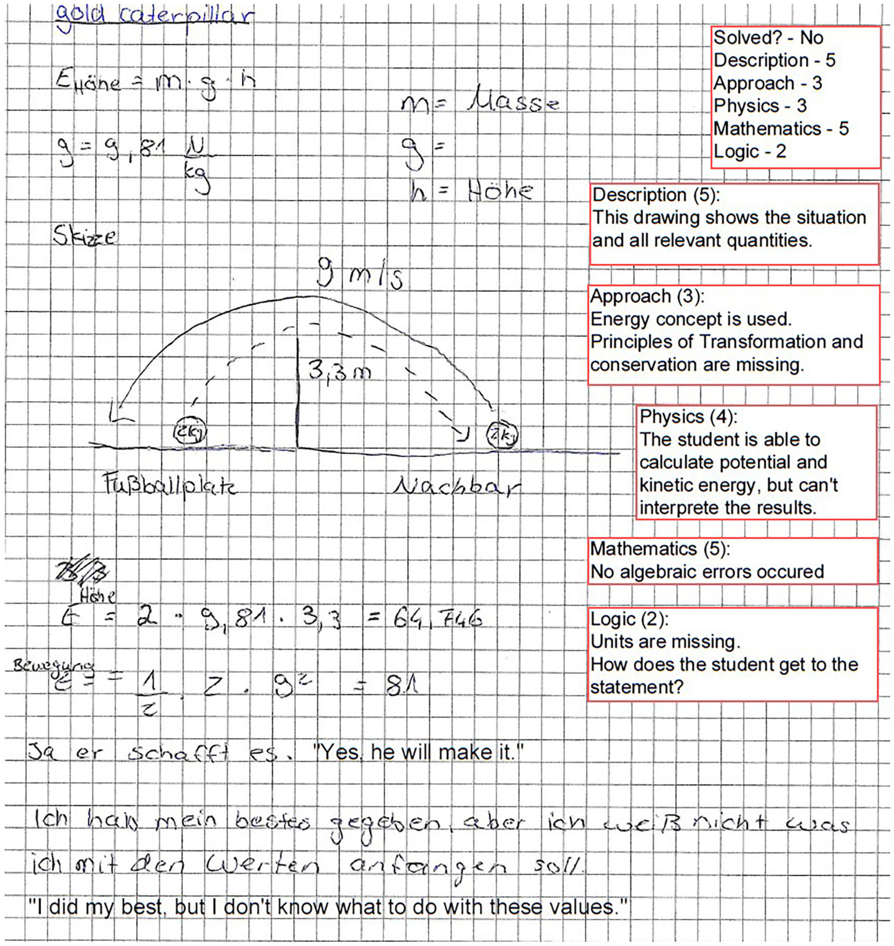


FIG. 4. Problem solution of the user "gold caterpillar." This problem solution is assigned to the *routine-exercised* group. It considers the two involved forms of energy separately and does not use the principle of energy conservation in order to solve the problem.

Students from the *approach failed* group scored mediocrely in the energy test with a mean of 9.5 (SD 1.9). These students achieved mean scores of 3 (SD 1) correct items for the *Forms* and the *Transformation* conceptions. For the *Conservation* conception, they achieved a mean of 1.4 (SD 0.8), and for the *Degradation*, 2.0 (SD 0.9) correct items.

The *completely failed* group shows relatively good results for the conceptual knowledge, with a mean total score of 10.8 (SD 2.9). In particular, 3 students achieved 13 points in total with 3 or 4 correct items for the *Conservation* conception.

VI. DISCUSSION

A. Useful problem for assessment

We want to investigate what learning needs about problem solving can be analyzed using the MAPS rubric on secondary school problem solutions. First, we discuss briefly whether the examined task poses a problem suitable for assessment using MAPS as a prerequisite for our study. In Sec. IV B, we explain the design process and the aspects of the task that are supposed to make it a problem. However, in Sec. II, we explain how the definition of a problem relies on the solver. Therefore, the designed task needs to be analyzed with respect to the study participants.

At first, only 14 out of 51 students were able to solve the task correctly, indicating that it was not too easy for our participants. Taking a closer look at the problem solutions, we can find clear indications that the task indeed was a problem for the students. The students from the *routine-exercised* group were able to calculate the involved forms of energy, but failed in interpreting these results (e.g., see Fig. 4). This solution demonstrates exemplary that the student is able to calculate different forms of energy under given conditions, but she is unable to use the law of conservation of energy as a principle to solve this problem. Therefore, simple energy calculations might be a routine exercise for this student, but applying conceptual knowledge about energy conservation to new situations is not.

Tasks that directly prompt students to execute certain steps of an expertlike problem solution, like making a drawing, can impede the assessment [17]. Our task is designed without any explicit hints on the desired process.

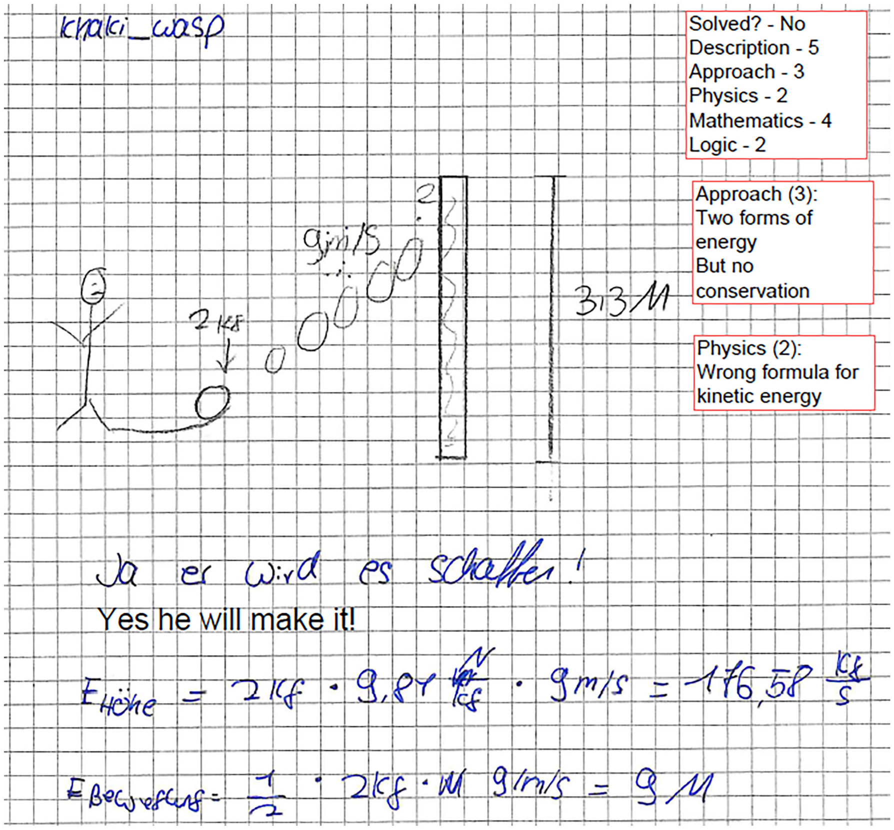


FIG. 5. Problem solution of the student "khaki wasp." This solution is assigned to the *novices* group. It considers the two forms of energy separately and contains physical errors.

As described in Sec. IV B, the problem can be solved by following different strategies. All of the explained strategies can be found in the students' solutions. This variety shows that the task provides sufficient scope for individual solutions and a need for active decision making without implicitly prescribing steps.

Overall, the problem solutions are rather short and the solution is relatively uncomplicated. This leads to high scores in the *Useful Description* category. In order to validly assess students' abilities in representing a problem solution, more complicated situations might be necessary. However, since our participants are in an early stage of their expertise development in physics, we could not use a more complicated problem, as the students would not be able to solve it, leading to floor effects. Even this problem was solved by only 14 out of 51 students.

Concerning the *Physics Approach* category, the presented problem shows strengths: In this category, the scores are relatively evenly distributed. Even though it is evident that the problem can be solved using the energy concept and the principle of energy conservation, there are decisions that our participants have to make during the problem-solving process. This decision making is an important skill for problem solving [17]. In routine exercises, the students are often required to find one unknown quantity and solve an equation for that quantity. In the used problem, the students were not asked to calculate a certain value, but they needed to compare two quantities in order to argue using the results. For that reason, there are multiple strategies for solving this problem, and the students need to make a decision on what they need to calculate. With this being a strength of our problem for assessing the *Physics Approach* category, it is also a weakness for the *Mathematical Procedures* category, because the students can reduce the necessary mathematical procedures to a minimum by simply calculating values with a known formula using a calculator. This leads to only a few students achieving medium scores, but many students at both ends of the scale.

The *Specific Application of Physics* category shows a slight tendency toward higher scores, but there are still 16 solutions that are at least partly inappropriate. The students were allowed to use a formula sheet; for that reason, in most of the solutions, the students used correct formulas to calculate the involved forms of energy. Since a wrong formula would cause deductions, the formula sheet influences the scores in this category. However, without the formulas given, many students would have canceled their solution process, because they usually do not know them by heart. Nonetheless, there are students failing to use the correct formulas with the correct quantities, causing low scores in this category.

Concerning the *Logical Progression* category, large differences between students can be found. Apparently, a well-structured problem-solving procedure is challenging for the participants, although we added instructions that they should document their solution clearly and pay attention to units. These instructions could have impaired the assessment since they are also hints for the participants on aspects that are taken into account for scoring. Such an influence toward higher scores cannot be observed in our results.

Based on these findings, we conclude the designed task is a useful physics problem for a rubric-guided assessment using MAPS. However, it must be borne in mind that both the task and the criteria for applying the rubric need to be adapted to the participants.

B. Analyzing student needs guided by the scoring

Docktor *et al.* [17] discussed their rubric in the context of analyzing students' needs by looking at their category scores. The developers of MAPS suggest that it can either provide information on further training needs for a group of students (e.g., a whole class) or be utilized to identify the specific abilities of individual students for targeted support [34]. However, there are only few examples of what patterns actually show useful information for formative assessment. For that reason, we analyzed different patterns in students' MAPS scores with respect to formative information.

1. Analyzing a group of students

For the analysis of the group of our participants, we used the median scores of each category. We did not analyze separate classes, because from some classes, only a few students participated in our study due to participation being voluntary. Also, separate analyses of year 9 and year 11 students were not done, because the problem-solving results show nearly no significant differences between these two groups. Medians were used because the MAPS scores are mainly ordinally scaled according to the developers [34]. However, we think it is debatable whether a score of 1 is actually better than a score of 0. A score of 0 is assigned if the solution does not include the respective category, and a score of 1 is assigned if everything concerning the respective category is incorrect. In both cases, the student's solution does not contain any correct aspects of a category. For the other scores, this differentiation was not done, for example, a score of 3 indicates that parts of the category are wrong, inappropriate, or missing [34].

In our study, the students achieved a median score of 4 for the *Useful Description* and the *Mathematical Procedures* categories. Thus, these categories show little room for improvement since students are already good at them; however, the problem used was relatively easy for representation and the mathematical procedures could be reduced to a minimum by choosing an easy physics approach. The lowest median scores of 2 were reached in the *Physics Approach* and *Logical Progression* categories, indicating a need for improvement among the participants. Many students have difficulties choosing the correct physical principles for a given situation. This finding is consistent with other research about students' problem solutions [5,16,26,28]. Also, the logical progression of the students' solutions needs to be improved by further training. The students could benefit from solving more problems like the one in this study, because they gain experience and improve their problem-scheme knowledge, possibly leading to a more structured way of solving problems. Additionally, students need to improve their documentation. Many deductions in the *Logical Progression* are caused by the solutions being unclear.

The energy test results of the participants are according to the learning progression of energy in secondary school [10]. Since the study took place in the learning unit where quantitative considerations of energy conservation are being taught, these two conceptions of the energy concept show overall lower results. Especially the conservation of energy seems to be challenging. The conceptions of *Forms* and *Transformation* that were taught earlier are better understood by the students. Here, they reached normally distributed results with means slightly above the middle of the scale (see Appendix B).

2. Docktor's patterns

As an example for the MAPS rubric indicating individual learning needs, Docktor *et al.* [17] pointed out that some of their participants showed low scores in the *Specific Application of Physics* but relatively high scores for the *Physics Approach* and the *Mathematical Procedures*. This pattern supposedly indicates that the students know the relevant principles for solving the problem and that they have the mathematical skills to use these principles, but they need further instruction on decision making to succeed in problem solving [17]. In our data, three students showed this pattern in their category scores. These students showed low to medium scores between 6 and 10 out of 20 for the conceptual knowledge in general and especially low scores of 0, 0, and 2 out of 5 for the *Conservation* conception. In contrast, they scored 4 or 5 out of 5 points for the *Transformation* conception. None of these three students wrote an answer using the conservation of energy or their calculated results. Two of them explicitly stated that they did not know how to interpret the results in regard to the problem situation. These findings suggest that the students not only fail at decision making but also did not fully understand the conservation of energy. That prevents them from successfully applying this principle to problem-solving. However, they have conceptual knowledge of energy being transformed and skills in calculating the different forms of energy, presumably enabling them to succeed in routine tasks. However, we found more than one interesting pattern in the MAPS scores.

3. Experts and unstructured experts

The experts show high scores in every category of the MAPS rubric. The unstructured experts also show high scores, with the exception of the logical progression category. The expert groups are also the most successful groups in solving the problem adequately. In regard to solving simple physics problems, these students do not seem to need much further assistance. The unstructured experts could benefit from being more structured in documenting their solution process. Especially when problems become more complex, they might face difficulties because they could lose track of their process.

Interestingly, about half of the students in these two groups achieved low scores in the energy test, indicating that they lack a solid conceptual knowledge about energy. This may indicate that these students have a solid problem-scheme knowledge that they can rely on during the task. This scheme knowledge enables them to solve the problem even without solid conceptual knowledge. In regard to formative assessment for everyday school teaching, this indicates a need for separate assessment of conceptual knowledge in addition to solving problem tasks, e.g., in exams.

The very low energy test results of some of the students could also be a sign of test fatigue. For example, the user "plum raccoon" answered only one item of the energy test correctly, which is even below a guessing rate. Perhaps this student did not work seriously on the energy test.

4. Routine-exercised and novices

The routine-exercised students' solutions use the energy concept in order to solve the problem. They identify at least one correct form of energy in the problem situation and use the correct formula to calculate the amount of energy. The same applies for the novices, yet the latter either use incorrect formulas or make other mistakes. However, both groups have in common that they fail to identify or use the principle of energy conservation. This is recognizable because the students either only use one form of energy and do not consider a transformation at all, or they calculate the gravitational potential energy and the kinetic energy separately and fail to argue with their results. Some of the students like "salmon bat" (see Fig. 3) or "gold caterpillar" (see Fig. 4) even state in their written solutions that they don't understand their calculated results and do not know how to use them.

In the energy test, the routine-exercised students performed well in all conceptions, with *Forms* (mean 3.3) and *Transformation* (mean 3.9) being better than *Conservation* (mean 2.5) and *Degradation* (mean 2.3). Thus, these students show solid conceptual knowledge. Combining the MAPS and energy test results indicates that these students understand the conservation of energy conceptually and that they are generally capable of solving energy tasks. However, they are not familiar with the practical implications of energy conservation for problem situations. These students would therefore benefit from practical training in order to generate problem-scheme knowledge for energy conservation problems.

The novice group shows relatively poor results in the energy test. Particularly for the conservation conception, these students performed very low, with 0 or 1 correct items out of 5. This low conceptual knowledge, together with the physical mistakes they made, shows that these students need to practice the energy formulas and that they need more conceptual teaching about energy conservation. After learning the concepts and becoming familiar with energy formulas, they can proceed to use this knowledge for problem solving [32].

5. Approach failed and completely failed

The students with MAPS score patterns of these groups were completely unable to solve the task. So, these students do need additional training on problem solving. Many students from these two groups did not apply the energy concept to the situation at all. This can either indicate that they need more conceptual knowledge about energy or that they need further instructions on how to apply the energy concept to problem situations.

In the energy test, the students from the approach failed group achieved mediocre scores. Especially concerning energy conservation, the students performed relatively poor with a mean of 1.4 correct items out of 5. These results indicate that the students additionally need conceptual training on the principle of energy conservation, before they can be trained to use this principle in physics problems. However, the students from the completely failed group show relatively high scores in the energy test. In particular, 3 students achieved a total of 13 correct items with scores of 3–4 for the conservation conception. This indicates that these students seem to have a solid conceptual knowledge of energy, but they fail to apply this knowledge to the problem situation. These students do not need additional teaching about the energy concept, but they need to improve their problem-scheme knowledge.

Three students returned a blank page or a written statement that they had absolutely no idea how to solve this problem. These scores make the analysis of students' needs difficult. Since the students' solutions do not contain any application of physics or mathematics, we cannot analyze if the student would have been able to use the energy formulas correctly. This indicates that these students definitely need training on decision-making for the *Devise a Plan* phase, but in order to individually assist these students, further assessments are necessary. An "atomized" approach [54,55], where the different phases of problem-solving are assessed separately, can help further analyze these students' needs.

VII. CONCLUSIONS AND PRACTICAL CONSIDERATIONS

In conclusion, the results of our study showed that the MAPS rubric can be applied to problem solutions with a lower degree of proficiency. The problem tasks and the scoring criteria need to be adapted to the participants and their anticipated skills. A lower level of conceptual and problem scheme knowledge can lead to difficulties in using the complete scale of the rubric, e.g., because the approach for the problem is relatively straight-forward to choose. The same applies for the mathematical procedures. Since the students in our study only know a few isolated formulas and were allowed to use a formula sheet, the procedures are relatively easy. As a result, the scoring of this category gets nearly binary as students either achieve high scores of 4–5 or low scores of 1. Nonetheless, the MAPS rubric was generally easy to apply, and it gave significant insights into how much expert-like behavior was demonstrated.

In our study, we were able to identify seven different patterns of MAPS scores. A qualitative comparison of problem solutions with similar MAPS scores revealed common possible needs for further instruction. This indicates that the MAPS rubric can indeed help instructors to analyze students' needs. Particularly when students seemed to fail applying conceptual knowledge, the energy test results helped to identify whether these students lacked the needed knowledge or whether they knew the concepts but fail to use them in problem solving. The application of standardized procedures like this rubric might ease formative assessment for teachers and for future automatic feedback tools.

For more reliable information, the assessment should contain more than one problem, because a student might not be able to find an approach for one problem and therefore cancel the solution process quickly, but might be able to solve another problem very well. For everyday school teaching, this might be challenging for teachers to provide. The two human coders in our study took approximately 4 min per problem solution after they were familiar with the rubric. If a teacher were, for example, to score three problem solutions per student in a class of 25 students, that would require an effort of nearly 5 h. Current developments in the field of artificial intelligence for physics problem-solving [55–57] have the potential to enable further automation of the scoring processes. This would allow digital learning environments to use a rubric for adaptive learning.

Assessment methods using qualitative analyses of complete problem solutions might be inappropriate for students who generally struggle to find an approach. These students cannot show other potential qualities of their physics abilities. For cases like this, an "atomized" approach [54,55] that assesses the phases of problem solving separately can be more appropriate. However, an "atomized" approach uses step-by-step instructions that exclude active decision making from the assessment. For these reasons, a combination of items concerning the subprocesses and qualitative analyses of complete problem solutions seems to be appropriate for a comprehensive assessment of problem-solving abilities.

ACKNOWLEDGMENTS

This work was supported within the Project "LernMINT—data-driven teaching in STEM subjects" by the Ministry of Science and Culture Lower Saxony (Grant No. 51410078) and by the German Academic Scholarship Foundation (Doctoral Scholarship). The publication of this article was funded by the Open Access Fund of Leibniz Universität Hannover. In the preparation of this manuscript, language editing was conducted with the assistance of ChatGPT, a large language model developed by OpenAI. ChatGPT was employed to enhance the language quality of the text while maintaining the integrity of the original content. We would like to express our sincere thanks to the reviewers for their constructive feedback during the review process, which helped improve the quality of this article.

DATA AVAILABILITY

The data that support the findings of this article are openly available [58].

APPENDIX A: COMPLETE RESULTS OF THE HUMAN RATING

TABLE III. Complete results of the consensus rating of the students' problem solutions using the MAPS rubric. All NA scores are NA (solver).

Username	year	Solved?	Description	Approach	Physics	Mathematics	Logic	Group
Beige wombat	11	Yes	Not applicable	5	4	5	4	Expert
Black butterfly	11	No	Not applicable	3	4	4	2	Routine-exercised
Black deer	9	No	4	2	3	4	2	Novice
Black pelican	11	No	5	5	3	4	4	Expert
Black raccoon	9	No	5	1	1	1	0	Approach failed
Black scallop	11	No	5	1	5	5	2	Routine-exercised
Blue crocodile	9	No	5	5	5	4	4	Expert
Brown lion	9	No	5	3	5	5	4	Routine-exercised
Crimson chicken	11	No	5	2	0	0	1	Approach failed
Cyan bat	11	No	5	0	5	4	0	Routine-exercised
Cyan bird	11	No	Not applicable	5	3	4	3	Docktor's pattern
Cyan bluewhale	9	Yes	5	3	4	5	4	Routine-exercised
Gold anaconda	11	No	Not applicable	5	5	5	3	Unstructured expert
Gold caterpillar	9	No	5	3	3	5	2	Novice
Gold crayfish	9	No	5	2	0	0	2	Approach failed
Gray lion	9	No	4	0	0	0	0	Approach failed
Gray mockingbird	11	No	3	2	3	5	3	Miscellaneous
Green clam	11	No	5	3	0	0	0	Approach failed
Khaki wasp	9	No	5	3	2	4	2	Novice
Lavender chipmunk	9	No	4	0	0	0	0	Approach failed
Lavender fish	11	Yes	5	5	5	4	5	Expert
Lavender swan	9	Yes	Not applicable	Not applicable	5	5	4	Expert
Lavender swordfish	9	No	0	0	0	0	1	Completely failed
Lavender wombat	11	No	5	1	3	2	2	Approach failed
Magenta hedgehog	9	Yes	Not applicable	Not applicable	5	5	3	Unstructured expert
Maroon pony	9	Yes	5	5	4	4	3	Unstructured expert
Olive flamingo	9	No	4	2	2	3	3	Approach failed
Olive snake	9	No	0	0	0	0	0	Completely failed
Orange camel	11	Yes	4	5	4	5	3	Unstructured expert
Pink anteater	9	No	0	0	0	0	0	Completely failed
Plum kangaroo	9	No	5	0	0	0	0	Approach failed
Plum raccoon	9	Yes	Not applicable	5	5	4	4	Expert
Purple bird	11	No	4	1	1	0	5	Miscellaneous
Purple firefly	9	No	2	0	0	0	1	Completely failed
Red dinosaur	11	Yes	5	5	4	5	5	Expert
Red octopus	9	No	0	2	2	3	2	Completely failed
Salmon bat	9	No	2	5	3	4	2	Docktor's pattern
Salmon dragon	11	Yes	4	Not applicable	4	5	3	Unstructured expert
Salmon leopard	9	Yes	5	5	5	4	4	Expert
Silver bison	11	No	5	2	4	4	2	Routine-exercised
Silver camel	11	No	5	1	4	4	4	Routine-exercised
Silver crocodile	9	No	4	1	4	4	4	Routine-exercised
Silver takin	9	Yes	4	Not applicable	5	5	2	Unstructured expert
Teal blackpanther	9	No	4	0	1	1	0	Approach failed
Teal lemming	11	No	5	1	0	0	1	Approach failed
Teal snake	11	Yes	5	5	5	5	5	Expert
Violet chameleon	9	Yes	5	5	3	5	4	Docktor's pattern
Violet gecko	9	No	3	2	5	0	1	Miscellaneous
Violet giraffe	11	No	5	0	0	0	0	Approach failed
White coral	11	No	3	2	2	5	2	Miscellaneous
White fish	9	No	5	3	3	4	2	Novice
Median scores			5	2	3	4	2	
Interrater agreement (α)		0.95	0.84	0.92	0.90	0.91	0.84	

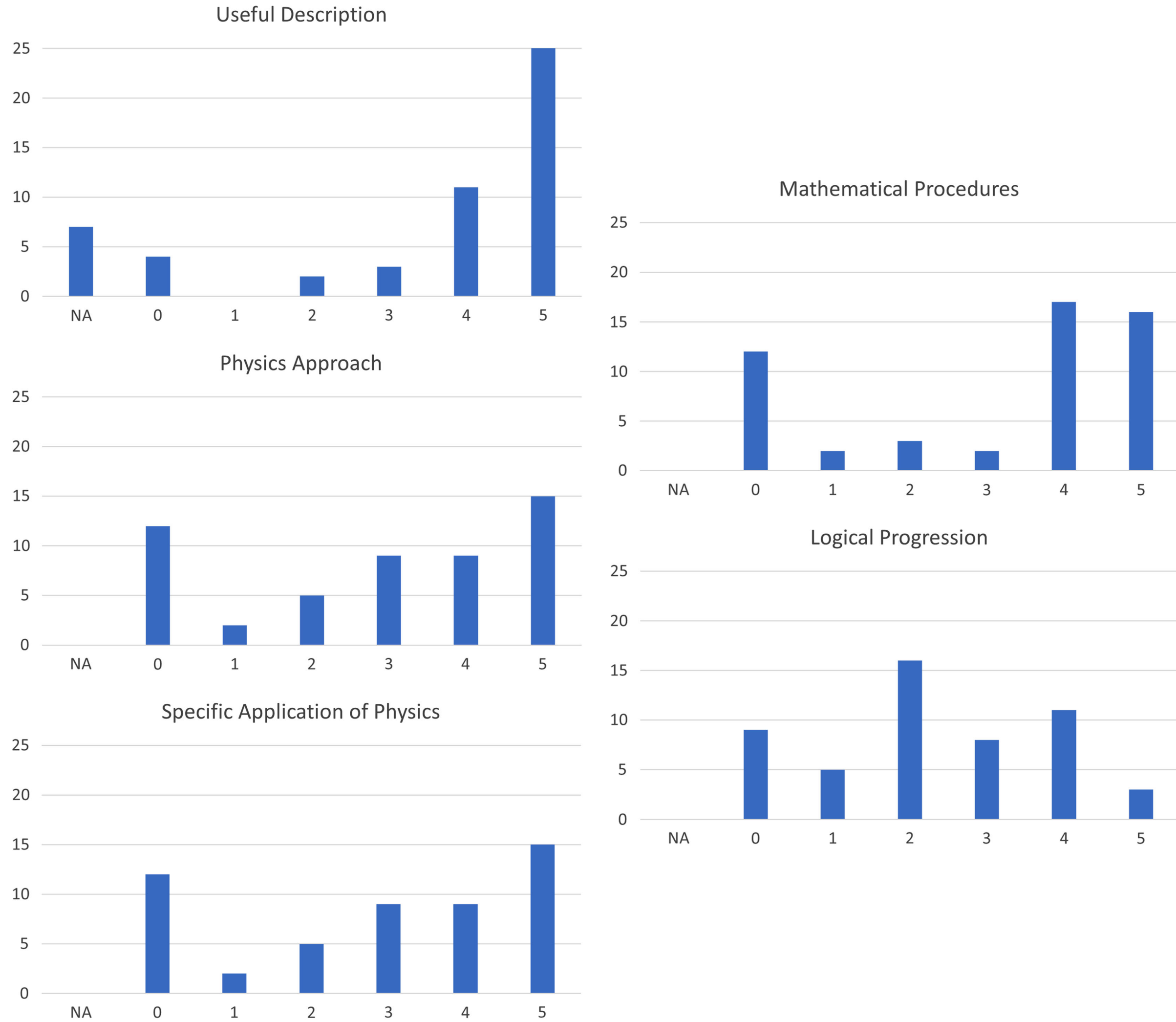


FIG. 6. Histograms showing the distribution of MAPS scores for the different categories.

APPENDIX B: RESULTS OF THE ENERGY CONCEPT ASSESSMENT

TABLE IV. Complete results of the energy concept assessment.

Username	Class	Forms	Transformation	Conservation	Degradation	Total
Beige wombat	4	5	4	3	1	13
Black butterfly	2	4	5	4	3	16
Black deer	1	3	5	1	2	11
Black pelican	3	3	3	2	3	11
Black raccoon	1	4	2	1	3	10
Black scallop	2	5	5	3	5	18
Blue crocodile	7	3	5	3	2	13
Brown lion	1	2	4	1	1	8
Crimson chicken	7	4	4	2	1	11
Cyan bat	2	3	2	3	2	10
Cyan bird	2	1	2	1	2	6
Cyan bluewhale	1	1	3	1	1	6
Gold anaconda	2	3	3	1	3	10
Gold caterpillar	1	0	2	0	2	4
Gold crayfish	7	4	4	2	2	12
Gray lion	1	1	4	1	2	8
Gray mockingbird	2	4	4	3	5	16
Green clam	5	4	3	2	2	11
Khaki wasp	1	0	4	0	2	6
Lavender chipmunk	1	2	3	1	4	10
Lavender fish	3	4	5	4	4	17
Lavender swan	1	3	3	1	1	8
Lavender swordfish	7	3	5	4	1	13
Lavender wombat	6	3	3	1	1	8
Magenta hedgehog	7	0	0	0	0	0
Maroon pony	1	3	3	0	1	7
Olive flamingo	1	2	1	3	2	8
Olive snake	1	3	2	3	5	13
Orange camel	5	2	1	0	2	5
Pink anteater	1	3	2	3	5	13
Plum kangaroo	7	3	3	2	1	9
Plum raccoon	7	1	0	0	0	1
Purple bird	5	2	4	3	3	12
Purple firefly	7	2	2	1	1	6
Red dinosaur	5	2	1	0	2	5
Red octopus	1	2	3	1	3	9
Salmon bat	1	3	4	2	1	10
Salmon dragon	5	3	2	2	3	10
Salmon leopard	7	2	2	1	1	6
Silver bison	2	3	4	4	0	11
Silver camel	2	5	5	3	5	18
Silver crocodile	1	3	3	1	1	8
Silver takin	1	2	1	1	2	6
Teal blackpanther	1	4	5	1	3	13
Teal lemming	6	3	3	1	2	9
Teal snake	3	4	3	3	4	14
Violet chameleon	1	2	5	0	3	10
Violet gecko	1	2	2	1	1	6
Violet giraffe	4	2	3	0	1	6
White coral	2	4	2	4	3	13
White fish	1	3	1	0	2	6
Mean scores		2.7	3.0	1.7	2.2	9.6
Standard deviations		1.2	1.4	1.3	1.3	3.6

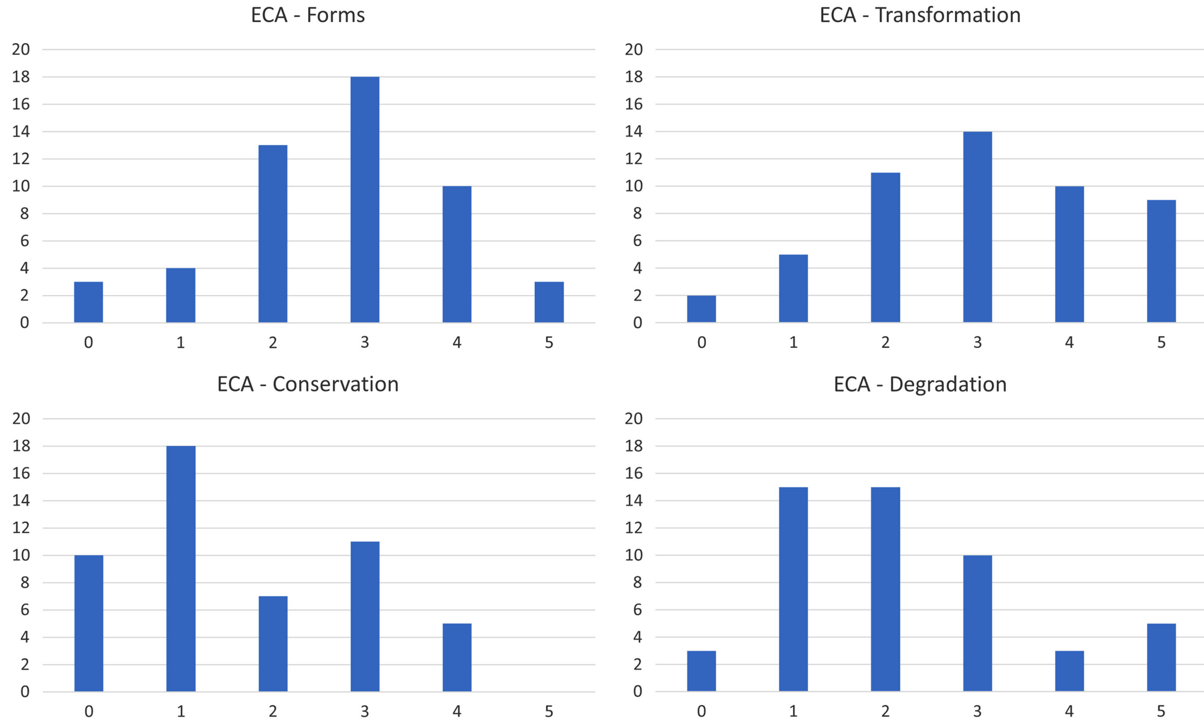


FIG. 7. Histograms of the number of correct items for the four different conceptions of energy within the energy concept assessment.

[1] OECD, *PISA 2022 Results (Volume I): The State of Learning and Equity in Education* (OECD Publishing, Paris, France, 2023), 10.1787/53f23881-en.

[2] J. Tuminaro and E. F. Redish, Elements of a cognitive model of physics problem solving: Epistemic games, Phys. Rev. ST Phys. Educ. Res. **3**, 020101 (2007).

[3] T. Nilsen, C. Angell, and L. S. Grønmo, Mathematical competencies and the role of mathematics in physics education: A trend analysis of TIMSS advanced 1995 and 2008, Acta Didact. Norge **7**, 6 (2013).

[4] T. Tong, F. Pi, S. Zheng, Y. Zhong, X. Lin, and Y. Wei, Exploring the effect of mathematics skills on student performance in physics problem-solving: A structural equation modeling analysis, Res. Sci. Educ. **55**, 489 (2025).

[5] M. N. Kholid, F. Rofi'ah, N. Ishartono, M. Waluyo, S. Maharani, A. Swastika, N. Faiziyah, and C. K. Sari, What are students' difficulties in implementing mathematical literacy skills for solving PISA-like problem?, J. Higher Educ. Theory Pract. **22**, 181 (2022).

[6] A. C. Frenzel, R. Pekrun, A.-L. Dicke, and T. Goetz, Beyond quantitative decline: Conceptual shifts in adolescents' development of interest in mathematics, Dev. Psychol. **48**, 1069 (2012).

[7] P. Potvin and A. Hasni, Interest, motivation and attitude towards science and technology at k-12 levels: A systematic review of 12 years of educational research, Stud. Sci. Educ. **50**, 85 (2014).

[8] L. Steidtmann, T. Kleickmann, and M. Steffensky, Declining interest in science in lower secondary school classes: Quasi-experimental and longitudinal evidence on the role of teaching and teaching quality, J. Res. Sci. Teach. **60**, 164 (2023).

[9] C. V. Schwarz, B. J. Reiser, E. A. Davis, L. Kenyon, A. Achér, D. Fortus, Y. Shwartz, B. Hug, and J. Krajcik, Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners, J. Res. Sci. Teach. **46**, 632 (2009).

[10] K. Neumann, T. Viering, W. J. Boone, and H. E. Fischer, Towards a learning progression of energy, J. Res. Sci. Teach. **50**, 162 (2013).

[11] H. Jang, Identifying 21st century stem competencies using workplace data, J. Sci. Educ. Technol. **25**, 284 (2016).

[12] *Complex Problem Solving: The European Perspective*, edited by P. A. Frensch and J. Funke (Psychology Press, New York City, NY, 1995), https://psycnet.apa.org/record/1995-97947-000.

[13] M. E. Martinez, What is problem solving, Phi Delta Kappan **79**, 605 (1998), http://www.jstor.org/stable/20439287.

[14] *The Nature of Problem Solving: Using Research to Inspire 21st Century Learning, Educational Research and Innovation*, edited by B. Csapo and J. Funke (OECD Publishing, Paris, France, 2017), 10.1787/9789264273955-en.

[15] M. U. Smith, *Toward a Unified Theory of Problem Solving: A View from Biology* (Routledge, Hillsdale, NJ, 1988), https://eric.ed.gov/?id=ED295978.

[16] A. H. Schoenfeld, *Mathematical Problem Solving* (Academic Press, Orlando, FL, 1985).

[17] J. Docktor, J. Dornfeld, E. Frodermann, K. Heller, L. Hsu, K. Jackson, A. Mason, Q. Ryan, and J. Yang, Assessing student written problem solutions: A problem-solving rubric with application to introductory physics, Phys. Rev. Phys. Educ. Res. **12**, 010130 (2016).

[18] K. Mosier, U. Fischer, R. R. Hoffman, and G. Klein, Expert professional judgments and "naturalistic decision making", in *The Cambridge Handbook of Expertise and Expert Performance*, edited by K. A. Ericsson, R. R. Hoffman, A. Kozbelt, and A. M. Williams (Cambridge University Press, Cambridge, England, 2018), Chap. 25, pp. 453–475, 10.1017/9781316480748.025.

[19] A. Price, S. Salehi, E. Burkholder, C. Kim, V. Isava, M. Flynn, and C. Wieman, An accurate and practical method for assessing science and engineering problemsolving expertise, Int. J. Sci. Educ. **44**, 2061 (2022).

[20] J. H. Larkin and F. Reif, Understanding and teaching problem-solving in physics, Eur. J. Sci. Educ. **1**, 191 (1979).

[21] D. Dörner and J. Funke, Complex problem solving: What it is and what it is not, Front. Psychol. **8**, 1153 (2017).

[22] G. Pólya, *How to Solve It* (Princeton University Press, Princeton, NJ, 1945).

[23] G. Friege, *Wissen und Problemlösen: Eine Empirische Untersuchung des Wissenszentrierten Problemlösens im Gebiet der Elektrizitätslehre auf der Grundlage des Experten-Novizen-Vergleichs* (Logos Verlag, Berlin, Germany, 2001).

[24] D. Ramalingam, R. Philpot, and B. McCrae, The PISA 2012 assessment of problem solving, in *The Nature of Problem Solving: Using Research to Inspire 21st Century Learning, Educational Research and Innovation*, edited by B. Csapo and J. Funke (OECD Publishing, Paris, France, 2017), Chap. 7, pp. 75–91, 10.1787/9789264273955-7-en.

[25] K. van Lehn, Problem solving and cognitive skill acquisition, in *The Foundations of Cognitive Science*, edited by M. Posner (MIT Press, Cambridge, MA, 1989), Chap. 14, pp. 527–580, https://psycnet.apa.org/record/1990-97026-014.

[26] F. Reif and J. I. Heller, Knowledge structure and problem solving in physics, Educ. Psychol. **17**, 102 (1982).

[27] G. Friege and G. Lind, Types and qualities of knowledge and their relations to problem solving in physics, Int. J. Sci. Math. Educ. **4**, 437 (2006).

[28] S. Lenzer, B. Smarsly, and N. Graulich, How do students become experts? An in-depth study on the development of domain-specific awareness in a materials chemistry course., Int. J. Sci. Educ. **42**, 2032 (2020).

[29] M. T. H. Chi, R. Glaser, and E. Rees, *Expertise in Problem Solving* (ERIC, Pittsburgh, PA, 1981), https://files.eric.ed.gov/fulltext/ED215899.pdf.

[30] J. P. Mestre, J. L. Docktor, N. E. Strand, and B. H. Ross, Conceptual problem solving in physics, in *Psychology of Learning and Motivation: Advances in Research and Theory* (Academic Press, San Diego, 2011), Vol. 55, p. 269, 10.1016/B978-0-12-387691-1.00009-0.

[31] J. L. Docktor, N. E. Strand, J. P. Mestre, and B. H. Ross, Conceptual problem solving in high school physics, Phys. Rev. ST Phys. Educ. Res. **11**, 020106 (2015).

[32] C. A. Ogilvie, Moving students from simple to complex problem solving, in *Learning to Solve Complex Scientific Problems*, edited by D. H. Jonassen (Routledge, New York City, NY, 2007), Chap. 7, pp. 159–186.

[33] G. Taasoobshirazi and J. Farley, Construct validation of the physics metacognition inventory, Int. J. Sci. Educ. **35**, 447 (2012).

[34] J. Docktor, Development and validation of a physics problem-solving assessment rubric, Ph.D. thesis, University of Minnesota, 2009, https://hdl.handle.net/11299/56637.

[35] R. Warnakulasooriya, D. J. Palazzo, and D. E. Pritchard, Time to completion of web-based physics problems with tutoring, J. Exp. Anal. Behav. **88**, 103 (2007).

[36] Y.-J. Lee, D. J. Palazzo, R. Warnakulasooriya, and D. E. Pritchard, Measuring student learning with item response theory, Phys. Rev. ST Phys. Educ. Res. **4**, 010102 (2008).

[37] OECD, *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy* (OECD Publishing, Paris, France, 2013), 10.1787/53f23881-en.

[38] R. Kelly, E. McLoughlin, and O. E. Finlayson, Analysing student written solutions to investigate if problem-solving processes are evident throughout, Int. J. Sci. Educ. **38**, 1766 (2016).

[39] *Protocol Analyses*, edited by K. A. Ericsson and H. A. Simon (MIT Press, Cambridge, MA, 1993), 10.7551/mitpress/5657.001.0001.

[40] M. W. van Someren, Y. F. Barnard, and J. A. Sandberg, *The Think Aloud Method: A Practical Guide to Modelling Cognitive Processes* (Academic Press, London, 1994), https://hdl.handle.net/11245/1.103289.

[41] A. Jonsson and G. Svingby, The use of scoring rubrics: Reliability, validity and educational consequences, Educ. Res. Rev. **2**, 130 (2007).

[42] P. Tschisgale, M. Kubsch, P. Wulff, S. Petersen, and K. Neumann, Exploring the sequential structure of students' physics problem-solving approaches using process mining and sequence analysis, Phys. Rev. Phys. Educ. Res. **21**, 010111 (2025).

[43] M. E. Omiles and E. J. D. Ramirez, Peer learning framework in the new normal, J. Pract. Stud. Educ. **6**, 14 (2025).

[44] G. Pelobillo, Assessment development in measuring students' context-rich problem-solving ability in physics, Int. J. Instr. **15**, 855 (2022).

[45] G. Zavala, A. Dominguez, and R. R. Gallegos, ACE: Innovative educational model to teach physics and mathematics for engineering students, in *Proceedings of the 2013 ASEE Annual Conference and Exposition*, ASEE Conferences (Atlanta, Georgia, 2013), 10.18260/1-2–19147.

[46] Z. A. Cashata, D. G. Seyoum, and F. E. Gashaw, The effect of using blended jigsaw-4 problem-solving instruction on preservice physics teachers' problem solving skill, Anatol. J. Educ. **8**, 35 (2023).

[47] M. R. A. Taqwa, R. Faizah, L. Rivaldo, D. E. Saputri, F. N. Aini, and M. I. Shodiqin, Students' problemsolving ability in temperature and heat concepts, J. Phys. Conf. Ser. **1339**, 012132 (2019).

[48] S. Shanta, Investigation of problem solving skills among 12th grade engineering students, Ph.D. thesis, Virginia State University, 2017, 10.13140/RG.2.2.28511.92323.

[49] S. Shanta and J. G. Wells, T/e design based learning: Assessing student critical thinking and problem solving abilities, Int. J. Technol. Des. Educ. **32**, 267 (2022).

[50] J. Funke, *Problemlösendes Denken [Problem Solving Thinking]* (Kohlhammer, Stuttgart, Germany, 2003), 10.17433/978-3-17-022830-6.

[51] M. Dudzinska, *Lernen mit Beispielaufgaben und Feedback im Physikunterricht der Sekundarstufe 1: Energieerhaltung zur Lösung von Aufgaben Nutzen* (Logos Verlag, Berlin, Germany, 2019).

[52] R. Duit, Teaching, and learning the physics energy concept, in *Teaching, and Learning of Energy in K–12 Education*, edited by R. F. Chen, A. Eisenkraft, D. Fortus, J. Krajcik, K. Neumann, J. Nordine, and A. Scheff (Springer International Publishing, Cham, Switzerland, 2014), pp. 67–85, 10.1007/978-3-319-05017-1_5.

[53] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 2nd ed. (Sage Publications, Los Angeles, CA, 2004), 10.4135/9781071878781.

[54] M. Brandenburger, *Was beeinflusst den Erfolg beim Problemlösen in der Physik? Eine Untersuchung mit Studierenden* (Logos Verlag, Berlin, Germany, 2016).

[55] A. Meyer, T. Bleckmann, and G. Friege, Automatic feedback on physics tasks using open-source generative artificial intelligence, Int. J. Sci. Educ. 1 (2025).

[56] P. Tschisgale, P. Wulff, and M. Kubsch, Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory, Phys. Rev. Phys. Educ. Res. **19**, 020123 (2023).

[57] F. Kieser, P. Tschisgale, S. Rauh, X. Bai, H. Maus, S. Petersen, M. Stede, K. Neumann, and P. Wulff, David vs. Goliath: Comparing conventional machine learning and a large language model for assessing students' concept use in a physics problem, Front. Artif. Intell. **7**, 1408817 (2024).

[58] See Supplemental Material at http://link.aps.org/supplemental/10.1103/3bs7-fnrd for alphabetically ordered scans of every student solution collected within this study.