



## OPEN ACCESS

## EDITED BY

Konstantinos T. Kotsis,  
University of Ioannina, Greece

## REVIEWED BY

Salman Rashid,  
Yogyakarta State University, Indonesia  
Selma Riyasni,  
Padang State University, Indonesia

## \*CORRESPONDENCE

André Meyer  
✉ a.meyer@idmp.uni-hannover.de

RECEIVED 03 December 2025

REVISED 20 January 2026

ACCEPTED 22 January 2026

PUBLISHED 13 February 2026

## CITATION

Meyer A and Friege G (2026) An atomized approach to assessing energy problem solving in physics using multidimensional item response theory.  
*Front. Educ.* 11:1759878.  
doi: 10.3389/feduc.2026.1759878

## COPYRIGHT

© 2026 Meyer and Friege. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# An atomized approach to assessing energy problem solving in physics using multidimensional item response theory

André Meyer\* and Gunnar Friege

Department of Mathematics and Physics, Institute for Mathematics and Physics Education, Physics Education Group, Leibniz University Hannover, Hanover, Germany

**Introduction:** Problem solving is a central competence in STEM education, yet many secondary school students struggle to coordinate the multiple skills required for successful problem solving. Early assessment of problem-solving skills can support individual feedback during this pivotal phase of schooling. However, existing assessment approaches focus mainly on complete problem solutions, which are resource-intensive and cannot adequately capture skills of students who fail in early phases of the problem-solving process.

**Methods:** To address this gap, the atomized problem-solving test (APST) was developed as a digital instrument that independently assesses four problem-solving subprocesses: Representation, Planning, Execution, and Evaluation. The APST was evaluated in two consecutive studies with a total of 800 German secondary school students within a web-based learning environment on energy conservation. Multidimensional item response theory (MIRT) was used to examine item quality and dimensional structure, complemented by supplemental assessments of conceptual knowledge, school grades, and rubric-based analyses of written problem solutions.

**Results:** The analyses supported a four-dimensional structure aligned with the theoretical design of the APST. The items showed acceptable model fit and reliable measurement of the intended subprocesses. All APST dimensions were moderately associated with conceptual knowledge of energy and with school grades in physics and mathematics, while no meaningful correlations were found with gender or native language. Evaluation emerged as a distinctive subprocess, showing strong associations with other subprocesses—particularly Execution—alongside evaluation-specific skills.

**Discussion:** The results indicate that the APST enables valid and reliable assessment of problem-solving subprocess skills in secondary physics education. At the same time, the findings underscore limitations of atomized assessments for measuring general problem-solving competence, as independent decision making is not assessed. The prominent role of Evaluation highlights its integrative function within the problem-solving process and points to important implications for both assessment design and future research.

## KEYWORDS

assessment, energy, multidimensional item response theory, physics education, problem solving, secondary school teaching

## 1 Introduction

Problem solving is a complex and challenging task that is considered to be one of the most important skills students have to achieve in STEM education (Jang, 2016). It requires multiple cognitive abilities (Träff et al., 2019). Especially quantitative physics problem solving requires solid conceptual knowledge and mathematical skills that have to be applied to unfamiliar situations (OECD, 2023; Tong et al., 2025; Nilsen et al., 2013; Tuminaro and Redish, 2007). Students are firstly introduced to such quantitative considerations in secondary physics teaching. They are taught about the principle of energy conservation in the context of mechanics and thermodynamics. In this context, they learn how to apply energy formulas to physics problems. This is a pivotal stage of secondary physics education, however, a declining interest and motivation for STEM subjects can be observed during that time (Potvin and Hasni, 2014; Frenzel et al., 2012).

The work presented in this article was done as part of a project that aims to develop a digital learning environment for training these important quantitative problem-solving skills. In order for that learning environment to be effectively adapted to the individual needs of the students, an automatic assessment of problem solving is essential (e.g., Plass and Pawar, 2020; Lee et al., 2024).

In general, a distinction is made between problem solving and routine exercises (e.g., Smith, 1991). This distinction does not implicate that a problem task is necessarily more difficult than a routine task, but it is made based on the necessary cognitive processes. A problem can be defined as a task where a defined beginning state has to be transformed into a desired end state without an immediately apparent solution path (e.g., Martinez, 1998; Csapó and Funke, 2017; Dörner and Funke, 2017). Solving a problem task requires active decision making on which concepts and principles to use (Mosier et al., 2018; Price et al., 2022). A routine exercise, in contrast, is a task where the solver knows from the beginning how to get to the solution. For that reason, it depends on the solver, whether a task poses a problem or a routine task for them (Martinez, 1998).

For decades, various models for problem-solving processes were defined. These models have in common that they divide problem-solving processes into different phases of subprocesses. For example, Pólya (1945) defined the following four steps of solving a mathematics problem: “understand the problem,” “make a plan,” “carry out the plan,” and “look back.” Friege (2001) defined four similar phases, that he called “problem representation,” “development or selection of a problem scheme,” “elaboration of a solution,” and “evaluation of the solution.” The PISA problem-solving assessment is based on a similar process model as well, although this model uses a slightly different division of the subprocesses. In the PISA model, there are the subprocesses “exploring and understanding,” “representing and formulating,” “planning and executing,” and “monitoring and reflecting” (Ramalingam et al., 2017). During the PISA assessment, students need to answer items including multiple choice, drag and drop, and written solution formats using computers (OECD, 2013).

The PISA problem-solving model has also been used for qualitative analysis of written problem solutions (Kelly et al., 2016). Such qualitative analyses are very close to the process under

examination which enables valid and reliable assessment. They can be used to gather insights into the problem-solving process, e.g., to explore the sequential structure (Tschisgale et al., 2025) or to compare expert-like and novice-like problem solutions (Docktor et al., 2016). Rubrics like the Minnesota Assessment of Problem Solving (MAPS) can be used to quantify the results of qualitative analyses in order to make them accessible for comparisons to other assessments (Docktor et al., 2015).

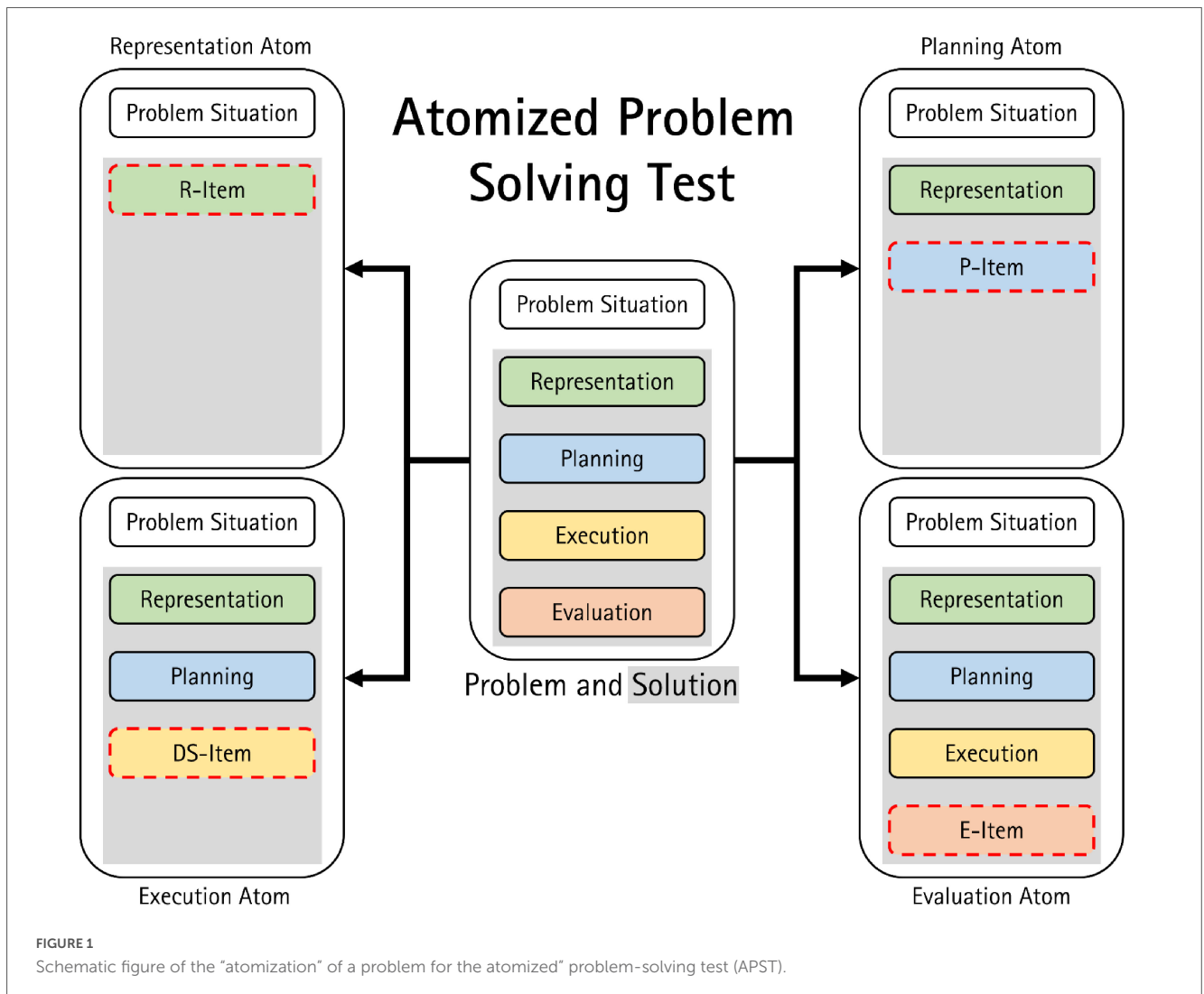
For deeper insights into the cognitive processes of problem solvers, qualitative analyses of written problem solutions are frequently supplemented with verbal data like interviews or think aloud protocols (e.g., Chiu et al., 2022). In interview studies, the participants are asked to explain their solution after they wrote it down, whereas participants in think aloud studies are prompted to verbalize their thoughts during the process (Ericsson and Simon, 1993). These kinds of studies are considered the “gold standard” of problem-solving process analysis (Docktor et al., 2016), but they require a lot of effort, so it is difficult to assess large populations using these methodologies.

Another difficulty of qualitative assessments based on problem solutions is that students might produce incomplete data. For example, a student that fails to devise a plan for the problem solution will not get to the execution or evaluation phases. Therefore, the participant's skills in these subprocesses cannot be analyzed using their written solution. Brandenburger (2016) designed a test instrument that separately assesses the four subprocesses of problem solving. This approach can overcome the abovementioned difficulties of incomplete solutions, because the skills for the subprocesses are assessed independently. However, this approach does not assess the decision making necessary for problem solving, since the participants are prompted step-by-step to execute certain parts of the solution process. Also, the tests were conducted using pen and paper at the university level. There is currently no assessment tool available following this approach, that is applicable for digital learning environments on secondary school level and enables automatic evaluation.

For our project, we aim to develop a digital assessment tool, that separately and independently assesses secondary school students' skills for the different subprocesses of problem solving. We call this an “atomized” approach. The research question for the presented validation study is: *Can the sub-dimensional structure of problem-solving skills be assessed using an atomized test instrument?*

## 2 Materials and methods

A problem-solving test was designed as part of a digital learning environment about physics problem solving concerning the conservation of energy in secondary school lessons. For validation of this test, two consecutive studies were conducted. The first study served as a preliminary study for item design. It was conducted in 13 years 9 through 11 classes from three secondary schools (German “Gymnasium”). In total, 270 students participated of which 138 were males, 127 were females and six students with non-binary gender identity. The mean age was 15.9 (SD 3.2) years. The students answered one of eight test sets with eight items each. A total of 40 different items were examined in this study. The test sets were implemented in a web-based learning environment



and answered using computers or tablets during a regular physics lesson. The results were quantitatively analyzed using a Rasch analysis and further examined qualitatively. The results from these analyses led to several changes that are explained in further detail in section "2.1 Design of the problem-solving tasks."

The second study was conducted in 24 years 9 through 11 classes from 12 secondary schools [German "Gymnasium" and "Gesamtschule (KGS)"] with a total of 530 students (253 males, 252 females, 11 divers, 14 NA). The mean age was 16.3 (SD 0.9) years. In this study, the students answered a personal questionnaire, a problem-solving pretest, and an energy pretest. Then, they did a problem-solving training that is unrelated to this validation study. Thereafter, the students answered a problem-solving posttest and an energy posttest. The posttests were answered by 230 students. All of these instruments were implemented in a web-based learning environment and accessed using computers or tablets. A subsample of participants additionally solved a problem task using pen and paper during the lesson of the pretests or the lesson of the posttests. The students decided voluntarily if they wanted to solve this additional task. In total, 51 students (28 females and 23 males; age: 15.6 SD 1.1) participated in this supplemental assessment.

## 2.1 Design of the problem-solving tasks

### 2.1.1 Initial item design based on theoretical considerations

At first, the items for the atomized problem-solving test (APST) were designed. For this purpose, the relevant official curriculum and commonly used school textbooks were examined. The subject area about quantitative energy considerations is taught between year 9 and 11 of lower saxony's secondary schools. During this teaching unit, students are supposed to learn the formulas for thermal energy, gravitational potential energy and kinetic energy. Besides understanding the interconnections of these formulas, students begin to engage quantitatively with physics problem solving concerning simple mechanics and thermodynamics (Ministry of School and Culture Lower Saxony, 2015). Based on these insights, various problem tasks were designed and discussed with colleagues of our institute. In the end, eight problem tasks were determined.

These eight problems were subsequently "atomized," meaning that each problem task was divided into four sub-tasks concerning the phases of problem-solving processes. Accordingly, for each problem there are tasks for *Representation*, *Planning*, *Execution*,

and *Evaluation*. Figure 1 illustrates the resulting structure of the APST. The complete problem solution is shown in the center of the figure and is divided into the four subprocesses of problem solving. The different APST atoms, displayed on either side of the figure, contain the problem situation together with an increasing number of subprocess solutions. For example, the *Representation* atom includes only the problem situation and the *Representation* item, whereas the *Evaluation* atom incorporates the solutions from the *Representation*, *Planning*, and *Execution* subprocesses in addition to the *Evaluation* item.

For designing these sub-tasks, didactical considerations were balanced against technical constraints. The goal was to design tasks, that reliably evaluate the phases of problem-solving processes, are digitally accessible without requiring special equipment, and can be assessed automatically. Task formats were inspired by the test items of [Brandenburger \(2016\)](#), who used a similar approach for assessing problem solving. But, since her test was used for assessment at the university level and it was conducted using pen and paper, major adjustments were necessary to design a digital test for secondary school level. The resulting task formats are summarized in [Table 1](#).

For the *Representation* phase, students had to select the correct drawing out of four options. Each option accurately depicted the problem situation, but the given and searched quantities were manipulated. A visual representation with a drawing is commonly used in physics problem solving. From a technical perspective, it is difficult to automatically assess students' drawings. For that reason, the multiple choice (MC) item with different drawings was considered a feasible alternative.

**TABLE 1** Item types for the atomized" problem-solving test (APST) tested in study 1 and study 2.

| Item type            | Description   |
|----------------------|---|
| <b>Study 1</b>       |   |
| Representation (R)   | Multiple choice with drawings as options  |
| Planning (P)         | Always two sub-items<br>PX_1: multiple choice with written plans as options<br>PX_2: multiple choice with formulas                                |
| Execution (D)        | Freely written equations with step-by-step instructions   |
| Self-explanation (S) | Complete problem solution that the students need to explain in a short text   |
| Evaluation (E)       | One or two questions about a complete problem solution  |
| <b>Study 2</b>       |   |
| Representation (R)   | Multiple choice with drawings as options  |
| Planning (P)         | Always two sub-items<br>PX_1: multiple choice with written plans as options<br>PX_2: multiple choice with equations as options (only one correct) |
| Execution (DS)       | Problem solutions with omissions that the students fill out with equations  |
| Evaluation (E)       | One or two questions about a complete problem solution  |

The items for the Planning phase already contained a correct drawing as a visual representation of the problem situation. This was done, because the items are supposed to assess precisely one subprocess of problem solving independently from the other subprocesses. With providing a correct drawing as a visual representation, students' success in the Planning items is less influenced by their representation skills. The students had to choose an appropriate approach for solving the problem in two MC questions. The options of the first MC task were short texts explaining approaches. In these questions, the students were supposed to identify the physics concepts and interconnections relevant for the problem. The MC format was used for this question, since formulating an approach with its quantitative manipulations without actually performing them is an unfamiliar task for students. It requires a lot of creativity and mental flexibility that average students might not be capable of ([Tschisgale et al., 2023](#)). The second MC question contained various equations and the students should decide which of them are useful for following the selected approach. The Planning items are designed similarly to the test instrument of [Brandenburger \(2016\)](#).

The Execution phase was evaluated using two task types. In both task types the students were provided with a correct visual representation and a written plan appropriate for solving the problem. In the first type of execution items (D-items), students should follow step-by-step instructions to perform the mathematical operations for solving the problem. This type of item focused on mathematical competencies with the step-by-step instructions providing the plan for the problem solution. Based on the considerations of [Pólya \(1945\)](#), the "carry out the plan" phase is characterized by a finite set of operations that the problem solver knows to be necessary. Therefore, that phase becomes similar to solving a routine exercise. In the other type of items for this phase (S-items), the students were provided with a complete mathematical solution without any comments and were then prompted to explain the given solution. This kind of explanations was derived from self-explanations ([Chi et al., 1989](#)) that are known to be productive in the context of worked examples ([Atkinson et al., 2000](#); [Dudzinska, 2020](#); [Hilbert et al., 2008](#)). The self-explanation items were designed to assess the understanding of a given problem solution.

For the *Evaluation* items, students were provided with a complete problem solution containing a drawing and explaining comments. Below the solution, they were asked questions about the model-like considerations that were used within this solution. For example, the students were asked, how the result would be different under real-world conditions in order to reflect on assumptions that were made for the given solution. As another example, there are questions proposing changes to the problem situation like a hill, that a car runs down, being twice as high. The students were then asked to estimate the new result without calculating, but arguing physically. The students needed to answer these questions in short written texts. The idea of this kind of evaluation task was to assess whether the students are able to identify, understand, and explain certain model-like considerations that are frequently used in this kind of problem solving. This is considered to be the necessary skill for evaluating whether a problem solution is appropriate in a given situation or not.

### 2.1.2 Item revision based on empirical results from the preliminary study

The results of 270 students each answering eight items (two representation tasks, two planning tasks, one D-item, one S-item, and two evaluation tasks) were used for analyses. This was only a preliminary study and there were some issues with the overlap of items in the different test sets. For that reason, the item parameters were simply examined using descriptive statistics like the proportion of correct solutions and a basic one-dimensional Rasch analysis. Additionally, the students' answers were inspected qualitatively guided by item parameters like infit values that were obtained from the Rasch analysis (e.g., Neumann, 2014). As a brief summary, these analyses indicated the following revisions to be appropriate.

The *Representation* items were answered correctly by between 75 % and 90 % of the participants, so they seem to be relatively easy. However, besides that, they seem to adequately fit the test instrument, because the infit values do not indicate problems for ability estimation using these items. The infits of the items are between 0.77 and 1.16 with an area of 0.8–1.2 being acceptable (Bond and Fox, 2007). As a result, the *Representation* items were not revised fundamentally, but the most difficult items were selected and minor changes were made to the quality of the drawings.

The first part of the *Planning* items, where the students select the written approaches, seems to be adequate as well. The ratios of correct solutions for these subitems are between 20 % and 36 % and the infits between 0.80 and 1.09. For the second type of subitem, where the students had to choose equations, it was noticeable that the students often only chose one equation even though multiple equations were correct and useful. Even with student answers being rated as correct, that only chose useful equations but not necessarily all of them, the correct answer ratios are between 14% and 40%. The infits are between 0.79 and 1.25. For that reason, the *Planning* items were slightly amended so that in the second MC question, there are now equations that mathematically represent the selected approach as options. In this amended version, there is always exactly one correct option.

The most substantial changes were made for the items assessing the *Execution* phase. The infits of the D-items with step-by-step instructions were relatively low between 0.69 and 0.92. Also, these items were only answered correctly by between 5% and 29% of the participants, so they seem to be difficult for students. The qualitative analysis while rating the students' answers to these items revealed that the students did not follow the instructions closely enough. This led to difficulties for valid rating, because many students solved the problem, but - strictly speaking - they did not answer the prompts for the separate steps. A similar observation was made for the self-explanations in the S-items. Here, the students rather described the formulas line by line without explaining the physical meanings or plans behind them. Therefore, many student answers were not wrong, but they also did not meet the expectations. As a result, the self-explanation items were excluded completely from the APST. The D-items with the step-by-step instructions were changed to a format similar to an uncomplete worked example (e.g., Atkinson et al., 2000; Hilbert et al., 2008). In these new DS-items assessing the *Execution* phase, the students are provided with a problem description, a correct drawing of

the problem situation, and a commented solution with omitted equations that the students are prompted to add.

The *Evaluation* items appeared to be adequate. With correct answer ratios between 16% and 49% they are rather difficult, but also cover a wide range of difficulties. Infit values between 0.78 and 1.14 indicate no major issues for the Rasch model. The qualitative analysis of students' answers revealed a large variety of correct and incorrect evaluations. Especially the incorrect answers revealed interestingly precise which students had misconceptions about the conservation of energy and assumptions that are frequently made for school-like problem solving. Thus, the *Evaluation* items were not changed fundamentally and the best fitting items were used for the further validation.

In addition to the described improvements of existing items regarding the item types, four new problems were designed and "atomized." Subsequently, two test sets each containing eight items (two per item type) were designed guided by the item parameters from the Rasch analysis. These test sets were used for the second study. The reworked APST was supplemented by the following assessments for further validation.

## 2.2 Qualitative analysis of written problem solutions

The problem-solving skills of a subsample of participants were also assessed using complete written solutions to a physics energy problem. In the problem task, the students were supposed to determine whether a football can be kicked over a fence. For solving the problem, the students needed to use the principle of energy conservation and the formulas for kinetic energy and gravitational potential energy.

The written problem solutions were analyzed by two coders using the Minnesota Assessment of Problem-Solving (MAPS) rubric (Docktor et al., 2016). The MAPS rubric has five different categories: *useful description*, *physics approach*, *specific application of physics*, *mathematical procedures*, and *logical progression*. For all of the categories, a score between zero and five is assigned to the solution. The scores are ordinally scaled with five being the best score indicating expert-like skills regarding the category. Additionally, the coders rated every student's solution as either solved correctly or not solved (correctly) leading to an additional score of zero or one for the problem solution. The two coders discussed cases with differing scores and defined a consensus rating.

For this article, the quantitative correlations between MAPS scores and APST results are analyzed. Further details on problem design and the qualitative analyses of the written solutions are reported in (Meyer et al., 2025b).

## 2.3 Energy test

Conceptual knowledge is known to influence the problem-solving skills for domain-specific problems (e.g., Friege and Lind, 2006). Since the presented assessment tool focuses on quantitative energy problems, the conceptual knowledge about energy was assessed. For this purpose, we used an energy test that has been used in multiple projects of our institution before (e.g., Dudzinska,

2020). It consists of 20 multiple choice items based on the energy concept assessment (ECA) (Neumann et al., 2013). The items cover four conceptions of energy: *forms*, *transformation*, *conservation*, and *degradation*. These conceptions are known to form a learning progression for energy in secondary physics education (Duit, 2014; Neumann et al., 2013). The participants of this study are supposed to have qualitatively learned the *forms* and *transformation* conceptions, and are in the process of learning quantitative aspects of the energy concept like *conservation* and *degradation*.

For this study, we chose five items for each of the four conceptions, ranging from relatively easy to relatively difficult items based on Rasch analyses that were conducted by Neumann et al. (2013) and within our institution. In total, the energy test items are validated using answers from more than 2,000 students.

## 2.4 Personal questionnaire

Because the studies were conducted completely anonymously, the students answered a questionnaire to provide us with personal information. In the questionnaire, the students are asked about their age, which gender they identify with, and if German was their mother tongue. Additionally, they were asked about their school career: what type of school they visit, which year they are in, and what their last grades in physics, mathematics, and German were.

## 2.5 Data analysis

Different procedures were used for the quantitative analysis of the APST items. At first, data screening was done using descriptive statistics. Subsequently, a multidimensional item response theory (MIRT) analysis was applied using the R-library “mirt” (Chalmers, 2012). Item response theory (IRT) is a class of various statistical models that can be used to estimate the probability of a specific response pattern based on a latent trait of the item and the person answering it (Bond and Fox, 2007). In the case of educational assessment, IRT is mostly used to determine the likelihood of a student answering an item correctly based on their skill (latent trait of the person) and the item’s difficulty (latent trait of the item) (Rost, 2004). An advantage of IRT, compared to classical test theory is, that the estimated item parameters are independent from the population they were based on (Bond and Fox, 2007). For that reason, an IRT-validated assessment tool can be utilized to analyze person abilities in various populations. In the case of MIRT, the item parameters and person abilities are estimated on multiple dimensions.

A *common-item equating to a calibrated pool design* was used for the MIRT analyses, meaning that the pretest and posttest results from study two were combined in order to place all items on a common scale (Kolen and Brennan, 2014, pp. 215–219). For that purpose, a first model was calculated using only the common-items that were part of the pretest and the posttest. The item parameters from this model were analyzed using differential item functioning (DIF) to examine if there are significant differences between the calculated item parameters in the pretest group and the posttest group. If there are significant differences for an item, this item is not suitable as common-item, because it is unstable. Afterwards, all of the items are used in a grouped model with the item parameters of

the stable common-items fixed (fixed parameter calibration) (Kolen and Brennan, 2014, pp. 182–183).

Since the basic dimensionality of the instrument was defined by the four phases of problem-solving during the item design, no exploratory factor analysis (EFA) was conducted before the MIRT. A four-dimensional model following the types of items with pairwise covariances was assumed and a 2-parameter logistic (2PL) model (Hambleton and Swaminathan, 1985) was calculated using the quasi-monte-carlo expectation-maximization (QMCEM) algorithm. In a 2PL model, two item parameters are estimated: item difficulty and discrimination. The MIRT model syntax is accessible in the [Supplementary Datasheet 1](#).

After MIRT modeling, a confirmatory factor analysis (CFA) was performed using the “lavaan” package (Rosseel, 2012) in R to examine the factor structure in further detail. The exact model selection can be an important aspect of an instrument’s empirical validation (Immekus et al., 2019). Three different model structures were tested: correlated factors, higher-order, and bifactor. As a commonality, these model structures assume that the four dimensions of the instrument are distinct, yet related. In a correlated factors model, the factors are simply correlated. In a higher-order model, there is one predominant factor (e.g., problem solving) and the four distinct factors (e.g., representation, planning, execution, and evaluation) are subdimensions of this overriding factor. Following this structure, every item would measure one of the four mentioned subprocesses and together these subprocesses are combined to the overall problem-solving process. In a bifactor model, every item loads on one of the four distinct dimensions and additionally on one primary factor. This can be interpreted as every item simultaneously measuring one of the subprocesses and problem solving in general.

For quantitative validation of the presented instrument, the person ability scores from the MIRT analysis were utilized. Person ability scores are a metric scale of a person’s ability for each of the four dimensions. Since the MIRT analysis does not provide an ability score for problem solving in general, the sum of correct items in the APST was used. It was then analyzed whether the four subdimension skills or the total APST scores correlate with the conceptual knowledge about energy, the school grades, and MAPS scores. Correlations were calculated using the “psych” package (Revelle, 2007) in R.

## 3 Results

The proportion of students that answered an item correctly and the distribution of total scores in the APST were used as descriptive statistics. The total scores in the pretest are normally distributed with a mean score of 5.9 (SD 2.9) out of 12 correct items per student. The ratios of correct solutions per item are widely spread between 19% (item DS8) and 81% (item R5). Usually, it is recommended to aim for a ratio between 20% and 80% when designing an assessment tool (Rost, 2004). Items with correct answer ratios below 20% might be too difficult and items that are answered correctly by more than 80% of participants are potentially too easy.

The test sets contained four common-items: R5 for the *Representation* phase, P4 (P4\_1+P4\_2) for the *Planning* phase, DS7 (DS7\_1+DS7\_2) for the *Execution* phase, and E6 for the *Evaluation* phase. These items were used for a MIRT model using grouped

pretest and posttest data. The DIF analysis revealed significant differences between the item parameters in the pre-test group and the post-test group for item DS7 (DS7\_1:  $p < 0.01$ ; DS7\_2:  $p < 0.05$ ), meaning that the item parameters for item DS7 could not be estimated reliably. For that reason, DS7 cannot be considered a stable common item. For the other items, no significant differences for the item parameters were detected ( $p > 0.05$ ). As a result, the items R5, P4, and E6 were fixed as stable common-items for fixed parameter calibration.

Subsequently, a MIRT model including all items was estimated. Infit statistics were used to guide a qualitative analysis of the items and students' answers like in the preliminary study 1. Most items showed acceptable fit; however, item P4 did not conform to the model. Both subitems of P4 exhibited infit values of approximately 0.4.

In MIRT analyses, infit values between 0.5 and 1.5 are generally considered productive for measurement, whereas values below 0.5 indicate limited contribution to the assessment and values above 2.0 are regarded as degrading (Linacre, 2002). The low infit values observed for items P4\_1 and P4\_2 therefore suggest that the items contribute little information to the model.

A qualitative analysis of the corresponding problem task supports this interpretation. The task underlying P4 required students to calculate the energy needed by a crane to lift a weight. In its atomized form, the solution plan consisted solely of computing the gravitational potential energy of the load and adding a given amount of degraded energy. This presents a routine exercise rather than a problem task for the target population. Consequently, the items P4\_1 and P4\_2 did not adequately align with the intended construct of problem-solving subprocesses and were therefore excluded from the final MIRT model.

The item DS7\_1 exhibited an even lower infit value of 0.1, indicating very limited contribution to the assessment. However, the qualitative analysis of this item did not reveal an obvious substantive reason for its lack of productivity. Nevertheless, re-estimating the model without DS7\_1 resulted in a substantial improvement of the global fit indices. This finding indicates that DS7\_1 adversely affected parameter estimation. On this basis, it was classified as a harmful item for measurement and excluded from the APST.

The final MIRT model was calculated using grouped pre-test and post-test data and the common-items R5 and E6. This model converged normally within 0.0001 tolerance after 373 QMCEM iterations. The model fit parameters (log-likelihood =  $-3576.6$ ; AIC = 7241.1; BIC = 7445.0) were the lowest compared to various alternative models that were estimated during the analyses and thus this model is the best fit for the data. Comparing different models in order to analyze the dimensionality is a commonly used procedure in IRT studies (e.g., Wu and Adams, 2006; Zöttl et al., 2011). It is noteworthy that these fit parameters can only be utilized for comparing different model specifications, but they are not useful for absolute argumentations.

### 3.1 Factor structure

The MIRT model was estimated using a four-dimensional model structure based on the four phases of problem-solving that

inspired the item design. Analyses of pairwise covariances between all of the four factors were enabled. The factor loadings of the items to the postulated subdimensions are between 0.47 and 0.92 which can be interpreted as moderate to high factor loadings. The proportional variances of the four subdimensions are between 9% for the *Representation* and 17% for the *Execution* with a total of 55% of variance being explained. The factor correlations are mostly moderate between 0.20 (R~P) and 0.89 (D~E). In summary, the MIRT analyses supported a four-dimensional structure. Four further analyses, confirmatory factor analyses (CFA) were executed like explained in the methods section.

The CFA for the bifactor model did not converge and was therefore found inappropriate for the test structure. The CFA of the correlated factors model [ $\chi^2(98) = 104.8$ ;  $p > 0.3$ ; CFI = 0.99; TLI = 0.99; RMSEA = 0.01] and the higher-order model [ $\chi^2(100) = 106.7$ ;  $p > 0.3$ ; CFI = 0.99; TLI = 0.99; RMSEA = 0.01] converged with good global fit indices (Hu and Bentler, 1999). A Chi-Squared-difference test showed no significant difference in the global fit of these two models [ $\Delta\chi^2(2) = 2.84$ ;  $p > 0.2$ ] and the AIC and BIC parameters are nearly identical. So, in regard to the global fit, the correlated factors model and the higher-order model are equally suitable for the APST subdimensions.

As of the local fit, the higher-order model shows no significant factor loading from the *Evaluation* factor to the higher-order factor *Problem Solving* ( $\beta = 0.02$ ;  $p > 0.9$ ). This indicates that the four subdimensions of the APST do not seem to load on a single-dimensional higher-order factor *Problem Solving*. The *Representation*, *Planning*, and *Execution* items can be summarized by one higher-order factor. However, the *Evaluation* variable cannot be explained by the same higher-order factor. Additionally, the correlation between the higher-order factor *Problem Solving* and the subdimension *Evaluation* is very high ( $\rho = 0.99$ ;  $p < 0.001$ ) and the variance of the *Evaluation* subdimension without *Problem Solving* is less than 0.02. This indicates that these two factors seem to assess nearly the same construct and that *Evaluation* skills cannot be separated from the higher-order factor *Problem Solving* in the APST.

In the correlated factors model, all of the APST dimensions show significant covariances. Especially, the covariance, between *Execution* and *Evaluation* is very strong ( $\rho = 0.8$ ;  $p < 0.001$ ). The covariances of *Representation* and *Planning* as well as of *Planning* and *Execution* are weak ( $0.1 < \rho < 0.3$ ;  $p < 0.05$ ). The other covariances are moderate ( $0.3 < \rho < 0.5$ ;  $p < .01$ ).

### 3.2 Item parameters

As described above, the infit values for every item were used as a guide for qualitative analyses and led to the exclusion of the items P4\_1, P4\_2, and DS7\_1. In the final MIRT model, most of the remaining items showed infit values between 0.5 and 1.0 which is well within the accepted range of 0.5–1.5 (Linacre, 2002). Only the *Planning* item P10\_1 is slightly out of this range with an infit of 0.46. This item might be mildly unproductive, but lower infit values do not indicate degrading items. Since the qualitative analysis of item P10 did not reveal any misfitting content, it was left in the test set. The MIRT analysis generated difficulty and discrimination parameters for each item

**TABLE 2** Item parameters of the multidimensional item response theory (MIRT) analysis: item discriminations for the four subdimensions and item difficulties.

| Item   | Representation | Planning | Execution | Evaluation | Difficulty |
|--------|----------------|----------|-----------|------------|------------|
| R1     | 1.71           |          |           |            | -1.24      |
| R5     | 1.74           |          |           |            | -1.50      |
| R9     | 1.6            |          |           |            | -1.09      |
| P3_1   |                | 2.55     |           |            | 0.01       |
| P3_2   |                | 1.93     |           |            | -0.26      |
| P10_1  |                | 3.07     |           |            | 0.7        |
| P10_2  |                | 2.41     |           |            | 0.71       |
| DS7_2  |                |          | 1.44      |            | -0.08      |
| DS8    |                |          | 1.35      |            | 1.44       |
| DS11_1 |                |          | 1.87      |            | -0.76      |
| DS11_2 |                |          | 2.10      |            | -0.21      |
| E2_1   |                |          |           | 3.11       | 0.12       |
| E2_2   |                |          |           | 0.99       | 0.33       |
| E6     |                |          |           | 0.90       | -0.83      |
| E12_1  |                |          |           | 1.83       | 1.11       |
| E12_2  |                |          |           | 1.46       | -0.37      |

(see Table 2). The discrimination parameters are divided into the four dimensions of the APST, so every type of item discriminates on a different scale. The discrimination values vary between 0.90 and 3.11. The *Representation* and the *Execution* items (DS-items) show more homogeneous discrimination parameters between 1.4 and 2.0, whereas the *Evaluation* items show the greatest variety. The *Planning* items show overall relatively high discriminations between 1.93 and 3.07. This is a considerable variation in the item discrimination.

Item difficulty in IRT analyses is a dimensionless, metric scale that allows to compare the difficulty of items within the same test. Like most of the other parameters, the difficulty parameters cannot be compared through different IRT models. The APST item difficulties vary in a range from -1.50 to 1.44. The *Representation* items are on the easy end of that scale with difficulties between -1.50 and -1.09. The other item types show reasonable variation with some items being on the easier half of the scale and some items on the more difficult half of the scale. *Evaluation* items are the most difficult ones. The detailed results can be found in Table 2.

### 3.3 Person parameters and correlation analyses

The MIRT model also estimated person ability scores for each participant for each of the four subdimensions of the APST. Since it was not possible to implement the higher-order structure into the MIRT analysis, the overall problem-solving skill can only be estimated using the number of correct items like in classical test theory. Correlations of the estimated person ability scores with the different covariates that were assessed during study two were analyzed.

The MAPS rubric was used as a second problem-solving assessment with a subsample of 51 participants. Unfortunately, 23 of these participants did not answer the APST appropriately, so

their results could not be used for analyses. For that reason, only 28 results were used for this correlation analysis, so the results (see Table 3) are rather exploratory. Especially, the *Evaluation* and *Execution* items show significant, moderate to strong correlations with the MAPS scores.

For the analyses of correlations between the energy test and APST as well as the information from the personal questionnaire and the APST, the sample size was  $n = 530$ . This makes the results (see Table 4) more reliable and all of the correlations are highly significant ( $p < 0.01$ ). All of the APST dimensions are moderately correlated with the total score of the energy test and the scores for the *Conservation* items from the energy test.

The correlations between the APST dimensions and the reported school grades for mathematics and physics are weak to moderate and negative. In Germany, school grades range from 1 to 6 with 1 being the best possible grade. So, in that case, the negative correlation is actually a positive correlation, because participants with better grades do also score higher in the APST. The native language of the participants as well as their gender have no correlations to the APST results ( $\rho < 0.1$ ).

## 4 Discussion

In this study, items for a digital and atomized problem-solving assessment, the APST, were developed. The final set of items, including sample solutions and estimated item parameters, is provided in the [Supplementary Datasheet 2](#). The item design is grounded in theoretical models of problem-solving processes and operationalizes four subprocesses that recur across multiple frameworks (e.g., Pólya, 1945; Friege, 2001; Ramalingam et al., 2017).

With respect to physics content, the APST focuses on energy conservation in alignment with the relevant secondary-school curriculum (Ministry of School and Culture Lower Saxony, 2015).

**TABLE 3** Spearman correlation matrix for atomized" problem-solving test (APST) person abilities and Minnesota Assessment of Problem Solving (MAPS) scores.

| <i>n</i> = 28  | Solution | Description | Approach | Physics | Mathematics | Logic | Total |
|----------------|----------|-------------|----------|---------|-------------|-------|-------|
| Representation | −0.08    | 0.01        | −0.11    | −0.17   | 0.03        | −0.19 | −0.16 |
| Planning       | 0.09     | 0.00        | 0.12     | 0.18    | 0.28        | 0.14  | 0.21  |
| Execution      | 0.43*    | 0.17        | 0.26     | 0.16    | 0.44*       | 0.19  | 0.31  |
| Evaluation     | 0.57**   | 0.25        | 0.39*    | 0.16    | 0.44*       | 0.28  | 0.38* |

\*:  $p < 0.05$ ; \*\*:  $p < 0.01$ .

**TABLE 4** Spearman correlation matrix for atomized" problem-solving test (APST) person abilities and energy test scores.

| <i>n</i> = 530 | Forms  | Transformation | Conservation | Degradation | Total  |
|----------------|--------|----------------|--------------|-------------|--------|
| Representation | 0.30** | 0.33**         | 0.44**       | 0.27**      | 0.47** |
| Planning       | 0.26** | 0.27**         | 0.32**       | 0.22**      | 0.37** |
| Execution      | 0.25** | 0.25**         | 0.39**       | 0.26**      | 0.41** |
| Evaluation     | 0.31** | 0.33**         | 0.43**       | 0.30**      | 0.49** |

\*\*: $p < 0.01$ .

As participants in the present study were not necessarily instructed in quantitative problem solving across other physics domains, like forces, the instrument was deliberately restricted to this content area. In addition, the APST builds on established design principles of an existing test instrument for problem-solving skills (Brandenburger, 2016). Together, these arguments support the content validity of the APST.

To examine criterion validity, conceptual knowledge about energy and problem-solving skills were assessed using validated test instruments. In addition, school grades in mathematics and physics were collected as indicators of general skills concerning these subjects. Problem solving is known to be influenced by conceptual knowledge and the ability to apply this knowledge to unfamiliar situations (Tuminaro and Redish, 2007; OECD, 2023). Accordingly, conceptual knowledge can be expected to predict problem-solving performance (Friege and Lind, 2006).

The analyses revealed significant correlations between all APST subdimensions and all conceptions assessed in the energy test. In particular, the *Conservation* conception showed stronger associations with APST results than other conceptions of energy. This finding is theoretically plausible, as the APST was explicitly designed for a learning environment that emphasizes the quantitative application of the energy conservation principle.

Moreover, the *Representation* and *Evaluation* subdimensions show stronger correlations with conceptual knowledge than *Planning* and *Execution*. In the *Representation* phase, students are required to identify relevant physics concepts and principles within a problem context, which presupposes well-developed conceptual understanding. Similarly, evaluating a complete solution necessitates situating a model-like solution within broader physical considerations, again drawing on conceptual knowledge. In contrast, the *Planning* and *Execution* phases rely more strongly on procedural skills and problem schemes (Friege, 2001). *Planning* additionally requires abstract thinking and imagination, as students must anticipate a solution path without yet performing it (Tschisgale et al., 2023), whereas *Execution* is characterized by the mathematical processing of planned steps (Pólya, 1945). It is therefore expected that these phases depend less strongly on conceptual knowledge than *Representation* and *Evaluation*.

The MAPS rubric (Docktor et al., 2016) was employed as an external measure of problem-solving skills to explore its relationship with APST scores. However, the corresponding correlation analyses are substantially limited by the small number of participants available due to artifacts in the data collection process. Within these limitations, exploratory analyses indicate that MAPS scores, primarily the mathematics rubric and the problem-solving success, show tentative associations with the *Execution* and *Evaluation* subdimensions of the APST. In contrast, no statistically significant associations were observed for the *Planning* and *Representation* subdimensions.

For the *Representation* subdimension, a negative trend can be observed for the association with MAPS scores. This finding should be interpreted with particular caution. One possible explanation is that the R-items of the APST are comparatively easy and therefore answered correctly by a large proportion of participants, which may restrict variance. Alternatively, this trend may reflect differences in the assessment approaches of APST and MAPS rather than substantive differences in representation skills. In the MAPS assessment, students solve problems without explicit guidance in order to minimize interference with the problem-solving process (Docktor et al., 2016). As a result, students are not required to explicitly perform the representation phase. In cases where no drawing or other representation is produced but the solution process is otherwise correct, the highest MAPS score is still assigned. Consequently, high MAPS scores may occur even when representation skills are not directly assessed. In contrast, the APST explicitly requires all participants to engage in the representation phase as a core component of the assessment.

Given the limited statistical power of the analyses, the MAPS-based findings should be interpreted as preliminary rather than confirmatory. Further research is needed to adequately examine the associations between MAPS and APST.

School grades in mathematics and physics are correlated significantly with the APST scores, indicating that students with stronger overall skills tend to perform better in the APST. In contrast, no significant correlations were observed with grades in German, native language, or gender. The absence of these correlations suggests that no systematic language-related or gender-related bias was observed in the APST results.

The MIRT analyses provided support for the four-dimensional structure assumed in the test design. Model fit is acceptable, since the algorithms converged within the usual QMCEM tolerance of 0.0001 and the fit parameters are within acceptable ranges. Moderate to high factor loadings among items within each subdimension indicate that the items consistently assess their intended constructs. Together with the correlations between the four subdimensions, these findings support the assumption that the problem-solving subprocesses require distinct yet interrelated skills.

The CFA results further support a correlated factors structure of the APST items, indicating empirically related yet distinguishable dimensions. Although a higher-order factor can be statistically specified, its interpretation requires caution. In particular, the higher-order factor shows substantial redundancy with the *Evaluation* subdimension as indicated by the near-zero loading of *Evaluation* on the higher-order factor, the extremely high correlation between those two factors, and the resulting minimal residual variance. These findings suggest that the APST subdimensions do not constitute a general problem-solving construct.

From a conceptual perspective, general problem solving inherently involves decision making, such as selecting strategies, representations, and solutions (Mosier et al., 2018; Price et al., 2022). By atomizing the solution process and guiding participants step-by-step, the APST excludes such decision making from the assessment. Accordingly, the results from the higher-order factor model suggest that problem solving is more than the sum of its subprocesses.

On the other hand, these findings also highlight the special role of *Evaluation* within the problem-solving process. Evaluating a given solution requires understanding the problem situation, comprehending the underlying plan, and judging the appropriateness of the execution, in addition to evaluation-specific competencies. This interpretation is further supported by the observed correlations between the APST subdimensions, with *Evaluation* showing moderate associations with *Representation* ( $r = 0.31$ ) and *Planning* ( $r = 0.36$ ), and a strong association with *Execution* ( $r = 0.66$ ). Given the focus on quantitative problem solving, the strong link between *Execution* and *Evaluation* is theoretically plausible, as procedural and mathematical skills play a central role in assessing solution correctness.

Taken together, these findings underscore the conceptual distinction between assessing subprocess skills and measuring problem-solving competence in a broader sense. While the APST provides detailed information about students' proficiency in specific subprocess skills, the present results should not be interpreted as evidence for assessment of a general problem-solving factor. Further research is required to examine how the subprocess skills, particularly *Evaluation*, relate to general problem-solving literacy.

The item parameters of the MIRT analysis indicate that the APST items span a broad range of difficulty levels. An exception are the *Representation* items, which are consistently easy and exhibit ceiling effects, with up to 81% of correct responses. As a consequence, these items show limited sensitivity to individual differences in representation skills and contribute less to discrimination at higher ability levels. Understanding and representing the problem situation is often a comparatively straightforward subprocess that expert-like problem solvers tend to

perform implicitly rather than explicitly (Friege, 2001). In addition, the APST in its current form does not require participants to independently construct a representation, as the multiple-choice format already provides response options. Both aspects likely reduce task difficulty and restrict variance in item responses.

Despite these limitations, the *Representation* items were retained in the current instrument because they operationalize a theoretically central phase of the problem-solving process and ensure coverage of all four subprocesses assumed in the underlying framework, while allowing for automated assessment. At the same time, the observed ceiling effects indicate that the present items are not optimal for assessing higher levels of representation skills. More complex problem situations or alternative task formats, such as prompting students to generate their own representations, may be necessary to increase sensitivity. Future technical developments may enable automatic assessment of such alternative task formats, for example through artificial intelligence (AI)-based analyses of student generated drawings (Lee and Zhai, 2025).

In conclusion, the presented study provides evidence that the APST can assess four theoretically grounded subprocesses of problem solving as described in established models (Pólya, 1945; Friege, 2001). By focusing on energy conservation problems, the APST is particularly suited for use in secondary physics education, where quantitative problem-solving skills are still developing. However, to what extent the assessed subprocess skills predict more general problem-solving literacy remains an open question, and the generalizability of the instrument to other physics concepts requires further investigation.

A key strength of the APST in its current form is its suitability for automatic assessment across all item types, e.g., using open-source AI for automatic feedback on the *Evaluation* items (Meyer et al., 2025a). This enables large scale summative assessments as well as formative assessments in everyday school teaching. However, similar to the test instrument proposed by Brandenburger (2016), the APST does not require active decision making. Consequently, while it may inform students about their skills for different subprocesses of problem solving, the APST is not sufficient on its own to assess general problem-solving literacy. A combined use of atomized instruments such as the APST and assessments based on complete problem solutions, ideally supplemented by verbal data, may therefore offer a more comprehensive approach to evaluating problem-solving processes.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The studies involving humans were approved by Regionales Landesamt für Schule und Bildung (RLSB) Hannover [Approval-Nr. H1R.10-81402-(103/2024) and H1R.10-81402-(116/2023)]. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

## Author contributions

AM: Resources, Validation, Conceptualization, Writing – review & editing, Data curation, Methodology, Writing – original draft, Formal analysis, Software, Funding acquisition, Visualization, Investigation, Project administration. GF: Supervision, Methodology, Writing – review & editing, Conceptualization, Resources, Project administration, Validation, Visualization.

## Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported within the Project “LernMINT - data-driven teaching in STEM subjects” by the Ministry of Science and Culture Lower Saxony (grant number 51410078) and by the German Academic Scholarship Foundation (Doctoral Scholarship/No grant number). The publication of this article was funded by the Open Access Fund of Leibniz Universität Hannover.

## Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

- Atkinson, R. K., Derry, S. J., Renkl, A., and Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Rev. Educ. Res.* 70:181. doi: 10.2307/1170661
- Bond, T. G., and Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Brandenburger, M. (2016). *Was Beeinflusst den Erfolg Beim Problemlösen in der Physik? Eine Untersuchung mit Studierenden. [What influences success in problem-solving in physics? A study with students]*. Berlin: Logos Verlag Berlin GmbH. German
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *J. Stat. Soft.* 48, 1–29. doi: 10.18637/jss.v048.i06
- Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., and Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cogn. Sci.* 13, 145–182. doi: 10.1207/s15516709cog1302\_1
- Chiu, B., Randles, C., and Irby, S. (2022). Analyzing student problem-solving with MATch. *Front. Educ.* 6:769042. doi: 10.3389/feduc.2021.769042
- Csapó, B., and Funke, J. eds (2017). *The Nature of Problem Solving*. Paris: OECD Publishing.
- Docktor, J. L., Dornfeld, J., Frodermann, E., Heller, K., Hsu, L., Jackson, K. A., et al. (2016). Assessing student written problem solutions: A problem-solving rubric with application to introductory physics. *Phys. Rev. Phys. Educ. Res.* 12:010130. doi: 10.1103/PhysRevPhysEducRes.12.010130
- Docktor, J. L., Strand, N. E., Mestre, J. P., and Ross, B. H. (2015). Conceptual problem solving in high school physics. *Phys. Rev. ST Phys. Educ. Res.* 11:020106. doi: 10.1103/PhysRevSTPER.11.020106
- Dörner, D., and Funke, J. (2017). Complex problem solving: What it is and what it is not. *Front. Psychol.* 8:1153. doi: 10.3389/fpsyg.2017.01153
- Dudzinska, M. (2020). *Lernen mit Beispielaufgaben und Feedback im Physikunterricht der Sekundarstufe I: Energieerhaltung zur Lösung von Aufgaben nutzen. [Learning with example problems and feedback in secondary school physics*

## Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. ChatGPT and DeepL were used as generative AI tools for language editing of the submitted article in order to improve readability and clearness. The final text was checked for factual accuracy.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/feduc.2026.1759878/full#supplementary-material>

- lessons: Using conservation of energy to solve problems*. Berlin: LOGOS Verlag. German
- Duit, R. (2014). “Teaching and learning the physics energy concept,” in *Teaching and Learning of Energy in K – 12 Education*, eds R. F. Chen, A. Eisenkraft, D. Fortus, J. Krajcik, K. Neumann, J. Nordine, et al. (Cham: Springer International Publishing), 67–85.
- Ericsson, K. A., and Simon, H. A. (1993). *Protocol Analysis*. Cambridge, MA: The MIT Press.
- Frenzel, A. C., Pekrun, R., Dicke, A.-L., and Goetz, T. (2012). Beyond quantitative decline: Conceptual shifts in adolescents' development of interest in mathematics. *Dev. Psychol.* 48, 1069–1082. doi: 10.1037/a0026895
- Frieg, G. (2001). *Wissen und Problemlösen: Eine empirische Untersuchung des wissenszentrierten Problemlösens im Gebiet der Elektrizitätslehre auf der Grundlage des Experten-Novizen-Vergleichs. [Knowledge and problem solving: An empirical investigation of knowledge-centered problem solving in the field of electricity based on expert-novice comparison.]*. Berlin: LOGOS Verlag. German
- Frieg, G., and Lind, G. (2006). Types and qualities of knowledge and their relations to problem solving in physics. *Int. J. Sci. Math. Educ.* 4, 437–465. doi: 10.1007/s10763-005-9013-8
- Hambleton, R. K., and Swaminathan, H. (1985). *Item Response Theory*. Dordrecht: Springer Netherlands.
- Hilbert, T. S., Renkl, A., Schworm, S., Kessler, S., and Reiss, K. (2008). Learning to teach with worked-out examples: A computer-based learning environment for teachers. *J. Comp. Assis. Learn.* 24, 316–332. doi: 10.1111/j.1365-2729.2007.00266.x
- Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equat. Model. A Multidisc. J.* 6, 1–55. doi: 10.1080/10705519909540118
- Immekus, J. C., Snyder, K. E., and Ralston, P. A. (2019). Multidimensional item response theory for factor structure assessment in educational psychology research. *Front. Educ.* 4:45. doi: 10.3389/feduc.2019.00045

- Jang, H. (2016). Identifying 21st Century STEM competencies using workplace data. *J. Sci. Educ. Technol.* 25, 284–301. doi: 10.1007/s10956-015-9593-1
- Kelly, R., McLoughlin, E., and Finlayson, O. E. (2016). Analysing student written solutions to investigate if problem-solving processes are evident throughout. *Intern. J. Sci. Educ.* 38, 1766–1784. doi: 10.1080/09500693.2016.1214766
- Kolen, M. J., and Brennan, R. L. (2014). *Test Equating, Scaling, and Linking*. New York, NY: Springer.
- Lee, G., and Zhai, X. (2025). Realizing visual question answering for education: GPT-4V as a multimodal AI. *TechTrends* 69, 271–287. doi: 10.1007/s11528-024-01035-z
- Lee, G.-G., Latif, E., Wu, X., Liu, N., and Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Comp. Educ. Art. Intell.* 6:100213. doi: 10.1016/j.caeai.2024.100213
- Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Trans.* 16:878.
- Martinez, M. E. (1998). What is problem solving? *Phi Delta Kappan* 79, 605–609.
- Meyer, A., Bleckmann, T., and Friege, G. (2025a). Automatic feedback on physics tasks using open-source generative artificial intelligence. *Intern. J. Sci. Educ.* 1–26. doi: 10.1080/09500693.2025.2499220 [Epub ahead of print].
- Meyer, A., Fischer, S., and Friege, G. (2025b). Analyzing problem-solving in secondary physics education: A rubric-guided approach to explore individual learning needs. *Phys. Rev. Phys. Educ. Res.* 22:010111. doi: 10.1103/3bs7-fnrd
- Ministry of School and Culture Lower Saxony (2015). *Kerncurriculum für das Gymnasium Schuljahrgänge 5-10 Naturwissenschaften*. [Core curriculum for secondary schools, grades 5-10, natural sciences]. German: Ministry of School and Culture Lower Saxony. German
- Mosier, K., Fischer, U., Hoffman, R. R., and Klein, G. (2018). “Expert professional judgments and “naturalistic decision making,” in *The Cambridge Handbook of Expertise and Expert Performance*, eds K. A. Ericsson, R. R. Hoffman, A. Kozbelt, and A. M. Williams (Cambridge, MA: Cambridge University Press), 453–475.
- Neumann, K. (2014). “Rasch-Analyse naturwissenschaftsbezogener Leistungstests,” in *Methoden in der Naturwissenschaftsdidaktischen Forschung*, eds D. Krüger, I. Parchmann, and H. Schecker (Berlin: Springer), 355–369.
- Neumann, K., Viering, T., Boone, W. J., and Fischer, H. E. (2013). Towards a learning progression of energy. *J. Res. Sci. Teach.* 50, 162–188. doi: 10.1002/tea.21061
- Nilsen, T., Angell, C., and Gronmo, L. S. (2013). Mathematical competencies and the role of mathematics in physics education: A trend analysis of TIMSS Advanced 1995 and 2008. *ADNO* 7, 1–21. doi: 10.5617/adno.1113
- OECD (2013). *PISA 2012 Assessment and Analytical Framework*. Paris: OECD Publishing.
- OECD (2023). *PISA 2022 Results*. Paris: OECD Publishing.
- Plass, J. L., and Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *J. Res. Technol. Educ.* 52, 275–300. doi: 10.1080/15391523.2020.1719943
- Pólya, G. (1945). *How to Solve it*. Princeton, NJ: Princeton University Press.
- Potvin, P., and Hasni, A. (2014). 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, 85–129. doi: 10.1080/03057267.2014.881626
- Price, A., Salehi, S., Burkholder, E., Kim, C., Isava, V., Flynn, M., et al. (2022). An accurate and practical method for assessing science and engineering problem-solving expertise. *Intern. J. Sci. Educ.* 44, 2061–2084. doi: 10.1080/09500693.2022.2111668
- Ramalingam, D., Philpot, R., and McCrae, B. (2017). “The PISA 2012 assessment of problem solving,” in *The Nature of Problem Solving*, eds B. Csapó and J. Funke (Paris: OECD Publishing), 75–91.
- Revelle, W. (2007). *psych: Procedures for Psychological, Psychometric, and Personality Research. R package version 2.5.6*. Northwestern University, Evanston, IL. Available online at: <https://CRAN.R-project.org/package=psych>
- Rossee, Y. (2012). lavaan: An R package for structural equation modeling. *J. Stat. Soft.* 48, 1–36. doi: 10.18637/jss.v048.i02
- Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion. [Textbook Test Theory - Test Construction]*. Bern: Verlag Hans Huber. German
- Smith, M. U. (1991). *Toward a Unified Theory of Problem Solving: Views from the Content Domains*. Hillsdale, NJ: L. Erlbaum Associates.
- Tong, T., Pi, F., Zheng, S., Zhong, Y., Lin, X., and Wei, Y. (2025). Exploring the effect of mathematics skills on student performance in physics problem-solving: A structural equation modeling analysis. *Res. Sci. Educ.* 55, 489–509. doi: 10.1007/s11165-024-10201-5
- Träff, U., Olsson, L., Skagerlund, K., Skagenholt, M., and Östergren, R. (2019). Logical reasoning, spatial processing, and verbal working memory: Longitudinal predictors of physics achievement at Age 12-13 Years. *Front. Psychol.* 10:1929. doi: 10.3389/fpsyg.2019.01929
- Tschisgale, P., Kubsch, M., Wulff, P., Petersen, S., and Neumann, K. (2025). Exploring the sequential structure of students’ physics problem-solving approaches using process mining and sequence analysis. *Phys. Rev. Phys. Educ. Res.* 21:010111. doi: 10.1103/PhysRevPhysEducRes.21.010111
- Tschisgale, P., Wulff, P., and Kubsch, M. (2023). 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. doi: 10.1103/PhysRevPhysEducRes.19.020123
- Tuminaro, J., and Redish, E. F. (2007). Elements of a cognitive model of physics problem solving: Epistemic games. *Phys. Rev. ST Phys. Educ. Res.* 3:020101. doi: 10.1103/PhysRevSTPER.3.020101
- Wu, M., and Adams, R. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Math. Ed. Res. J.* 18, 93–113. doi: 10.1007/BF03217438
- Zöttl, L., Ufer, S., and Reiss, K. (2011). “Assessing modelling competencies using a multidimensional IRT approach,” in *Trends in Teaching and Learning of Mathematical Modelling*, eds G. Kaiser, W. Blum, R. Borromeo Ferri, and G. Stillman (Dordrecht: Springer), 427–437.