Journal of Research in Science, Mathematics and Technology Education

https://jrsmte.com/

# Difficulty-Generating Features of Text-based Physics Tasks

**Knut Schwippert, Kendra Zilz and Dietmar Höttecke**

University Hamburg, Germany

**Abstract:** Influencing the difficulty of performance tasks is of great interest in science education as in several other subjects. In the context of the VAMPS project, difficulty-generating features with respect to the cognitive demand of text-based physics tasks were systematically varied at three levels. Based on preliminary work and two pilot studies presented here briefly, a model was developed by which cognitive requirement was varied according to three features. The viability of this model was empirically tested with a sample of n = 414 secondary school students. The feature *cognitive activity* proved to be a significant factor influencing the empirically measured difficulty of tasks. With the help of the feature *number of information obtained from task stem* and *number of subject-specific mental procedures*, no systematic influence on task difficulty could be shown. The influence of the test persons' individual prior knowledge on the actual task difficulty is generally assumed to be a confounding factor. Overall, the present study contributes to a better understanding of construct representation in assessments of subject-specific proficiency and empirically confirms that a systematic variation of the task feature *cognitive activity* on three levels affects task difficulty.

**Keywords:** *Test items, Item difficulty, Teaching and learning electricity*
**DOI:** https://doi.org/10.31756/jrsmte.721

## Introduction

Design and development of performance tasks in science education like in many other subjects are crucial for diagnosing and assessing students' subject-specific competencies and providing learning-supportive feedback. Successfully solving written performance tasks requires both subject-specific cognitive as well as linguistic skills. The latter are particularly crucial for developing a mental model of the context presented in the task as well as activating situation-specific prior knowledge (Schnotz, 2006; Schmidt-Barkow, 2010). In addition, individual prerequisites such as general knowledge in a subject as well as familiarity with and interest in the content covered also influence the probability of solving the performance tasks. There is no coherent research evidence on how linguistic complexity of performance tasks influence their difficulty and how they interact with subject-specific cognitive requirements. Therefore, one of the goals of VAMPS (German acronym: Variation of tasks - mathematics, physics, language) is to investigate the impact of subject-specific cognitive requirements (SCR) and linguistic complexity of physics performance tasks on empirical task difficulty in an experimental design.

In two distinct sub-projects of VAMPS, performance tasks were developed in physics and mathematics. Each task comprises:

a) one extensive task stem followed by
b) one question or instruction followed by
c) one multiple-choice-single-select item (1 correct answer + 4 distractors).

(a), (b) and (c) together are referred to as a task. Ultimately, these tasks aim to allow the analysis of the role of subject-specific cognitive and linguistic skills in the completion of performance tasks and their respective

contributions to generating empirical difficulty. Assuming that the effect of linguistic variation is more likely to be evident when the stem text is longer, they are comparatively long compared to other studies. Linguistic variation in VAMPS is limited to the task stems (a), but not to (b) or (c).

To systematically distinguish the impact of linguistic variation in task stems from other sources of variance, VAMPS employs a rigorous approach. During the development of tasks, the project deliberately separates linguistic variation in task stems (a) from subject-specific cognitive requirements in questions, instructions (b), and items (c). The comprehensive integration and analysis of the difficulty-generating effects arising from both subject-specific cognitive and linguistic features will occur in a final study of VAMPS.

To achieve this aim, task development in VAMPS is directed by two models, one addressing linguistic complexity while the other focuses subject-specific cognitive requirements. The model delineating linguistic complexity was extensively explained and justified (Heine et al., 2018). This model systematically varies multiple linguistic surface features of a text across three distinct levels. This variation contributes to decreasing transparency and accessibility of a text across the three levels. To put it briefly, the higher the level of linguistic surface features of a text, the more difficult it is to read and understand.

This article reports design and results of the "SCR Study," an exploration into the isolated variation of subject-specific cognitive requirements in physics. Additionally, we share insights from two pilot studies. Results concerning the variation of linguistic complexity will be reported in a separate paper. The SCR study presented in this paper contributes to a better understanding of construct representation in assessments. In the overall context of the VAMPS project, it aims to vary subject-specific cognitive requirements in a model-based and empirically supported manner. Items that meet this requirement will be used in the final project study, briefly mentioned here as a preview, to reintegrate the variation of linguistic complexity and subject-specific cognitive requirements.

**Factors Influencing Difficulty of Tasks in Assessments**

To provide a comprehensive context for the approach and findings of the SCR study, we delve into contemporary research to elucidate whether and how the attributes of performance tasks contribute to empirical difficulty.
Until today, numerous task features have been investigated for their potential to predict empirical difficulty or clarify the variance of item parameters. The analysis of difficulty-generating features is considered highly significant for test development in general (Walpuski & Ropohl, 2014). The difficulty of tasks can be attributed to surface features and the process quality of their execution, along with associated cognitions. When students work on a task, task features and person features interact (Prenzel et al., 2002; Kauertz, 2008). However, as *difficulty* generally depends on prior knowledge and experience, Prediger and Aufschnaiter (2023) conclude that research should consider it as an intersubjectively valid construct without taking familiarity into account. Concerning formal task features, it is known that context effects such as position effects (Nagy et al., 2017) influence task difficulty, and open response formats increase task difficulty compared to closed tasks (e.g., MC-items) (Gut-Glanzmann, 2012; Le

Hebel et al., 2017; Prenzel et al., 2002; Härtig, Heitmann & Retelsdorf, 2015; Mesic & Muratovic, 2011). For the construction of items in competency tests, it has been recommended to include essential information required for accurate responses in the task stem. This approach aims to refine the measurement and focus less on subject-specific knowledge and more on its application (e.g. Kauertz, 2008). The comparison of chemistry tasks with and without such subject-specific information presentations shows differences in the factor "Chemistry knowledge" for predicting person parameters in regression models, with an increasing influence of subject-specific knowledge for tasks without information presentation as expected (Ropohl, Walpuski & Sumfleth, 2015). The plausibility of distractors also influences task difficulty (Hartig & Frey 2012). Task difficulty can be reduced by providing visual information (e.g., Gut-Glanzmann, 2012; Prenzel et al., 2002). However, highly complex illustrations (Solano-Flores & Wang, 2015) or decorative pictures (Carney & Levin, 2002) can also increase task difficulty, while the presentation of tables have been shown to reduce difficulty (Stiller et al., 2016).

Regarding content-related task features, difficulties can arise depending on the subject concerning domain-specific cognitive processes and knowledge (Prenzel et al., 2002). For example, in mathematics, the choice of a specific field of knowledge (e.g., geometry) or the type of mathematical work (e.g., technical task) can lead to gender-related item difficulty (Knoche & Lind, 2004). In science education, the hierarchical complexity of subject knowledge – ranked from everyday knowledge, facts, processes, linear causality to multivariate interdependence – can explain a significant amount of variance (e.g., $R^2 = 0.54$ to 0.57 for different content areas of chemistry; Bernholt & Parchman, 2011). The construction of complexity levels is cognitively informed and empirically ordered in terms of task difficulty (Bernholt & Parchman, 2011). Depending on the modeling of *complexity* of tasks, research has so far led to a variety of results. Kauertz (2008) models task complexity as a sequence of six levels from using a fact, multiple facts, a single connection, multiple unrelated connections, multiple connected connections, to an overarching concept. The levels do not consistently align, but analysis of correlations showed a weak relationship with item parameters ($\rho = 0.38$). Kauertz (2008, p. 101) explains that *complexity* accounts for 23% of the variance in task difficulty, while *cognitive activities* varied across four levels (reproduce, select, organize, integrate) fail to reach significance. Kauertz attributes his finding to the possibility that curricular content as represented in the test compared to content learned earlier may differ significantly (Kauertz, 2008, p. 119) which again points to the significant role of prior knowledge for the prediction of person parameters. In his work, the content-related *guiding idea* explains 17% of the variance in task difficulty; the overall model explains 30%. Kauertz's (2008) complexity levels were aggregated in subsequent studies (Neumann et al., 2013). In a study on modeling scientific communication competence (Ziepprecht et al., 2017), *cognitive activities* align with *complexity* as expected. Neumann et al. (2010) model *complexity* as the amount of information and how it is interlinked, presented at the lower level in the task stem but to be inferred at a higher level. In this way, task difficulty could not be systematically varied across *complexity*. In Neumann's (2011) investigation into the Nature of Science a consistent pattern in task difficulty is noted for *complexity*. However, in the case of *cognitive activities* modeled across four levels, aligning with Kauertz's (2008) competence model, only a tenuous correlation with achievement was identified when some levels were aggregated. For tasks in the context of experimental scientific work, item

parameters correlate more strongly with *complexity* (r = 0.69) than with *cognitive processes* (r = 0.36) (Mannel et al. 2009). In analyses of high school biology exam tasks, six ascending integrated combinations of cognitive requirements – from accessing information, integrating and determining subject knowledge to using information, expanding subject knowledge, and arguing – explain 41% of the variance in task difficulty.

Kulgemeyer und Schecker (2009) show with a structural model for physics communication that a cognitive contribution explains task difficulty distinctly from other factors. The contribution describes the sum of cognitive steps to be processed for task solution, without considering their quality. Wellnitz et al. (2012) were able to demonstrate that the amount of processed information and its degree of connection across the three subjects of biology, chemistry, and physics have a difficulty-generating effect. The study shows medium to strong effects of *cognitive processes* on item difficulty. However, *complexity* and *cognitive processes* were not operationalized in a consistent manner across various studies.

In a competence model for the energy concept (Neumann, Viering & Fischer, 2010), different subject-related sub-concepts of energy generate task difficulty in an expected sequence from "forms of energy" ascending to "energy conservation." In general, the use of abstract concepts and the consideration of technical terms contribute to task difficulty (Prenzel et al. 2002; Stiller et al. 2016).

As task stems need to be read, reading proficiency is usually assessed and shows the significant role of prior knowledge for reading comprehension (Kintsch & van Dijk, 1978). Härtig et al. (2022) demonstrate, regarding the so-called DIME model (Direct and Inferential Mediation Model of Reading Comprehension), that compared to the role of vocabulary used in a task, prior knowledge is more strongly associated with text comprehension. This can also be interpreted as an indication that content-specific task features influence their difficulty. Similarly, Jaeger and Müller (2019) show that linguistically mediated difficulty (varied through readability index) in solving physics tasks has only marginal significance mediated through cognitive load. Experimental studies systematically examining linguistic properties of test items with regard to difficulty show inconsistent findings (Kieffer et al., 2009; Höttecke, Feser, Heine & Ehmke, 2018). Hackemann (2023) concludes in his recent research overview that only small effects of linguistic complexity on text comprehension and students' task performance can be identified. However, Cruz Neri, Guill and Retelsdorf (2021) show that linguistic properties of items interact with reading proficiency in a way that individuals with higher reading proficiency can benefit from extended texts. Overall, the empirical evidence regarding the role of linguistic features in predicting item difficulty is not consistent.

**Students Background and Attitudes Related to Subject-specific Performance**

In several large-scale assessments, it is documented that students' performance in reading, mathematics, and science tests are related to their personal attributes. Next to gender effects, student attitudes like interest in the subject as well as the general grade level of the students show effects on performance in assessments. On average, girls in international assessments perform better in reading, whereas boys outperform girls in mathematics and science.

Students from higher grades outperform students from lower grades. And in general, the more students are interested in a subject the better they perform (OECD, 2021, 2023; Mullis et al., 2020). Furthermore, it is proven, that general knowledge in a subject field is a good predictor to understand specific aspects within the field (Kunter, Baumert & Köller, 2007).

The current state of research results in the following summary: The impact of various formal item features as well as content aspects on a task's difficulty can be considered established. This is not the case for linguistic features, although an influence should not be entirely ruled out at present. Theoretical findings from different studies do not provide coherent evidence on the role of *complexity*, and *cognitive activities* for the difficulty of tasks which will both be explored in this paper. However, research aimed at establishing an empirical foundation for requirement levels in assessments is still in an early stage across various subject didactics (Prediger & Aufschnaiter, 2023).

**Research Question and Model for Variation of Cognitive Requirements in this Study**

Given the contradictory findings highlighted in the research review on the influence of *cognitive activities* and task *complexity* on the prediction of task difficulty, the SCR study as outlined in this paper addresses the research question of whether, and if so to what extent, *cognitive activities* and task *complexity* contribute to task requirement (a task appears to be more or less demanding) and can explain the variance in task difficulty. The main research interests of this study are:

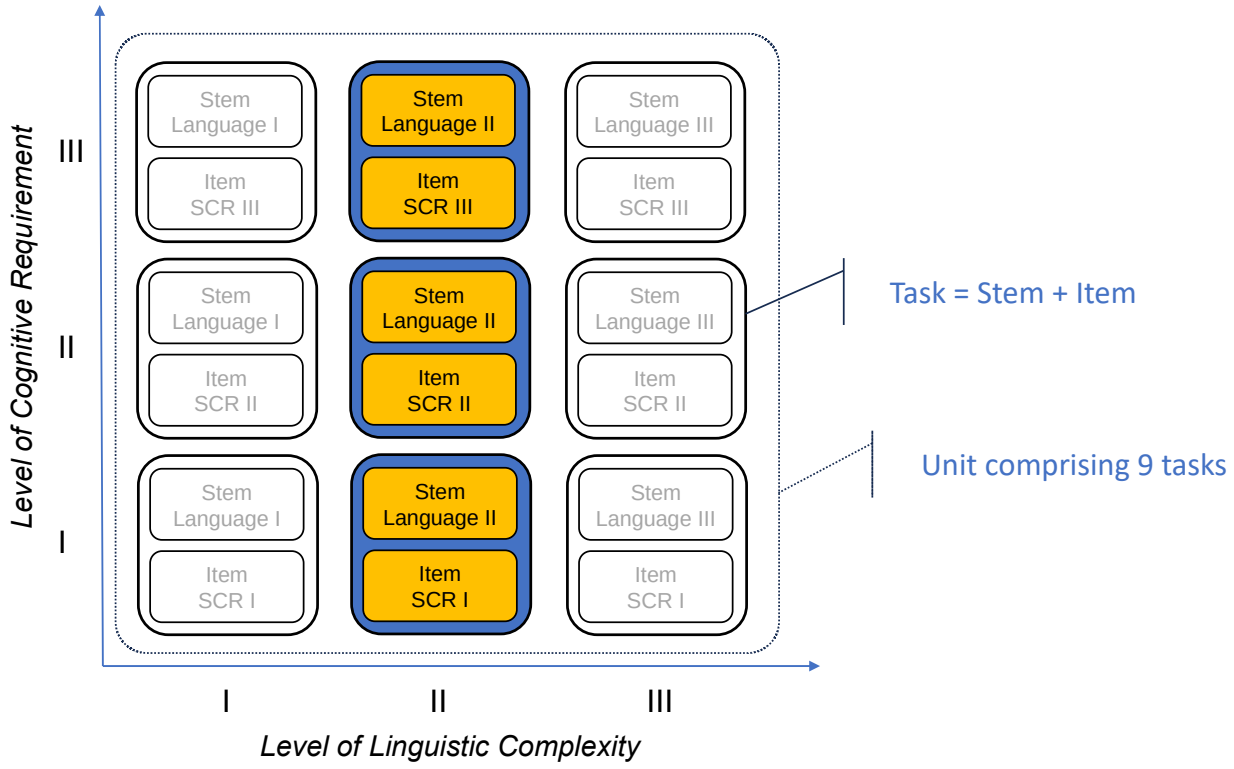
(1) Is it possible to create a one-dimensional Physics test for a specific content area with items that focus on the same item stem by a systematic variation of the *level of cognitive requirement*?
(2) To what extent does *cognitive requirement* modelled as cognitive activity and taks *complexity* influence the difficulty to solve the physics tasks?
(3) How closely are students' characteristics linked to solving physics tasks?
(4) To what extent are general linguistic and global physics proficiency related to the possibility to solve the specific physics tasks?

To investigate this questions, units of tasks were designed based on a model that integrates features related to both *cognitive activity* and *task complexity*, as well as linguistic features in assessment tasks. As mentioned above, the variation of linguistic features according to Heine et al. (2018) will be explored in a follow-up study and presented in a subsequent investigation. Therefore, in the current SCR study linguistic features are kept constant at the intermediate level (level II) within the unit between the tasks (Fig. 1).

Next, we will detail the construction of units of tasks, task stems, followed by an exposition of how we modelled *cognitive requirements* in the corresponding items.

**Figure 1**

*Integrated model of linguistic complexity and cognitive requirements. Shown is a unit of 9 tasks. It varies a task stem at 3 linguistic levels and the task itself (text-instruction + multiple-choice items) at 3 levels of subject-specific cognitive requirements (SCR levels). Only tasks from the column in the middle are relevant for the SCR study reported here.*




**Development of the Physics Test: Construction of Tasks**

Content of our Physics test relates exclusively to electricity. Electricity is one of the most important curricular elements in physics lessons at secondary level. Task in the Physics test address technical terms (e.g., alternating current circuit), concepts (e.g., separation of electric charges), phenomena (e.g., lightning during thunderstorm), or well-established school experiments (e.g., measuring strength of an electric current), usually taught in grade 8. Although the core curricula in Hamburg do not precisely specify the grade in which electricity should be taught, experienced teacher trainers in the second phase of German teacher education have recommended electricity as a core theme for this study because the study participants (especially 9th and 10th graders) are very likely to have been taught this curricular element.

All stem texts report an occurrence in a narrative manner, introduce acting individuals, and focus on a physical phenomenon from everyday life (e.g., electric shock) or a typical context from physics class (e.g., demonstration of electric conductors and non-conductors).

Each stem comprises a text of 24 propositions with approximately 200 words. The task stems in this study are highly controlled but relatively long. As mentioned above, this is due to the fact that the task development in the overall project VAMPS should in principle also allow for variation in linguistic complexity, even this was not the case in the study reported here. A decorative image illustrating the task context is attached to each task stem. However, images were chosen with great care, as it is known that they can affect the difficulty of a task (e.g., Carney & Levin, 2002). The stem is followed by an item containing a question or instruction and, and finally, multiple-choice answer options (1 attractor, 4 distractors). The attractor position varies unsystematically across all items. Distractors should appear as plausible as possible. Moreover, attractors and distractors should not differ visually (e.g., in text length or the mention of technical terms only in the attractor) to minimize the risk of successful task completion through test wisdom.

**Variation of Cognitive Requirements**

If a person starts reading the stem of a task, a varying amount of information will be inferred, selected, and interconnected depending on the instruction or the intended questions to be asked. In this case, we assume, that relevant prior knowledge needs to be recalled, activated, and processed repeatedly. As mentioned above, the level of cognitive requirement in our project is exclusively realized through the instruction and the multiple-choice options, but not by any variation of the stem. This led to the construction of three items assigned to each task stem. These three items are initially varied on increasing levels of cognitive requirements – I, II, III – while the linguistic complexity is held constant (Fig. 1).

If, in addition to a decorative-illustrative element, further visual elements are required for the task (e.g. electric circuit diagram), their level of abstraction increases from iconic (drawing of an electric circuit with light bulbs) to symbolically simple (e.g. diagram of a simple electric circuit) to symbolically complex (e.g. diagram of a complex and branched electric circuit).

The variation of levels of cognitive complexity in this study is based on the following three task features within a unit:

(1) The first feature concerns the type of *cognitive activity* classified on three levels. At the lowest level, a reproduction or recalling of knowledge is required. At an intermediate level, students are asked to establish relationships, apply a physics concept, classify elements, evaluate or explain a physical context, all at a simple level. At the highest level, a generative performance is required. Here, students are asked to apply more complex physics concepts, generate abstractions, make analyses or generalizations, or develop evaluations/explanations, all in a more complex manner. Cognitive activity modelled in this study is aligned with the well-known taxonomy of cognitive activities by Anderson et al. (2001). Not all cognitive activities proposed there were realized in this study (e.g., we did not consider *comparison* as a cognitive activity), so the focus was on a selection of activities that have proven particularly relevant in the construction of physics assessments. A systematic variation of types of knowledge suggested by Anderson et al. (2001, p. 29) such as

factual knowledge, procedural, conceptual, and metacognitive knowledge was not relevant in this study. As a task may elicit multiple cognitive activities simultaneously, the particular item was coded by prioritizing the highest discernible level of cognitive activity.

(2) As tasks in our study require relatively long task stems to be read, a cognitive sub-task involves reading comprehension of the stems, including the development of text-adequate situational models (Schnotz, 2006), identification of relevant information, and their retention in working memory. Therefore, the second feature by which cognitive requirements were varied is the *number of information entities which must be derived from the task stem* to answer the item correctly. Theoretically, we assume that task difficulty increases with the *number of information entities* because an increase in cognitive load is expected (Sweller et al. 2011). Information entities were considered as facts (e.g., applied voltage is 4.0 V), terms (e.g., electric shock), concepts (e.g., electric current is flowing), objects and their properties (e.g., battery with 1.5 V), or processes (e.g., observation of a light bulb when circuit is manipulated). Attributes are deemed relevant as long as they convey pertinent information, (e.g., two independent switches). This concept of information units aligns with Kauertz's definition of a *fact* (2008, p. 36) but has undergone refinement through discursive discussions throughout the development of tasks in this study.

(3) The third feature by which cognitive requirements was varied in this study relates to the *number of subject-specific mental procedures* needed to answer an item correctly. The assumption was made that, in principle, a right answer should be discernible without consulting the multiple-choice options. The assumption that the level of cognitive requirement increases with the number of subject-specific mental procedures is plausible, although – and this is likely true for most studies using multiple-choice items – the chain of reasoning when identifying an attractor as correct is not known to us with certainty. The subject-specific mental *cognitive procedures* were understood as the stepwise processing of information entities. Information entities had to be either extracted from the task stem or activated as prior knowledge. The processing of information entities is understood as either subordination or coordination. An overview of the model of cognitive requirements is presented in Table 1.

The initial coding of the tasks with regard to the three task features was discussed and communicatively validated several times by the authors of this paper during the entire process of task development. The objective of these discussions was twofold: Firstly, it was essential to reliably estimate the level of the highest cognitive activity needed for solving a task. Moreover, the scope, relevance, and extent of all information units had to be determined, and a concise sequence of mental procedures established—from recalling or discerning information to identifying an attractor. Secondly, a developmental and discussion phase ensued, during which the tasks were independently coded a second time by the second and third authors of this paper. Any remaining discrepancies primarily pertained to the extent of an information unit. These were further deliberated upon, resulting in a shared understanding, and an entirely and consistently rater agreement for all tasks.

**Table 1**

*Model of Cognitive Requirements*

| | |
|---|---|
| **I Cognitive activity** | |
| **1. 1. Reproducing** | |
| *Reproducing* | Recalling or identifying a straightforward technical/physical factual context, fact, concept, or technical term |
| **2. Establishing Connections** | |
| *Application simple* | Utilizing / applying a physics-specific concept / procedural knowledge in a familiar context |
| *Creating* | Synthesize single elements into something new |
| *Evaluating simple* | Forming a simple judgment based on criteria |
| *Classifying* | Assigning multiple elements to a category |
| *Explaining simple* | Assigning a cause to a scientific context/phenomenon |
| **3. Generating** | |
| *Application complex* | Utilizing / applying a subject-specific concept / procedural knowledge in a non-familiar context |
| *Abstracting* | Identifying abstract-structural properties of a subject-specific context |
| *Evaluating complex* | Forming a complex judgment based on criteria |
| *Explaining complex* | Assigning multiple (interlinked) causes to a physical context/phenomenon |
| **II Information Derived from Stem** | |
| | 1 – 8 information entities |
| **III Number of Subject-Specific Mental Procedures** | |
| | 1 – 8 procedures |

**Development of Test Items**

The development of items was based on two pilot studies. In Pilot Study 1, conducted in spring 2020, the levels of cognitive requirements 1, 2, and 3 of 35 units of tasks were developed and examined with the aim of evaluating the tasks in terms of students' (reading) comprehension. A total of n = 58 9th-grade students from three schools in Hamburg were divided into small groups of 2–3 participants. Task stems and a corresponding item at one of the three levels of cognitive requirements were presented to the groups for processing. Prior to this, students were instructed to talk as much as possible during the activity. The conversations were audio-recorded and transcribed, and group work was also observed by a trained researcher using a guideline.

Hints regarding reading comprehension emerged during the oral reading and paraphrasing of the task stems by the students. Moreover, students were asked to indicate perceived text difficulty of task stems on a 5-point scale (1–5) (M = 3.67 +/- 2.01). These were largely rated as understandable (Kesten, 2020, p. 47). Only three stems classified as

particularly difficult were substantially revised or discarded for subsequent studies. Problems with student processing were identified for 13 items, and formulations were subsequently revised to increase clarity and precision. Unattractive distractors (selected by < 5% of individuals) were identified and revised for 10 items. Pilot Study 1 also provided initial indications of a test difficulty which overall turned out to be slightly too high. Furthermore, the processing behavior and its results were analyzed to determine to what extent students used the task stems when working on our multiple-choice items. We have expected students firstly reading a stem, then reading the instruction, and finally returning to the stem. It was observed that students either did not reread the task stems (31 cases), reread parts of them (33 cases), or even reread them intensively (24 cases) (Kesten, 2020, p. 49). Upon inquiry, it became apparent that even the task stems that were not reread during task processing were well understood by the students. Overall, there were indications that task stems were actually used for item processing (Kesten, 2020).

Subsequently, all three levels of cognitive requirement of the tasks within a unit were presented to the students, although they had only intensively worked on an item on a single level before. There was a tendency for students to rate the level of cognitive requirement as the most difficult if they had been intensively worked on it before, regardless of the item classification according to our model of cognitive complexity. For items not intensively worked on, students assessed their difficulty only based on surface features. Therefore, in Pilot Study 1, it was evident that students were not suitable raters for item difficulty, a finding that had already been shown elsewhere for teachers (Impara & Plake, 1998). A validation of levels of cognitive requirements by students was therefore excluded for all subsequent developmental steps.

In Pilot Study 2 in spring 2021, the revised items were examined using a quantitative design. A total of 90 tasks from 30 units, differing in terms of SCR classification, were presented to a sample of n = 727 students from 9th and 10th-grade classes at secondary schools in Hamburg. 15 tasks were presented to each student, ensuring that a task stem was presented to each person not more than once. The items were rotated in blocks across 18 test booklets, so that each item appeared in 3 test booklet variations at variable positions, to avoid cumulating position and fatigue effects on individual items. For each item 115 to 122 observations were made. The test items of Pilot Study 2 were evaluated using the Rasch model (1PL). All test items showed good item fit values (weighted MNSQ between 0.8 and 1.2), with neither floor nor ceiling effects (percentages correct between 5% and 95%). Although the EAP reliability is low at 0.401, it is still acceptable for the developmental phase.

Pilot Study 2 also indicates that the test is challenging for the intended sample. Items with insufficient psychometric quality, those with solution frequencies indicating excessive item difficulty, or those deviating significantly from the expected ordering SCR I < SCR II < SCR III were removed from the item pool. Difficulty and distractor analyses led to further revisions for 46 tasks, with the goal of aligning difficulties of the 3 items per task unit as expected along the SCR levels and ensuring that distractors were chosen by at least 5% of the test takers.

Overall, the two pilot studies provided information for the verification and revision of task stems and items in terms of perceived comprehensibility, clarity, precision, fluid workability, adequate psychometric quality, fit to the Rasch model, clarity of instructions, and a sufficient selection of distractors (> 5%).

**Sample, Measures and Test Implementation of the SCR Study**

The SCR study was conducted in 2021 at secondary schools in Hamburg with social indeces 2 to 6[1]. A total of 414 students, comprising 14 grade-9 classes and 5 grade-10 classes, participated in this study. Students in grades 9 and 10 should have already learned the content addressed in our test (typically grade 8). All tasks presented in test booklets relate to the topic of electricity. From here on, we refer to this test as the "Physics test". In addition, a questionnaire related to socio-demographic information, a test section to assess basic *language proficiency* (Cloze test), and a test to assess *global proficiency in physics* (adapted TIMSS test) were presented. The Cloze test represents a segment of the DCLL+3, a standardized language proficiency test for grades 7 and 8.[2] The test is based on a classic, c-test-like deletion principle, where the front halves of 25 words are deleted at syllable boundaries in a text of 100 words. The adapted TIMSS test is based on published items from TIMSS. In addition to the items available in German, some items were translated into German. Furthermore, open-ended items were converted into closed ones. The test properties have been examined in a separate study (Feser & Höttecke, 2023). Since the test is very mechanics-oriented, in the SCR study it was supplemented with additional items from an instrument on electric circuits (Engelhardt & Beichner, 2004). The adapted TIMSS test used in the SCR study includes items on electricity (6), mechanics (3), optics (2), magnetism (2), and thermodynamics (2).

For every task in the Physics test, students were not only asked to answer the related items but also to indicate their level of familiarity with the task's content and express their interest in the specific topic.

The SCR study is based on a quasi-experimental design with a priori classifications of the items along the three features of cognitive requirement as dependent variables and item parameter as independent variable. For the Physics test, items on the three SCR levels were developed for each task unit. The test booklets were constructed in such a way that the 15 items presented in each booklet contained an equal number of tasks at SCR levels I, II, and III. The assignment of an item to a SCR level was based on the rule that the task feature (*cognitive activity*, *number of information entities derived from stem*, *number of subject-specific procedures*) rated highest determined the overall rating at a SCR level. The students always worked on a maximum of one of the three SCR item variants of a task unit. Each task was rotated into 3 of the total 12 test booklets. The test booklets were randomly assigned to students. After completion of the test, a total of 414 completed datasets were recorded for further data processing. Each of the 60 items in the Physics test (20 task units with three tasks at SCR levels I–III) was completed by at least

[1] The social index called KESS is assigned to schools in Hamburg and describes the socio-economic composition of the student body on a scale from 1 (disadvantaged) to 6 (privileged). See: https://www.hamburg.de/bsb/hamburger-sozialindex/
[2] The DCLL+3 is a standardized test and is offered by the Test Development and Diagnostics Department of the Hamburg Institute for Educational Monitoring and Quality Assurance.

100 and a maximum of 107 test subjects. The Cloze test and the adapted TIMSS test were completed by the entire sample.

Due to unreadable or sporadically missing answers to background information and information about familiarity or interest in the tasks, some information is missing, which were considered as missing (listwise deletion) during the analyses. In the item response models used (1PL), missing answers were treated as incorrect responses.

## Results

The Physics test, the Cloze test, and the adapted TIMSS test were evaluated using the Rasch model (1PL) in separate models. All test items showed good item fit values (weighted MNSQ between 0.8 and 1.2), with neither floor nor ceiling effects (percentage correct between 5% and 95%). In the Physics test (60 items; EAP reliability 0.446), only one challenging item (Ph083a) exhibited a negative point-biserial correlation ($r_{bis}$ = -0.09) of the correct answer with the estimated performance values of the participants; however, since the corresponding values of the easy ($r_{bis}$ = 0.38) and medium ($r_{bis}$ = 0.13) difficulty levels showed positive values, and the fit value was within an acceptable range at 1.07, this item was retained in the overall test.

By using the Rasch model for scaling the Physics test with acceptable parameters for the included items the first research question can be answered positively. It is possible to create a one-dimensional Physics test for a specific content area with items that focus on the same tasks by variating the features cognitive level and complexity systematically.

The distribution of item difficulties and the achieved test scores (steam-and-leaf plot) indicate, for the Physics test, that the tasks, measured against the average person's abilities, tend to be somewhat challenging. Taking this into account, the low reliability can be accepted in the context of the development of the tests but needs to be improved for following studies. The Cloze test exhibits an EAP reliability of 0.649 for 18 items, and the adapted TIMSS test, comprising 16 items, shows an EAP reliability of 0.615. In these tests the person's abilities tend to be somewhat higher compared to the corresponding item difficulties.

To get an impression of convergent and discriminant ability of the developed Physics test, the students' performances in this test are predicted by the Cloze and the adapted TIMSS tests (see Table 2). As single peredictor the adapted TIMSS score (Model I) shows a standardized regression coefficient with a low effect size (< 0.3) and explains 5.4 percent of the variance of the Physics test. The Cloze test score as a single predictor (Model II) has a medium effect size of (0.3 < beta < 0.5) and explains 11.0 percent of the variance of the Physics test. Whereas the standardized regression coefficients in the combined Modell III both predictors show low effect sizes (< 0.3) but together explain 12.9 percent of the variance of Physics test. The change of the beta coefficients across Model I, II and III indicate a substantial collinearity of both predictors. In the combined Model II the general language skills (Cloze test beta = 0.291) exhibit a higher correlation compared to global physics proficiency (adapted TIMSS test

beta = 0.150). These findings do not align with expectations, as a higher correlation of the Physics test with the adapted TIMSS test as a measure of students' global physics proficiency was anticipated, compared to the general language test. This could be an indicator that the extensive text-based stems of the items seem to be linguistically more demanding than expected. The described difference (steam-and-leaf plot) between item difficulty and persons abilities underlines this assumption.

For each unit, three systematically varied tasks for each of the cognitive requirement levels I–III were developed. It was assumed that the more difficult the tasks are, students give correct answers less frequently. As described above, three features of the tasks were distinguished: Cognitive activity on three levels (1–3), number of subject-specific mental procedures (1–8), and number of information entities from the stem (1–8) (Tab. 1). Since the counts for the number of information entities and the number of subject-specific mental procedures are low for codes 6, 7, and 8, they have been aggregated into "6 and more". The frequencies of these features for the 60 (3 cognitive variations * 20 units) tasks are summarized in Table 3.

**Table 2**

*Regression models, prediction of results in Physics test by adapted TIMSS test and Cloze test results separately and combined (z-standardized)*

| | **Model I** | | **Model II** | | **Model III** | |
|---|---|---|---|---|---|---|
| | **beta** | **p-value** | **beta** | **p-value** | **beta** | **p-value** |
| Intercept | 0.000 | 1.000 | 0.000 | 1.000 | 0.000 | 1.000 |
| Adapt. TIMSS score | 0.236 | <0.001 | | | 0.150 | 0.002 |
| Cloze test score | | | 0.336 | <0.001 | 0.291 | <0.001 |

Dep. Var.: Physics test score

**Table 3**

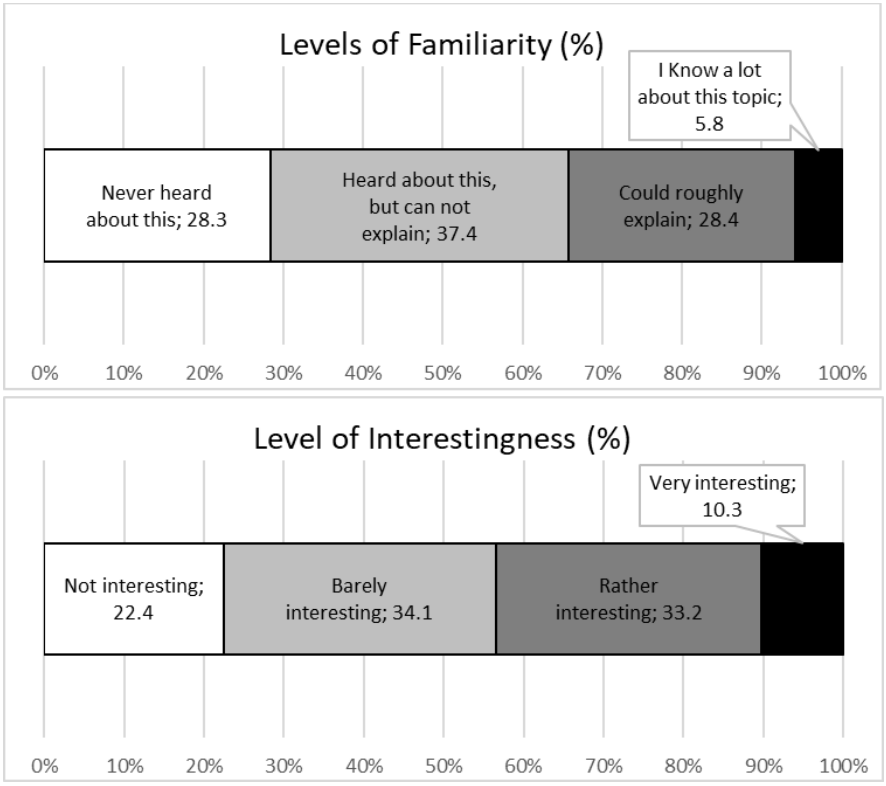
*Frequency of the task features of cognitive requirement – cognitive activity (1-3), number of information entities (1-6) and number of mental procedures (1-6) for n = 60 tasks*

| | **Cognitive activity** | | **Number of information entities derived from task stem** | | **Number of subject-specific mental procedures** | |
|---|---|---|---|---|---|---|
| code | n | % | n | % | n | % |
| 1 | 16 | 26.7 | 9 | 15.0 | 12 | 20.0 |
| 2 | 24 | 40.0 | 14 | 23.3 | 13 | 21.7 |
| 3 | 20 | 33.3 | 15 | 25.0 | 15 | 25.0 |
| 4 | | | 7 | 11.7 | 8 | 13.3 |
| 5 | | | 5 | 8.3 | 4 | 6.7 |
| > = 6 | | | 10 | 16.7 | 8 | 13.3 |
| Total | 60 | 100.0 | 60 | 100.0 | 60 | 100.0 |

Analysis of item difficulties in the Physics test also considered students' self-reported familiarity with the content of the task (low 1 to high 4; for example: "Did you already know something about the topic 'Operation of multimeters' before?") and their interest in the context presented by the task (low 1 to high 4; for example: "How interesting do you find the topic 'Operating multimeters'?"). Measures determined in the Cloze test and the adapted TIMSS test (standardized) are also included in the analysis.

**Figure 2**

*Students' self-reported level of familiarity with the context of tasks and their interest (in percent)*




From Figure 2, it can be observed that approximately 28% of the students indicated that they are not familiar at all with the context of the tasks on average, and only 6% reported having advanced knowledge. Regarding the interest in a task's content, slightly more than half (57%) of the respondents expressed little to no interest in the context presented in the task, while about 43% stated moderate to high interest.

To explore the connection between task features and the likelihood of correct answers, we utilized a Generalized Linear Mixed Model (GLMM). In this model, the dichotomously coded response to a task (0: incorrect, 1: correct) is predicted by both item and student features. Acknowledging the potential variability in task difficulty, the model is structured as a multilevel model. This entails treating the three tasks on the three SCR levels, all belonging to one task unit (the middle column in Fig. 1), as a cluster.

**Table 4**

*Generalized linear mixed models (GLMM) to predict item response based on item features and person characteristics*

| | | **Model I** | | **Model II** | | **Model III** | |
|---|---|---|---|---|---|---|---|
| Effect | Code | Estimate | Pr > \|t\| | Estimate | Pr > \|t\| | Estimate | Pr > \|t\| |
| Intercept | | -1.659 | < 0.001 | -2.498 | < 0.001 | -2.520 | < 0.001 |
| Cognitive activity | 1 | 1.361 | < 0.001 | 1.402 | < 0.001 | 1.428 | < 0.001 |
| Cognitive activity | 2 | 0.815 | < 0.001 | 0.780 | < 0.001 | 0.800 | < 0.001 |
| Cognitive activity | 3 | 0.000 | – | 0.000 | – | 0.000 | – |
| Number Inform. | 1 | 0.029 | 0.951 | 0.078 | 0.873 | 0.054 | 0.913 |
| Number Inform. | 2 | -0.010 | 0.982 | 0.097 | 0.821 | 0.085 | 0.846 |
| Number Inform. | 3 | 0.453 | 0.223 | 0.544 | 0.153 | 0.540 | 0.163 |
| Number Inform. | 4 | -0.426 | 0.322 | -0.405 | 0.369 | -0.421 | 0.355 |
| Number Inform. | 5 | 0.515 | 0.163 | 0.583 | 0.118 | 0.602 | 0.109 |
| Number Inform. | > = 6 | 0.000 | – | 0.000 | – | 0.000 | – |
| Number procedures | 1 | 0.168 | 0.785 | 0.057 | 0.927 | 0.069 | 0.914 |
| Number procedures | 2 | 0.296 | 0.544 | 0.221 | 0.663 | 0.245 | 0.631 |
| Number procedures | 3 | 0.320 | 0.488 | 0.257 | 0.583 | 0.275 | 0.561 |
| Number procedures | 4 | 0.490 | 0.293 | 0.419 | 0.383 | 0.445 | 0.356 |
| Number procedures | 5 | 0.439 | 0.368 | 0.395 | 0.430 | 0.417 | 0.406 |
| Number procedures | > = 6 | 0.000 | – | 0.000 | – | 0.000 | – |
| Gender (f) | 1 | | | -0.033 | 0.551 | -0.065 | 0.275 |
| Gender (m) | 2 | | | 0.000 | – | 0.000 | – |
| Grade Level | 9 | | | -0.021 | 0.760 | 0.073 | 0.305 |
| Grade Level | 10 | | | 0.000 | – | 0.000 | – |
| Familiarity | | | | 0.326 | < 0.001 | 0.311 | < 0.001 |
| Interestingness | | | | 0.099 | 0.002 | 0.082 | 0.010 |
| adapt. TIMSS Score | | | | | | 0.103 | 0.001 |
| Cloze Test Score | | | | | | 0.195 | < 0.001 |
| Variance explained | | $R^2$ = 9.10% | | $R^2$ = 12.09% | | $R^2$ = 13.46% | |

Depend. Var.: Item - response right-wrong (0;1)

$R^2$: calculated following Nakagawa & Schielzeth (2013); Nakagawa, Johnson, & Schielzeth (2017)

In Model I (Table 4), the prediction of item solutions is restricted to item features. In this case, neither the *number of information entities* nor the *number of mental procedures* show a systematic relationship with *item difficulty*. It was

expected that the parameters specified for codes 1 to 5, in comparison with the reference group (6 = 6 and more), would systematically increase from 5 to 1 – implying that the lower the *number of information entities* or *number of mental procedures*, the higher the likelihood of a correct answer (on the logit scale). Only the highest level of *cognitive activity* exhibits the expected relationship with higher likelihoods of correct answers by lower demands of cognitive activities.

The answer to research question 2, to what extent do the task features *cognitive activity* and *complexity* influence the difficulty to solve the physics tasks, is that only the addressed cognitive activity varies the difficulty of the tasks as expected.

The results of Model II show the answer to research question 3, about how closely individual characteristics of students are related to the likelihood of solving a physics task. In this model, gender, grade level, and the self-reported prior knowledge and interest described by the students are included as additional predictors. Regarding gender and grade level, both are not significant. As expected, higher familiarity with the context of a task corresponds to a significantly higher probability of giving a correct answer. The same applies to self-reported interest, which also indicates higher probabilities of correct answers as it increases.

As a further extension of the analyses, Model III includes the z-standardized performance scores of students in the Cloze test and adapted TIMSS test to find an answer to research question 4: To what extent are general linguistic and global physics proficiency related to the likelihood to solve the specific physics tasks? Even after controlling for task features and students' self-reported familiarity and interest, it is evident that the general language proficiency nominally predicts the probability of correct answers better than global physics proficiency, corroborating the findings of the initial regression model (Tab. 2).

## Discussion

The aim of the SCR study presented here basically was to systematically vary three cognitively demanding features of physics tasks and their effects on task difficulty. We have modeled cognitive requirements towards solving a physics task as a combination of *cognitive activity* as well as *task complexity*. Complexity in our study was realized as the *number of information entities* to be derived from a task stem and the *number of mental processes* on the way to solving of a task. Based on comparison of three generalized linear mixed models, it turned out that *cognitive activity* was the only item feature which predicted item response. This result contributes to a body of research presented above which indicates *complexity* – even though operationalized in various ways across studies – as a highly inconsistent predictor of item response. Among the person features, language proficiency, global physics proficiency as well as students' familiarity with and interest in the context presented by a task proved to be predictive of item response.

The tasks were developed for 9th and 10th-grade students in physics class but were atypical due to rather text-heavy task stems. Nevertheless, two pilot studies as well as the SCR study have shown that students coped well with the extended texts. We assume that the students already had made several experiences of working on tasks with longer text during their school career in other subjects such as mathematics and therefore had no difficulties processing an amount of text higher than they might have expected in physics. During the test administration, the test administrators also took notes of the students' feedback. It was observed that the Physics test tasks, despite featuring unusually long task stems, were not further commented on by the students which indicates a high degree of acceptance. It can be assumed that although the format may seem unusual on first sight, it can be considered an ecologically valid task format in the context of physics instruction.

The coding of *cognitive requirement* based on the task features *number of information entities*, *number of mental procedures*, and *cognitive activity* proved to be challenging. In the course of this study, it was not possible to develop a strict hierarchical coding instruction that would allow these features to be precisely assessed. Both in the process of task development and in the process of re-coding the task features in the light of the response patterns provided by the students, it became evident that these features cannot be deterministically derived solely from the task stem, questions, and answer options. We suspect a fundamental problem in model-based task construction, which may contribute to an explanation of the unclear research situation outlined above.

A valid coding of the task feature *cognitive activity* – even though it turned out to predict item response in our study – reaches its limits, as the actually performed cognitive activity must be highly dependent on prior knowledge (Prediger & Aufschnaiter, 2023). In our model, *familiarity* with the context of a task proved to be predictive for item response. This result confirms the assumption, that an a-priori estimate of cognitive activity depends on students' prior knowledge as self-reported familiarity can be considered as a valid proxy variable for prior knowledge. Whether, for example, the mention of a physics concept is merely recalled and reproduced or requires a higher-order cognitive activity depends on students' prior learning and knowledge, not just on task features. Thus, a reproductive task may be particularly difficult if the declarative knowledge to be remembered is particularly abstract or barely represented in the curriculum (e.g., the distribution of electric charge in a thundercloud needs to be recalled), while a complex evaluation in a particularly familiar content area may even be easy (e.g., evaluating measures to extend the service life of a light bulb in an electric circuit). Accordingly, the taxonomy of cognitive activities our SCR study is based upon, proposed by Anderson et al. (2001), has been criticized for lacking discriminative power among the six levels of cognitive process complexity (Maier, Kleinknecht, Metz, & Bohl, 2010). As confirmed here, it is generally difficult to anchor thinking processes to the surface features of tasks, as their strongly inferential evaluation also involves knowledge about mental processes students use to solve a task (Kauertz, 2008, p. 23). We therefore assume the idea of considering tasks in assessments as an intersubjectively valid construct without taking familiarity into account (Prediger & Aufschnaiter, 2023) may be of limited practical relevance.

Regarding the task feature of *number of mental procedures*, based on information that students either recall or derive from the task stem, a sequence of mental steps leading to the attractor should occur. This feature, in addition to the *number of information entities* derived from the stem, in our study characterizes the complexity of the task. The assumption that the *number of mental procedures* and the *number of information entities* could predict item response is based on the assumption that test developers can actually predict cognitive strategies and chains of reasoning that students actually apply while working on a task. Again, this assumption presupposes that students' prior knowledge is known well. However, it is plausible to assume that test persons can arrive at task solution through various sequences of mental steps. For instance, test persons with high prior knowledge might entirely skip some of the expected mental steps, so that the empirically determined task difficulty differs from an expected level. This holds particularly true, as an attractor is more readily discernible when juxtaposed with distractors than when assessed in isolation. Additionally, the process becomes more effective when test persons can successively eliminate distractors, so that an attractor is finally recognized as the least inconclusive option. This idea is in line with the observation that the plausibility of distractors affects item difficulty (Hartig & Frey, 2012). In sum, various cognitive shortcuts for solving a task are conceivable, which problematizes the assumption of an anticipatable sequence of mental procedures that lead to an attractor.

The findings of this study suggest that cognitively challenging solution processes for students remain a desideratum. To identify, describe, and analyze cognitive solution paths but also (in the case of assumed multi-step solution paths) solution "shortcuts," cognitive labs resulting in thoroughly documented think-aloud processes, as we had started in our pilot studies, need to be conducted. The insights gained from such studies could then contribute to the systematic development of task difficulties or their coding based on various features.

In our SCR study, another noteworthy finding was the more prominent predictive role of language proficiency measured by the Cloze test compared to global physics proficiency assessed through the adapted TIMSS test. Despite their significance, neither language proficiency nor global physics proficiency achieved the predictive impact observed with familiarity, as indicated by self-reports on prior knowledge of the topic. The text-heavy nature of the task stems, though deemed ecologically valid based on student feedback, poses a challenge to the students in extracting relevant information. Consequently, it is plausible that language proficiency significantly influences item response in this study. On the contrary, global physics proficiency in our study necessitates knowledge spanning various fields of physics, in addition to the topic of electricity on which the Physics test was restricted. We attribute the dependence of knowledge on diverse physics fields to the widely varying curricula implemented in schools, especially considering the official curricula in Hamburg, which articulate competency objectives at the conclusion of grades 8 and 10 while staying reluctant in specifying precise lesson content. This lack of specificity contributes to a diversity of teaching approaches across schools and topics. Thus, the dominance of language proficiency over global physics proficiency in predicting item responses in our Physics test is further underscored by the specific focus of this study on the topic of electricity. The lower explanatory power of global physics proficiency, compared to

language proficiency, finds justification in this context. This interpretation is further supported by the robust predictive capability of the *familiarity* variable as a proxy for prior knowledge.

## Limitations

The study is subject to several limitations, starting with the "Model of Variation of Cognitive Requirements". Through variation of cognitive activity, the actual difficulty of a task can only be systematically influenced to a limited extent. Instead, the actual difficulty is strongly related to prior knowledge which is consistent with research (Ropohl et al., 2015; Prediger & Aufschnaiter, 2023). With the adapted TIMSS test used in this study, (content-specific) prior knowledge can only be partially represented and, therefore, only empirically controlled to a limited extent. The self-assessments of prior knowledge documented by the students appear to be empirical more robust but are, however, afflicted with known limitations of self-assessments (Brown, Andrade & Chen, 2015). The dependence on prior knowledge is particularly challenging for the subject of physics, as school- or class-specific curricula do not follow an obligatory canon and sequence of teaching. Thus, some students may have not yet covered the topic, covered it a long time ago, or just covered it recently. This limitation should be addressed in follow-up investigations by focusing on a defined and limited content area that has not been taught in the classes of the sample and is then conveyed in uniformly timed instructional sequences preceding the test. This way, the topical relevance of the theme is kept constant for all students and confounds less with expected or necessary prior knowledge. This limitation becomes particularly pronounced in the SCR study, as school closures, alternating classes with half class sizes, and suspended mandatory attendance in schools due to the pandemic in 2021 have led to significant restrictions in the sequencing of instruction and the treatment of the physics curriculum (from the 7th to the 9th grade). Thus, it cannot be reliably assumed that all participants in our study had sufficient prior knowledge to answer the items. This is further confirmed in the self-assessments of the test subjects regarding familiarity with the task contexts in the SCR study.

A reliable assessment of the *number of mental procedures* in the solution path proved to be highly limited, mainly due to the possibility that students may use distractors to solve the tasks. For future task development, it is advisable that distractors not only represent (plausible) incorrect answer options but also address pre-instructional concepts deliberately to systematically consider their difficulty-generating property, as is done, for example, in ordered multiple-choice items (Hadenfeldt et al., 2013).

Furthermore, the validation of the coding was conducted communicatively rather than through independent ratings. Independent ratings would have the advantage of reliability testing. However, all items had already undergone a shared and intensive discussion and were highly familiar to all authors, making an independent rating after this phase of discursive learning less meaningful.

## Acknowledgements

The authors would like to thank the editors and anonymous reviewers for their helpful feedback, all students involved in the study, Lena Heine (Ruhr University Bochum), Dominik Leiß and Timo Ehmke (Leuphana University Lüneburg) as well as all members of the Physics Education group (University of Hamburg) for their support and feedback. Special thanks go to Carina von der Geest for her support in developing the test items.

## Funding

The DFG (Deutsche Forschungsgemeinschaft) supported this study (grand number 417017613).

## References

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M.C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* New York: Longman.

Bernholt, S., & Parchmann, I. (2011). Assessing the complexity of students' knowledge in chemistry. *Chemistry Education Research and Practice, 12*, 167–173. https://doi.org/10.1039/C1RP90021H

Brown, G.T.L., Andrade, H.L., & Chen, F. (2015). Accuracy in student self-assessment: directions and cautions for research. *Assessment in Education: Principles, Policy & Practice, 22*(4), 444-457, https://doi.org/10.1080/0969594X.2014.996523.

Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review, 14*(1), 5–26. https://doi.org/10.1023/A:1013176309260

Cruz Neri, N., Guill, K., & Retelsdorf, J. (2021). Language in science performance: Do good readers perform better? *European Journal of Psychology of Education, 36*, 45–46. https://doi.org/10.1007/s10212-019-00453-5

Engelhardt, P. V., & Beichner, R. J. (2004). Students' understanding of direct current resistive electrical circuits. *American Journal of Physics, 72*(1), 98–115. https://doi.org/10.1119/1.1614813

Feser, M. & Höttecke, D. (2023). Development of a test in German language to assess middle school students' physics proficiency. *The Physics Educator, 5*(1), Article 2320002. https://doi.org/10.1142/S2661339523200020.

Gut-Glanzmann, C. (2012). *Modellierung und Messung experimenteller Kompetenz. Analyse eines large-scale Experimentiertests*. Berlin: Logos.

Hackemann, T. (2023). *Textverständlichkeit sprachlich variierter physikbezogener Sachtexte*. Diss., Universität Hamburg.

Hadenfeldt, J. C., Bernholt, S., Liu, X., Neumann, K., & Parchmann, I. (2013). Developing an instrument using ordered multiple choice items to access Students' understanding of the structure and composition of matter. *Journal of Chemical Education 90*(12), 1602–1608. https://doi.org/10.1021/ed3006192

Härtig, H., Heitmann, P., & Retelsdorf, J. (2015). Analyse der Aufgaben zur Evaluation der Bildungsstandards in Physik – Differenzierung von schriftsprachlichen Fähigkeiten und Fachlichkeit. *Zeitschrift für Erziehungswissenschaft, 18*, 763–779. https://doi.org/10.1007/s11618-015-0646-2

Hartig, J., & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. *Psychologische Rundschau, 63*(1), 43–49. https://doi.org/10.1026/0033-3042/a000109

Härtig, S., Bernholt, S., Fraser, N., Cromley, J. G., & Retelsdorf, J. (2022). Comparing reading comprehension of narrative and expository texts based on the direct and inferential mediation model. *International Journal of Science and Mathematics Education, 20,* 17–41. https://doi.org/10.1007/s10763-022-10302-5

Heine, L., Domenech, M., Otto, L., Neumann, A., Krelle, M., Leiß, D., Höttecke, D., Ehmke, T., & Schwippert, K. (2018). Modellierung sprachlicher Anforderungen in Testaufgaben verschiedener Unterrichtsfächer: Theoretische und empirische Grundlagen. *Zeitschrift für angewandte Linguistik, 69*, 69-96. https://doi.org/10.1515/zfal-2018-0017

Höttecke, D., Feser, M., Heine, L., & Ehmke, T. (2018). Do linguistic features influence item difficulty in physics assessments? *Science Education Review Letters*, https://doi.org/10.18452/19188.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*(1), 69–81. https://doi.org/10.1111/j.1745-3984.1998.tb00528.x

Jaeger, D., & Müller, R. (2019). Einflussfaktoren beim Lösen physikalischer Aufgaben. In C. Maurer (Hrsg.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Kiel 2018* (S. 293–296). Regensburg: Universität Regensburg.

Kauertz, A. (2008). *Schwierigkeitserzeugende Merkmale physikalischer Leistungsaufgaben*. Berlin: Logos.

Kesten, B. (2020). *Die Wirkungen sprachlicher und kognitiv-fachlicher Aufgabenmerkmale in Physik auf ihre Bearbeitung durch Schüler*innen.* Unpublished Master Thesis, University Hamburg

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research, 79*(3), 1168–1201. https://doi.org/10.3102/0034654309332490

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Knoche, N., & Lind, D. (2004). Eine differenzielle Itemanalyse zu den Faktoren Bildungsgang und Geschlecht. In M. Neubrand (Hrsg.), *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland. Vertiefende Analysen im Rahmen von PISA 2000* (S. 73–86). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-80661-1_5

Kulgemeyer, C., & Schecker, H. (2009). Kommunikationskompetenz in der Physik: Zur Entwicklung eines domänenspezifischen Kommunikationsbegriffs. *Zeitschrift für Didaktik der Naturwissenschaften, 15*, 131–153.

Kunter, M., J. Baumert and O. Köller (2007), "Effective classroom management and the development of subject-related interest", *Learning and Instruction*, Vol. 17/5, pp. 494-509, https://doi.org/10.1016/j.learninstruc.2007.09.002.

Le Hebel, F., Montpied, P., Tiberghien, A., & Fontanieu, V. (2017). Sources of difficulty in assessment: Example of PISA science items. *International Journal of Science Education, 39*(4), 468–487. https://doi.org/10.1080/09500693.2017.1294784

Maier, U., Kleinknecht, M., Metz, K., & Bohl, T. (2010). Ein allgemeindidaktisches Kategoriensystem zur Analyse des kognitiven Potenzials von Aufgaben. *Beiträge zur Lehrerinnen- und Lehrerbildung, 28*, 84–96.

Mannel, S., Sumfleth, E., & Walpuski, M. (2009). Schwierigkeitsbestimmende Faktoren von Aufgaben zu experimentell-naturwissenschaftlichen Arbeitsweisen für den unteren Leistungsbereich. In D. Höttecke (Hrsg.), *Chemie- und Physikdidaktik für die Lehramtsausbildung* (S. 380–382). Berlin: Lit.

Mesic, V., & Muratovic, H. (2011). Identifying predictors of physics item difficulty: A linear regression approach. *Physical Review Special Topics – Physics Education Research, 7.* https://doi.org/10.1103/PhysRevSTPER.7.010110

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science.* Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-results/

Nagy, G., Lüdtke, O., Köller, O., & Heine, J.-H. (2017). IRT-Skalierung des Tests im PISA-Längsschnitt 2012/2013: Auswirkungen von Testkontexteffekten auf die Zuwachsschätzung. *Zeitschrift für Erziehungswissenschaft, 20*(Suppl. 2), 229–258. https://doi.org/10.1007/s11618-017-0749-z

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining $R^2$ from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*, 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination $R^2$ and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface, 14*. https://doi.org/10.1098/rsif.2017.0213

Neumann, I. (2011). *Beyond physics content knowledge. Modeling competence regarding nature of scientific inquiry and nature of scientific knowledge*. Berlin: Logos.

Neumann, K., Viering, T., & Fischer, H. (2010). Die Entwicklung physikalischer Kompetenz am Beispiel des Energiekonzepts. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 285–298.

Neumann, K., Vollstedt, M., Lindmeier, A., Bernholt, S. Eckhardt, M., Harms, U., Härtig, H., Heinze, A., & Parchmann, I. (2013). Strukturmodelle allgemeiner Kompetenz in Mathematik und den Naturwissenschaften und Implikationen für die Kompetenzentwicklung im Rahmen der beruflichen Ausbildung in ausgewählten kaufmännischen und gewerblich-technischen Berufen. In R. Nickolaus, J. Retelsdorf, E. Winther & O. Köller (Hrsg.), Mathematisch-naturwissenschaftliche Kompetenzen in der beruflichen Erstausbildung. Stand der Forschung und Desiderate. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 26. Beiheft, 113–138.

OECD (2021), *21st-Century Readers: Developing Literacy Skills in a Digital World*, PISA, OECD Publishing, Paris, *https://doi.org/10.1787/a83d84cb-en*.

OECD (2023), *PISA 2022 Results (Volume II): Learning During – and From – Disruption*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/a97db61c-en.

Prediger, S. & Aufschnaiter, C.v. (2023, 2nd ed.). Umgang mit heterogenen Lernvoraussetzungen aus fachdidaktischer Perspektive: Fachspezifische Anforderungs- und Lernstufungen berücksichtigen. In T. Bohl, J. Budde & M. Rieger-Ladich (eds.), *Umgang mit Heterogenität in Schule und Unterricht. Grundlagentheoretische Beiträge und didaktische Reflexionen* (pp. 299–316). Bad Heilbrunn: Klinkhardt.

Prenzel, M., Häußler, P., Rost, J., & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft, 30*(2), 120–135.

Ropohl, M., Walpuski, M., & Sumfleth, E. (2015). Welches Aufgabenformat ist das richtige? – Empirischer Vergleich zweier Aufgabenformate zur standardbasierten Kompetenzmessung. *Zeitschrift für Didaktik der Naturwissenschaften, 21*, 1–15. https://doi.org/10.1007/s40573-014-0020-6

Schmidt-Barkow, I. (2010). Lesen – Lesen als Testverstehen. In V. Frederking et al. (eds.), *Taschenbuch des Deutschunterrichts* (pp. 218-231). Vol. 1, 9th ed., Baltmannsweiler: Schneider Hohengehren.

Schnotz, W. (2006). Was geschieht im Kopf des Lesers? Mentale Konstruktionsprozesse beim Textverstehen aus der Sicht der Psychologie und der kognitiven Linguistik. In H. Blühdorn, E. Breindl & U. H. Waßner (Hrsg.), *Text – Verstehen. Grammatik und darüber hinaus* (S. 222–238). Berlin: de Gruyter.

Solano-Flores, G., & Wang, C. (2015). Complexity of illustrations in PISA 2009 science items and its relationship to the performance of students from Shanghai-China, the United States, and Mexico. *Teachers College Record, 117*(1), 1–18. https://doi.org/10.1177/016146811511700103

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., Krüger, D., & Upmeier zu Belzen, A. (2016). Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education, 41*(5), 721–732. https://doi.org/10.1080/02602938.2016.1164830

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York: Springer. https://doi.org/10.1007/978-1-4419-8126-4

Walpuski, M., & Ropohl, M. (2014). Statistische Verfahren für die Analyse des Einflusses von Aufgabenmerkmalen auf die Schwierigkeit. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 385–398). Berlin: Springer. https://doi.org/10.1007/978-3-642-37827-0_30

Wellnitz, N., Fischer, H. E., Kauertz, A., Mayer, J., Neumann, I., Pant, H. A., Sumfleth, E., & Walpuski, M. (2012). Evaluation der Bildungsstandards – eine fächerübergreifende Testkonzeption für den Kompetenzbereich Erkenntnisgewinnung. *Zeitschrift für Didaktik der Naturwissenschaften, 18*, 261–291.

Ziepprecht, K., Schwanewedel, J., Heitmann, P., Jansen, M., Fischer, H. E., Kauertz, A., Kobow, I., Mayser, J., Sumfleth, E., & Walpuski, M. (2017). Modellierung naturwissenschaftlicher Kommunikationskompetenz – ein fächerübergreifendes Modell zur Evaluation der Bildungsstandards. *Zeitschrift für Didaktik der Naturwissenschaften, 23*, 113–125. https://doi.org/10.1007/s40573-017-0061-8

**Corresponding Author Contact Information:**

| | |
|---|---|
| **Author name**: | Dietmar Höttecke |
| **Department**: | Faculty of Education, Physics Education |
| **University, Country**: | University Hamburg, Germany |
| **Email**: | dietmar.hoettecke@uni-hamburg.de |

**Please Cite:** Schwippert, K., Zilz, K. & Höttecke, D. (2024). Difficulty-Generating Features of Text-based Physics Tasks. *Journal of Research in Science, Mathematics and Technology Education, 7*(2), 1-24 https://doi.org/10.31756/jrsmte.721



**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**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.

**Data Availability Statement**: The raw data supporting the conclusions of this article will be made available by the authors upon request, without undue reservation.

**Ethics Statement:** This study was reviewed and approved by the Institute for Educational Monitoring and Quality Development of the Hamburg. Ethical principles were fully observed. In particular, the study participants were informed about their rights and informed about the voluntary nature of participation, rights of withdrawal and principles of transparency.

**Author Contributions**: All authors listed have made a substantial, direct, and intellectual contribution to the work andapproved it for publication.

*Received: March 16, 2024 ▪ Accepted: May 27, 2024*