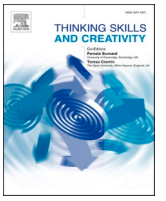
Thinking Skills and Creativity 50 (2023) 101387

Contents lists available at ScienceDirect

Thinking Skills and Creativity

journal homepage: www.elsevier.com/locate/tsc



Development of a formative assessment instrument to determine students' need for corrective actions in physics: Identifying students' functional level of understanding

Frits F.B. Pals [a], Jos L.J. Tolboom [b,*], Cor J.M. Suhre [a]

[a] University of Groningen, the Netherlands
[b] Netherlands Institute for Curriculum Development, SLO, P.O. box 502, Amersfoort, AM 3800, the Netherlands

ARTICLE INFO

Keywords:
Formative assessment
Secondary science education
Problem solving
Skill theory
Modelling cycle
Functional level of insight

ABSTRACT

In physics education, most teachers provide students feedback on their problem solutions through grades on written tests. The practice of feedback after a summative test does not often meet the needs of many students to improve their problem solving. In this paper we report on the development of a formative assessment instrument to allow teachers to provide more meaningful action-oriented feedback on students' performance on written tests.

Our research and development approach comprised three phases.The first phase consisted of a literature guided cognitive analysis of effective problem-solving strategies in the physics domain. This analysis resulted in the identification of three crucial episodes in students' problem-solving approaches during which students engage in specific cognitive activities. The second phase consisted of the design of an assessment instrument to monitor specific cognitive activities during the three crucial episodes when solving physics problems. This resulted in a rating scale with 11 levels to indicate students' efficacy. The third phase consisted of research of the validity, reliability, and practicality of the instrument. Here we trained three teachers to trace students' mistakes on different problems in the domain of kinematics and asked them to rate the mastery of 16 eleventh-grade pre-university students. In this phase we assessed the reliability and validity of this instrument by computing Krippendorff's alpha to indicate teachers' inter-rater reliability. Practicality of the instrument was assessed by examining the variation in students' level of mastery on problems of different complexity. Further research is needed to provide more detailed guidelines for how teachers can use the instrument in formative assessment (in contrast to summative assessment) to help students to develop correct solution methods and foster students' metacognition about problem solving in related physics areas (i.e., knowledge transfer).

1. Introduction

Many students in secondary education consider physics—as part of science—a difficult subject, because it relies on abstract concepts and requires applications of mathematical procedures to solve problems. Failure to apprehend connections between abstract

Abbreviations: RTO, rejected take-off; HAVO(Dutch), higher general secondary education; SR, self-generated representations; STEM, science, technology, engineering, and mathematics.

* Corresponding author.
E-mail address: j.tolboom@slo.nl (J.L.J. Tolboom).

https://doi.org/10.1016/j.tsc.2023.101387
Received 5 October 2022; Received in revised form 21 August 2023; Accepted 24 August 2023
Available online 25 August 2023

concepts and a lack of necessary mathematical knowledge can lead to frustration and disappointment for both teachers and students (Abend, 2018; Britner & Pajares, 2006; Chinn, 2020; Levering, 2010; Ryan et al., 2022; Schutz et al., 2020). Moreover, many science students continue to struggle and need additional tutoring to pass their examinations.

To support students' problem-solving abilities, science teachers must be able to obtain insight of students' solution approaches of physics as well as mathematics (Izquierdo-Acebes & Taber, 2023), and conduct rapid analyses to identify the specific knowledge that students lack to solve physics problems. Teachers benefit when they can identify this gap in knowledge. Because, tests are usually administered in written form, it is crucial for teachers to improve their sensitivity and ability to interpret students' descriptive work and to identify gaps in students' declarative, procedural, and conceptual knowledge. According to Vygotsky (1978), Fischer and Bidell (2006), Taber (2018), there is a range of cognitive activities that students can do well without support (Zone of Proximal Development) and beyond that there are cognitive activities that students consider hard to complete independently and need support to accomplish.

Despite analysing substantial research into various aspects of science students' problem solving, such as their application of appropriate mathematics procedures (Hsu et al., 2004; Ince, 2018; Maloney, 2011), many research findings do not transfer readily to daily classroom practice. Teachers are challenged by widely varying students' knowledge gaps and misconceptions (Chi, 2013; Marsh & Eliseev, 2019), and they often may lack sufficient time (Abend, 2018) and know-how to explore in depth students' cognitive errors in paper tests (Thiede et al., 2019). This challenge clearly demands the development of practical diagnostic instruments and teacher training to help teachers analyze students' test solution approaches and identify their methodological weaknesses, but few studies have sought to develop such tools (Pol et al., 2008; Tolboom, 2012; Chinn, 2020). In response, we seek to design and evaluate a transparent diagnostic instrument that increases science teachers' abilities to assess students whose physics knowledge is lacking which results in an erroneous action and to identify 'where' students' methods need improvement.

Specifically, we describe the design, development and implementation of a formative assessment instrument to monitor the problem-solving capacities of 16- to 17-year-old students in the field of kinematics. We focus on kinematics, as teachers note that the plurality of this domain—it has at least nine elements and several different formulas—often causes anxiety and cognitive confusion to students (Molin, 2021; Schoenfeld, 2016). Moreover, the elements (concepts of distance, velocity and acceleration) which are central to this domain are widely used in science, economics, and biology (Ottevanger et al., 2014; van der Zwaart, 2007).

In this paper, we focus on the initial stages of the development of the instrument for pinpointing the 'where' of the gap of students' knowledge in solving physics problems, the focus on possible support to students is subject to a sequential study. The usefulness of the to be developed formative assessment instrument depends on its accuracy in pinpointing students' needs for meaningful feedback and practicality for informing teachers how to supply this feedback in the classroom context. Our research was guided by the following research questions:

(1) Which cognitive steps and activities need teachers monitor to help students successfully solve physic problems?
(2) Can the cognitive diagnostic instrument of the code table reliably identify students' actual levels of problem-solving performance?
(3) Does the instrument provide teachers information helpful for deciding to apply whole classroom feedback or differentiated instruction depending on students' functional level of performance in solving problems tapping on different elements in the kinematics domain?

2. Method

2.1. Study setting

The research and development approach as deployed during this study comprised three phases. The first phase consisted of a literature guided analysis of defining the challenges teachers face in monitoring students' development in kinematics problem solving and resulted in three episodes to monitor of evaluating students' mastery of instructed problem-solving methods. The second phase comprised the design of the assessment instrument to monitor students' performance of specific cognitive activities during crucial episodes when solving physics problems. The third phase consisted of research of the validity, reliability, and practicality of the instrument for supporting teachers in supplying appropriate feedback to (specific groups of) students. Three teachers were trained to trace students' mistakes on different problems in the domain of kinematics and asked them to rate the mastery of 16 eleventh-grade pre-university students. We assessed the reliability and validity of this instrument by computing Krippendorff's alpha to indicate teachers' inter-rater reliability. Practicality of the instrument was assessed by examining the variation in students' level of mastery on problems of different complexity. If this examination unveils variation in mastery of different problems, this provides support for our expectation that teachers can use the instrument to decide between delivering whole classroom re-instruction or additional groupwise and individual instruction based on students' needs.

2.2. Phase 1: Defining the challenges teachers face in monitoring students' development in kinematics problem solving

To solve kinematics problems requires that students develop an integrated knowledge base of the abstract concepts of distance, displacement, velocity and acceleration, their usefulness in understanding and solving real world problems and how problems concerning these concepts can be modelled mathematically. During the initial instruction period, teachers need to ascertain that students are able to distinguish differences across the elements of kinematics, identify unique problem settings, and apply the appropriate mathematical procedures (Schoenfeld, 2016) to become expert in solving specific kinematics problems. The specific kinematics

concepts therefore deserve critical attention from teachers and students, because each concept is related to one or more distinctive solution methods. Unfortunately, after the initial preparatory instruction, not all science students recognize the relationships between the concepts of distance, displacement, velocity and acceleration as a coherent set (Powell et al., 2009) and how to model; novices in particular tend to fail to do so (Chi et al., 1981; Paas & van Merriënboer,2020).

2.3. Three crucial episodes to monitor in solving science problems

For most students, a step-by-step instruction approach focusing on crucial episodes in the solution of problems is key to helping students develop flexible problem-solving skills in the domain of kinematics and allowing teachers to monitor and support students' progress efficiently. We illustrate these crucial episodes through the solution steps of a realistic kinematics problem (Fig. 1), and we use this example throughout this paper to clarify several considerations concerning the diagnosis of students' problem solving.

Fig. 1 shows the (*v, t*) diagram, depicting velocity *v* as a function of the time *t*, part of a kinematics assignment from the Dutch National Examination in 2012: "Aircraft are regularly subjected to severe tests. An example of such a test is the Rejected Take Off test (RTO). During an RTO, an aircraft accelerates to the speed to take off. Then the brakes are applied as hard as possible" (College voor Toetsen en Examens, 2012). Students must answer the question: "Determine the acceleration on time t = 10 sec."

Physics is not just applying mathematical tricks; students must master deeper physics concepts to be able to correctly solve problems. They should then formulate a mathematical procedure, recall appropriate formulas, and perform the calculations to solve the problem by rearranging the equations of the mathematical problem and gaining insight into the usefulness of the rearranged equations for the application. After working out a mathematical solution, students have to transform a context bound solution. At last students have to answer a correct conclusion to the question, according to the modelling cycle (Freudenthal, 1978; Treffers, 1987; Greefrath & Vorhölter, 2016; PISA 2012, 2013). Teachers must be aware of the cognitive steps students have to take, starting with physics concepts, interpreting them mathematically, and transforming this again (back) into physics answers. With regard to the subject of kinematics, science students must overcome at least two cognitive barriers: (1) choosing which of the elements of kinematics correlates to the problem and (2) selecting the right relation among different formulas and combinations of relationships in the text and diagrams (Doorman, 2003). Because of the multiplicity of actions and choices required during problem solving, students often experience high cognitive loads with regard to these barriers (Ashman & Sweller. 2023; Sweller et al., 2011) and need time to develop insight into the differences across elements (Doorman & Gravemeijer, 1999; Fischer, 1980; Smith & Thelen, 2003), and they are prone to develop misconceptions and misinterpretations (Lingefard & Farahani, 2018). The mentioned educational problems listed above, indicate the need to take these factors into account when teachers diagnose written responses on tests about the subject of kinematics to answer the first question: 'Which cognitive steps and activities need teachers monitor to help students successfully solve physic problems?'

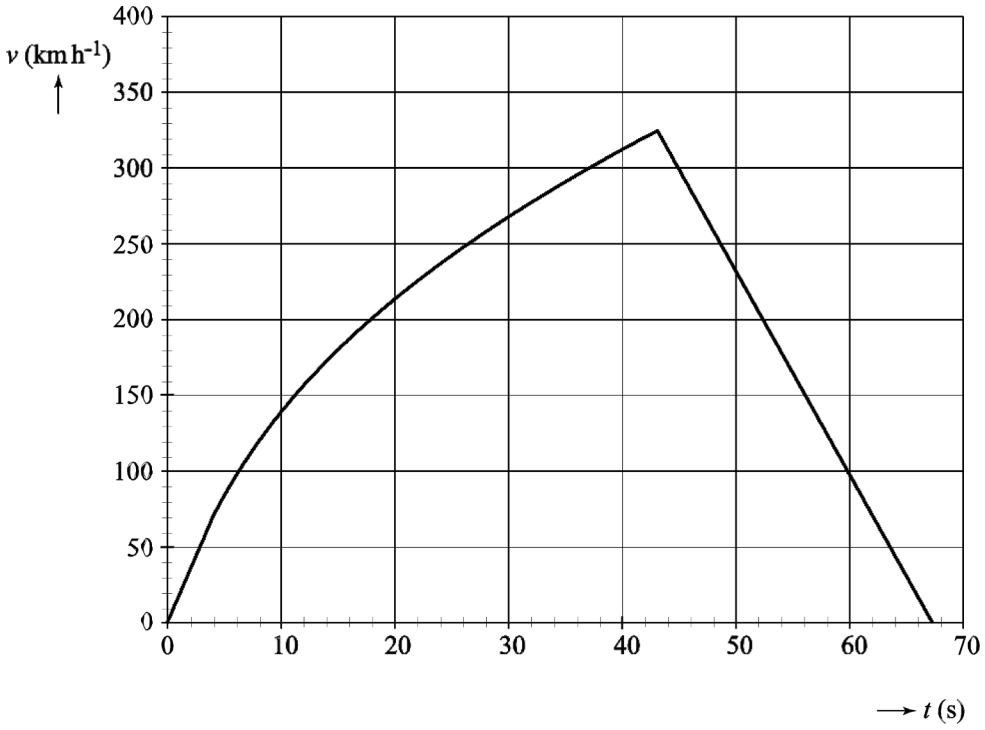


Fig. 1. (v,t) diagram of a RTO test (College voor Toetsen en Examens, 2012).

2.4. Phase 2: The design of a diagnostic instrument to assess students' cognitive problem-solving capacities

The results of the first phase of our study indicate that during problem solving, students' approach is reflected in three sequential episodes: reading, analyzing the task based on the retrieval of the appropriate concepts, and linking these to mathematical operations to produce an answer based on adequate calculations. Successful completion of these episodes taps on specific cognitive activities amongst others are described in the Skill Theory by Fischer and Bidell (2006).

Skill Theory is part of the Dynamic Systems approach; it formulates and assesses students' capabilities (and skill characteristics) and provides a useful framework for developing a diagnostic instrument that acknowledges the specific cognitive activities in which students need to engage during each of the three crucial episodes. According to Chi et al. (1981): "This knowledge base is arranged around 'problem schemata', each of which contains information necessary to solve a specific category of problems." The quality of the schemata that students generate while working on specific science problems can be observed and measured through short- and long-term developments (Fischer, 1980; Fischer & Bidell, 2006; Steenbeek & van Geert, 2020; van der Steen, 2014). while problem solving. According to van Geert (1994), "One of the advantages of a Dynamic Systems approach is that it offers a variety of different representational formats and a very liberal approach to how they should be used."

Whereas Fischer (1980), de Bordes et al. (2019) developed a hierarchical structure to determine a skill hierarchy of eight stages in students' progress toward the mastery of necessary actions and cognitive capabilities, we propose monitoring physics students' proficiency in solving problems by using a three-tier hierarchy that represents the three core episodes (Table 1). The abbreviation SR (self-generated representation) in this table refers to students' expressed thoughts in visual and mental problem representation, their selections of formulas appropriate to problem representations, and their use of appropriate calculations and validity checks (Wittrock, 1989; Fiorella & Mayer, 2016).

The sensorimotor level of insight covers encoding and transformation of information necessary to draw a graph that adequately represents a problem visually through coordination of reading (eyes) and writing or drawing (hand). For example, students must read an assignment accurately, then draw an appropriate line or select the proper information (Schwartz & Fischer, 2005). The representational level of insight relates to adequate deployment of students' declarative, procedural, and conceptual knowledge of the content of science (e.g., kinematics). In this case, students should recognize the nature of a problem, remember appropriate mathematical concepts and formulas, and select a proper formula connected to the tangent in the problem, as indicated by Fig. 1. Finally, the abstract level of insight relates to students' abilities to elaborate, calculate, and solve problems mathematically, such as choosing an appropriate formula and using maths correctly.

Originally according to Fischer (1980) and de Bordes et al. (2019), the first two categories of levels of insight shown in Table 1 were subdivided into three cognitive concepts: action, mapping and system (Table 2). The later category is called single abstractions according to Kisteman & Deutekom (2015) and van Vondel et al. (2018). So, there were 7 levels of insight to assess students' cognitive changes in solving science problems.

For this study, we have expanded the levels of insight in the abstract category from 7 to 11 levels of insight (Table 2) to design an assessment instrument aligned to specific cognitive activities corresponding with three crucial episodes during physics problem solving. The expanding is conforming the arithmetic procedure to solve maths problems.

That is, in the three columns in Table 2, the three tiers of cognitive episodes are divided into 11 levels of insight. Each level of insight takes a number (n), and a description. Sensorimotor and representational categories are divided into action, mapping, and system sections. Each action is a *single* comment that students report in their written work, whereas mapping represents comments that *relate* to each other. The system reflects the merging of all comments. The abstract section implies the completion of mathematical actions that lead to the solution of a physical problem.

In line with Fischer (1980) and de Bordes et al. (2019), we propose connecting levels of insight to students' abilities to solve problems and as Table 2 shows, to facilitate coding by teachers. For instance, level of insight 4 represents students' abilities to connect actual information to their prior knowledge (Mayer, 2014); students need to use what and where functions, which are located in different cognitive subsystems (Schnotz, 2002, 2014). These functions can support students' understanding and evoke associations with specific features in assignments that help them to choose possible heuristic ways to solve problems mathematically (level of insight 7). To use effectively students' prior knowledge, the associated "pathways" must be activated (Shuell, 1988; Lovell & Sherrington, 2020). We propose that the combination of such actions is necessary to connect with the abstract levels of insight, that is, level of insight 7–11. The concepts of the mathematical heuristic path are situated in these abstract levels of insight.

This study focusses to interpret and diagnose students' written tasks of kinematics with a corresponding level of insight while solving problems and proposes a cognitive diagnostic instrument that provides science teachers with a framework to assess students' levels of insight. The original model (Yan & Fischer, 2002) consists of 8 levels of insight in mathematical procedures; we adjust it to

Table 1
Code Table Developed by Fischer (1980) and de Bordes et al. (2019) Containing Three Tiers of Insights, with Numbers and Descriptions of Observations.

Level of Insight	n	Description of Observation
Sensorimotoric	1	SRs* that link visual information of the task
Representational	2	SRs that link visual information of the task to unobservable relations (thoughts)
Abstract	3	SRs that link to formulation of mathematical expressions

Notes. SRs* means: Self-generated representations.

Table 2
Code Table Developed by Fischer (1980) and de Bordes et al. (2019), Adapted to Include 11 Levels of Insight (Capability).

Level of Insight	n	Description of Observation
Sensorimotor Action	1	Signs of delineation, e.g., marks, symbols, or drawings, which cannot be coded in terms of levels of insight
Sensorimotoric Mapping	2	SRs* visualizing information of the task without relation and/or delineation between objects, or with relation without causality
Sensorimotor System	3	SRs visualizing information of the task with two or more relations and/or delineation between objects or relations with causality
Representational Action	4	SRs linking to not observable relations and/or a link between a SR and a formula
Representational Mapping	5	SRs linking to unobservable relations and/or a link between a SR and a formula (e.g., s = vt)
Representational System	6	SRs indicating the understanding of the solution method
Single Abstraction	7	Student shows understanding the regularity/pattern/choice of an adequate formula
Multiple Abstraction	8	Student shows understanding in correct applying the formula
Abstract Conducting	9	Student shows understanding in calculating a formula
Conducting	10	Student produces an answer
Abstract Mapping	11	Student correctly relates the answer to the assignment context

Notes. SRs* means: self-generated representations, and n = level of insight.

include up to 11 levels. Solving physics problems demands more than only mathematics, students must recognize the nature of the physics problem and formulate adequate answers, to reach level of insight 11.

As described by Fischer & Bidell (2006), the level of insight students can reach *without* help by teachers can be seen as the *functional* level of insight. The scope of this study is to determine the "where" the students' knowledge "ended" in the hierarchical scale; to determine this functional level of insight. The functional level of insight can be for instance level of insight 6; the student has not accomplished a task and based on the determined functional cognitive level of insight; teachers could target support students to reach the *optimal* level of insight (level of insight 11).

Determining the functional cognitive level of insight by applying the diagnostic instrument (Table 2) is different to the instrument and use of the developed 'rubric' by Reinhard and colleague's (2020, p. 010109-4), and had to be addressed. The theoretical basis, goal, and use are essentially different to our research. Because the approach of the research and principles of Reinhard and colleagues (2020) differ greatly from our research is it important to name the differences. We will do so in the discussion.

But the conclusion of their research to ours is the same: teachers should actually apply formative assessment and provide appropriate feedback as early as possible during a lesson series to instill domain-specific knowledge for being able to solve problems.

2.5. Phase 3: Research on the measurement qualities and practicality of the assessment instrument

2.5.1. Design

We instructed three experienced science teachers to apply the formative assessment instruments. We started the instruction with Table 1 (consisted in three episodes), and refined the instruction with Table 2, to determine students' level of insight in problem solving kinematics tasks. After the conducted experiment with the participants, we analysed data of teachers' agreement about the students' levels of insight across the nine assignment elements in the domain of kinematics and used Krippendorff's alpha to conclude the reliability and validity of the instrument. By examining the variation in students' level of mastery on problems of different complexity, the practicality of the instrument was assessed.

2.5.2. Participants

We conducted our study in a school in the rural north of the Netherlands. A total of 16 eleventh-grade pre-university students from one physics class participated in the study. The educational background of these students—nine girls ($M = 16.7$; $SD = 1.32$) and seven boys ($M = 17.1$; $SD = 0.35$)—was diverse, because in the Netherlands students can choose several possible courses in Grade 11 in secondary school. The students were familiar with the subject of kinematics; they had received instruction in the domain and taken a test less than two years earlier, in which they examined all components of the domain. Because the students did not receive prior information about the science topic, they could not prepare themselves. The only information students received was that the test consisted of assignments to prepare them for their final exam.

2.5.3. Teachers

Three teachers, aged 59, 62, and 67 years, each of whom had taught science for over 30 years, assessed the students' work independently, using the diagnostic instrument. They were authorized and professionally skilled to teach pre-university science students. Two teachers (B and C) received individual training of 30 minutes in length from the researcher (teacher A) and reviewed students' work independently.

2.5.4. Materials

The study domain is kinematics. To limit the diversity of elements in this subject domain, the test consisted of 21 assignments pertaining to motion, without the component of force. The 21 assignments were distributed evenly over nine categories of kinematics subjects. The difficulty of the test was similar to that of the national final examination.

2.5.5. Procedure

The test was administered to students in two sessions of 30 minutes, separated by a 5-minute break. For each assignment, students were required to write down their solution approach and calculations. Because the test was embedded in regular physics lessons, students received some explanation and answers about the tasks after each session. All students completed all the tasks (21). During test sessions, students were not allowed to ask questions or give verbal or non-verbal reactions, but they were permitted to use a science formula book (Verkerk et al., 2004).

2.5.6. Instructions

After a training session of 30 minutes by the researcher (teacher A), two teachers (B and C), working separately, applied the code table. One teacher (B) and the researcher (teacher A) coded the written solutions to the 21 assignments made by 16 students (total of 336 assignments). One teacher (C) coded 144 assignments. To apply the code table, in 3 tiers or 11 levels of insight, the teachers received the following instruction: "Determine the highest level of capability (functional level of insight) the student has worked out in this assignment and use the code tables, first Table 1, and then Table 2."

To prepare the teachers mentally to code the three tiers (Table 1), they were prompted with a question that reflects the nature of the problem: "In which area of the three levels do you think the student is 'located'/present?" In accordance with Fischer's (1980) procedures, the teachers had to observe the following decision rules:

(1) The highest level of insight, according to their judgment of the task, was the level of insight (capacity).
(2) The review of an upper level of insight includes lower levels of insight.
(3) On tasks that demand no calculations but only need to be explained or read, students receive a score of 11 if they understand, find, and use the basic elements and combined them into an adequate conclusion. When the same task is a part of or sets up an assignment, the code must be seen from the perspective of the overall assignment; if students have succeeded so far, the code must be level of insight 2 (at that moment in the conducting of the overall task; see decision rule 1).

2.5.7. Data analysis

To examine the overall reliability of the outcomes of the diagnostic instrument to determine students' cognitive functional level of insight, we assessed the inter-rater reliability of the three teachers using Krippendorff's alpha test (Hayes & Krippendorff, 2007), at the ordinal level of measurement. The test involves 11 levels of insight and 3 tiers. We assessed the validity of the instrument for aiding teachers' instruction practices by investigating the diversity in students' results across the nine categories.

3. Results

Science education is dealing with two different problems in the formative period: teachers are confronted by students' content gaps, but they may lack time and know-how to determine in depth students' cognitive level of insight in paper tests. Second, science students' need in additional tutoring to understand their own daily learning process. In this study, we try to address these different problems to help science teachers by analyzing students' test solution approaches, and to determine their functional level of insight, by designing and evaluating a practical diagnostic instrument. The goal is to investigate the reliability and the validity of the instrument, and if so, to describe the usefulness as formative assessment tool.

3.1. Assessment accuracy of students' functional level of insight on the test

Our second research question concerns the objective measurements of students' functional level of insight: '*Can the cognitive diagnostic instrument of the code table reliably identify students' functional levels of insight?',* the answer is: yes.

We first investigated the interrater reliability as a whole, by summarising teachers' overall inter-rater reliability about students' levels of insight when solving the 21 assignments connected to the nine problem elements of kinematics using Krippendorff's alpha for 3 tiers and 11 levels of insight.

Table 3
Krippendorff's Alpha Reliability Estimate, Mean, and SD, Across 11 Levels of Insight, Related to Nine Elements of Kinematics.

n	Elements	Krippendorff's alpha	Mean	SD
1	Slope interval $a_{average}$	.88	9.36	3.32
2	Slope one-point a_{max}	.83	8.51	3.07
3	Slope one-point $v \rightarrow a$*	.89	6.83	3.87
4	Surface Interval $v \rightarrow s$	.90	6.43	3.64
5	Read out	.70	9.27	2.98
6	Slope one-point $s \rightarrow v$	.93	8.51	3.61
7	Slope one-point v_{max}	.83	7.48	3.49
8	Explain	.85	9.64	3.06
9	Slope interval $v_{average}$	.97	8.91	3.10

Notes. * means: The assignment depicted in Fig. 2 belongs to this element.

The overall Krippendorff's alpha turns out to be .84 for the ordinal level of measurement at the 11 levels of insight, indicating strong inter-rater reliability. When the teachers used the code for 3 tiers, the Krippendorff's alpha value indicates acceptable inter-rater reliability in their cognitive diagnoses of students' developmental levels (79).

3.2. Assessment accuracy of students' levels of insight for each of the nine elements

Table 3 shows the Krippendorff's alpha values for nine separate elements in the field of kinematics included in the cognitive diagnostic instrument. The values indicate high agreement among the coders regarding students' mastery of different levels of insight in each category.

According to the average ratings of students' performance in nine problem elements, variation exists among students with regard to their levels of insight and the help they need in this class. As we expected, the means of the ratings of students' performance on assignments in elements 5 and 8 indicate that for these elements, students achieve the highest mean levels of insight. However, the standard deviations indicate that even for these relatively easy tasks, considerable variation arises among students' performance in all elements (and cognitive levels of insight). The means further show that tasks that demand calculation of the "slope in one point ($v \rightarrow a$)" (Fig. 1) and "surface interval ($v \rightarrow s$)" belong to the most difficult task elements. Students experience relative much cognitive barriers to solve correctly these tasks. The mean of students' level of insight of the task to calculate the slope in one point is the lowest (6.83) and the SD is the highest (3.87). The task to calculate the surface under an interval (M=6.43, SD=3.64) have nearly the same noteworthy results. Or, in other words, these two elements need special attention relative to other elements.

3.3. Practical validity for determining help needed by students

Using teachers' inter-rater scores on the three-tier levels of insight, we created a table for nine elements of kinematics (Table 4) that summarizes the nature of help students would need from their teachers to remedy their solution methods. As Table 4 shows, in most problem elements, a substantial number of students need additional tailored instruction.

These results indicate that teachers must be aware of the need to tailor their support carefully in each kinematics element. Students need support for their sensorimotor skill levels (i.e., making correct drawings) more than they need help with the representation and abstract categories though. The need for teacher support is evident: The bottom row of Table 4 shows that in each of the three cognitive tiers, nearly 70% of students needed help in at least one of the nine elements. The results related to the assignments "slope one point ($v \rightarrow a$)" (Fig. 1) and "surface interval ($v \rightarrow s$)" also confirm the predicted difficulty levels; the results show the lowest "no help needed" values. This table provides an informed basis for teachers to decide how to address students' diverse needs (finding the "where" of the gap in knowledge), such as an individual student or a group. To answer research question 3: '*Does the instrument facilitate teachers in providing differentiated instruction depending on students' functional level of insight as regards different problem elements in the kinematics domain?',* the answer is yes.

4. Discussion

The motivation of this study is to enhance teachers' awareness of students' levels of performance in solving specific kinematic tasks by using the developed diagnostic instrument (coding scheme in three episodes) and to reinforce their sense of the cognitive levels of insight that students have reached ("Zone of Proximal Development"; Vygotsky, 1978). This study presents the process of developing the instrument, in which the decision rules play a major role to achieve greater uniformity in assessing students' cognitive level of

Table 4
Mean and SD, Percentage of Help Needed and Percentage of No Help Needed in Three Tiers Coding of Elements of Kinematics.

	Descriptive three-tier coding		Help needed starting on level of insight in percentages			
Elements of Kinematics	Mean	SD	Senso-motoric	Representa-tional	Abstract	No help needed
Slope interval $a_{average}$	2.68	.67	18.8%	3.1%	3.1%	75.0%
Slope one-point a_{max}	2.51	.66	25.0%	18.8%	3.1%	53.1%
Slope one-point $v \rightarrow a$ *	2.22	.85	35.4%	18.8%	14.6%	31.3%
Surface Interval $v \rightarrow s$	2.26	.77	31.3%	28.1%	21.9%	18.8%
Read out	2.66	.62	14.6%	12.5%	2.1%	70.8%
Slope one-point $s \rightarrow v$	2.55	.77	18.8%	10.4%	14.6%	56.3%
Slope one-point v_{max}	2.45	.70	18.8%	28.1%	28.1%	25.0%
Explain	2.70	.66	14.6%	4.2%	0%	81.3%
Slope interval $v_{average}$	2.59	.69	18.8%	12.5%	6.3%	62.5%
Overall percentage			21.7%	14.6%	10.1%	53.6%
Percentage of students who need help in at least one elements of kinematics in each of the three tiers			75%	93%	69%	

Notes. * means: The assignment depicted in Fig. 2 belongs to this element.

understanding (Doyle & Ponder, 1978). At first sight, the instrument may seem complex for teachers, because students can make more than one mistake in one assignment. Considering students' descriptive answers, teachers also might note disharmonies between students' mathematically correct answers to assignments and the absence of internalization of the subject's concept (Sherin, 2001). If teachers are not aware of systematic determination of students' problem solving, they also might differ in their starting points of assessment. Refining teachers' training can reduce these differences, if they use this tool again and again in every task. Fortunately, the matter can be simplified, because in essence, there are only three cognitive categories (episodes) that teachers need to consider consecutively when assessing a student's work: (1) Does the student read and draw well? (2) Does the student remember well? and (3) Does the student calculate well? And in this order teachers need to decide on what would be appropriate feedback.

On an organizational level, application of this instrument to Lesson Studies might encourage teachers and colleagues to enhance their didactic attitudes (Lijnse, 2002). In general, causes of differences in inter-rater reliability (Opposs & Mapp, 2012)—often reflected in coding values that are farther or closer to each other—can be discussed (as shown in the assessment of Student I in 'Limitations and practical considerations of this study', which in turn should improve teachers' educational processes and scaffolding proficiency. It would be interesting to investigate the effect of teachers using this cognitive diagnostic instrument, compared with their own intuitive methods to assess students, as well as to investigate the potential improvement of science students' self-efficacy in problem solving (Britner & Pajares, 2006).

4.1. Comparing summative assessment by marks and formative assessment by the diagnostic instrument

To investigate students' level of problem-solving (according to the curriculum) often a summative instrument is used. A representation of a summative instrument is for instance developed and used by Reinhard and colleagues (2020) in an illustrative argument and enables us to clarify the difference between the differences of the summative approach of the research of Reinhard and colleagues (2020) and the formative approach in assessment of our research.

The first one is the theoretical foundation. The theoretical base of the 'rubric' of Reinhard and colleague's (2020) 'was created to mirror the aspects of expert like problem solutions enumerated in the literature'. The theoretical base of the diagnostic instrument is grounded in the Dynamic Systems Theory and Skill Theory of Fischer and Bidell (2006).

The second difference is the goal of the research. The goal of the investigation of Reinhard and colleague's (2020) is to 'perform a controlled study of students' use of expert problem-solving strategies on final exam problems' (with and without intervention). Our goal is to pinpoint student's level of insight in the process of problem-solving in the formative period and the validation of the used instrument and not to determine 'the problem-solving best practises or strategy' (by marks).

The third difference is about the categorisation and the applying of the 'rubric' and the diagnostic instrument. The 'rubric' developed and applied by Reinhard and colleagues (2020) is formulated in the same way as for the assessment of summative tests (as in many courses). The summation of scores indicates students' use of problem-solving best practices; a maximum of three scores in their investigation are assigned to (almost) each component of the described numbering of the task analysis. A consequence of this summation is that by these scores only a number is created. That means, it is not possible to interpret where the student (exactly) is in the process of the development of problem-solving. To give an example: A student can be graded 8×2 pts is 16 pts. According to the interpretation of the 'rubric', the student has satisfactorily completed the assignment. Another student grated 5×3 pts and 1×1 pt is 16 pts - has written down the correct concept (formula), but is not able to proceed in the mathematical elaboration. In short, this is the consequence in summative assessment: The product (16 pts) is the same, but the process differs markedly.

Our presented instrument is a hierarchical model subdivided into eleven cognitive levels of insight, which teachers can use practically as a formative assessment model to determine the *functional* cognitive level of insight, and as shown above: that is not attainable with a survey like that of Reinhard et al. (2022). May be, there is a contra dictionary in the use of the 'Force Concept Inventory pretest score as predictor' of that model as formative assessment model and the applying of summative assessment.

Despite the essential differences between the 'rubric' and the diagnostic instrument, the conclusion of both of our studies is the same: The necessity of formative assessment and providing appropriate feedback during a lesson series to instill domain-specific knowledge for being able to solve problems.

Our study offers some support for using the proposed diagnostic instrument as a tool for formative cognitive assessments in other subjects in which mathematics are applied, such as economics, biology, and chemistry. Generally, for science, technology, engineering, and mathematics (STEM) subjects, the instrument should lead to integrated formative didactics of disciplines to support students. The skills of recognizing the similarity of problems in three core sets of activity and solutions in another subject (transfer) are essential to problem solving (Whitelegg & Parry, 1999; Bati, 2022). It would be valuable to conduct further investigations of the contribution of this instrument with regard to the transfer of subjects in science education.

4.2. Limitations and practical considerations of this study

The use of the diagnostic instrument requires teachers to ask students to draw a graph of each problem and check its appropriateness for solving a specific problem. After checking a student's understanding of the nature of the problem, the teacher is recommended to check students' mathematical representation (e.g., linkage to a correct formula) and thirdly to identify students' use of correct mathematics, with the instructions in mind (Tables 1 and 2). The highest-identified functional level of insight can be considered as the beginning of the Zone of Proximal Development (Vygotsky, 1978; Fischer & Bidell, 2006) and is therefore the starting point for teachers' support. Although we established a high interrater reliability, the Krippendorf's alpha's indicate that there is still room for improvement of the diagnostic instrument as a means to support students' development. In the next paragraphs, we will therefore

investigate to what extent different ratings of teachers have serious consequences for what we consider adequate support to improve students problem solving. We addressed the consequences of different ratings of determinations in the paragraphs 'Considerations concerning the accurate determination of the feedback students need' and 'Considerations concerning the accurate determination between two results of one student'. The consequences show the need for sharpening the decision rules in the categories and tiers to increase the inter-rater reliability even more, in future surveys. The intention of this paper by developing an instrument to determine students' cognitive level of understanding is to determine the 'where' of the starting point of support for science teachers, and its implications.

For future use, the instrument can be improved by qualitive research of teachers' decisions of determinations (Doyle & Ponder, 1978). To augment drawing of provisional conclusions of the findings of this study, the instrument, with all its limitations and considerations, such as a small number of students and the application of greater variation across students, becomes the subject of future research. This will include elaborated teacher training for instance by Lessons Study or by qualitive research of teachers' decisions of determinations (Doyle & Ponder, 1978). In future research, teachers and researchers can practically and theoretically apply this instrument as formative testing, with and without interventions, to gain insight into the cognitive process of learning of students.

4.3. Considerations for further improvements concerning the accurate determination of the kind of feedback students need

Although teachers in general show high agreement in their rating of students' understanding of the steps students apply correctly when solving any of the 21 assignments, in seldom cases teachers' rating differed on a particular assignment. This obviously has consequences for the nature of the feedback a teacher would consider appropriate to provide to a student. We address this issue by examining teachers' ratings of the written answers on one of the 21 assignments, the RTO assignment. Table 5 shows teachers' ratings of students' understanding of the application of correct steps in solving this specific assignment. This table reveals that all three teachers disagree on only one student, namely student I. We discuss here the consequences of this difference for the nature of the feedback the teacher would provide to this particular student.

To solve the assignment, student I needed to draw a (correct) tangent line along the curve at a time of 10 seconds (see Fig. 2 for the correct tangent line); this sensorimotor motion (reading and drawing) is the beginning of the solution of the assignment, and the student must make connections to (prior) knowledge to produce this line. Student I did not produce the line.

Nor did Student I draw the two points that are necessary to produce the next steps, by drawing a triangle and choosing a formula to start calculating. The student displayed only the algebraic calculation (see Fig. 3) and the abstract effect of editing the problem. Although it was obvious that Student I knew what to do and how, the student did not show the points of reference to calculate. Therefore, at this point, what is the functional level of insight (and the starting level of support)?

Following the instructions we provided to teachers, the code should be level of insight 1: "Student I produce no sign (drawing)." In this case, the prescribed feedback to the student would be to draw the situation to understand the problem type at hand. Since two of the three teachers offered codes of levels of insight 6 and 8, they would offer feedback on the selection of the correct formula. This example shows that coding did not always result in the proper diagnosis of a students' level of understanding based on our instruction to the teachers, even though the results of this study show high overall inter-rater reliability. A possible explanation for this discrepancy may be that the coding depended on what the teachers read explicitly or implicitly in the descriptive answer of mathematics (Fig. 3) and correlated with the instructions. To avoid the aforementioned discrepancies in coding, we suggest to aid teachers by including more critical examples for deciding whether students need feedback in drawing or selecting a specific formula .

Table 5
Codes by Three Teachers for the RTO Assignment (Fig. 2), Mean and SD.

	Levels of Insight					
	3 Tiers			11 Levels		
Student	Teacher A	Teacher B	Teacher C	Teacher A	Teacher B	Teacher C
I	3	2	1	8	6	2
II	3	3	3	9	8	10
III	1	1	1	2	2	2
IV	3	3	3	11	11	11
V	3	3	3	9	8	10
VI	3	3	3	11	11	11
VII	1	2	1	3	6	3
VIII	1	1	1	3	3	2
IX	3	3	3	11	11	11
X	1	1	1	0	0	0
XI	3	3	3	8	9	9
XII	3	3	3	8	9	10
XIII	3	3	3	11	11	11
XIV	3	3	3	11	11	11
XV	3	3	3	11	11	11
XVI	1	1	1	2	3	2
Mean	2.38	2.38	2.25	7.38	7.5	7.25
SD	.96	.89	1	3.96	3.74	4.4

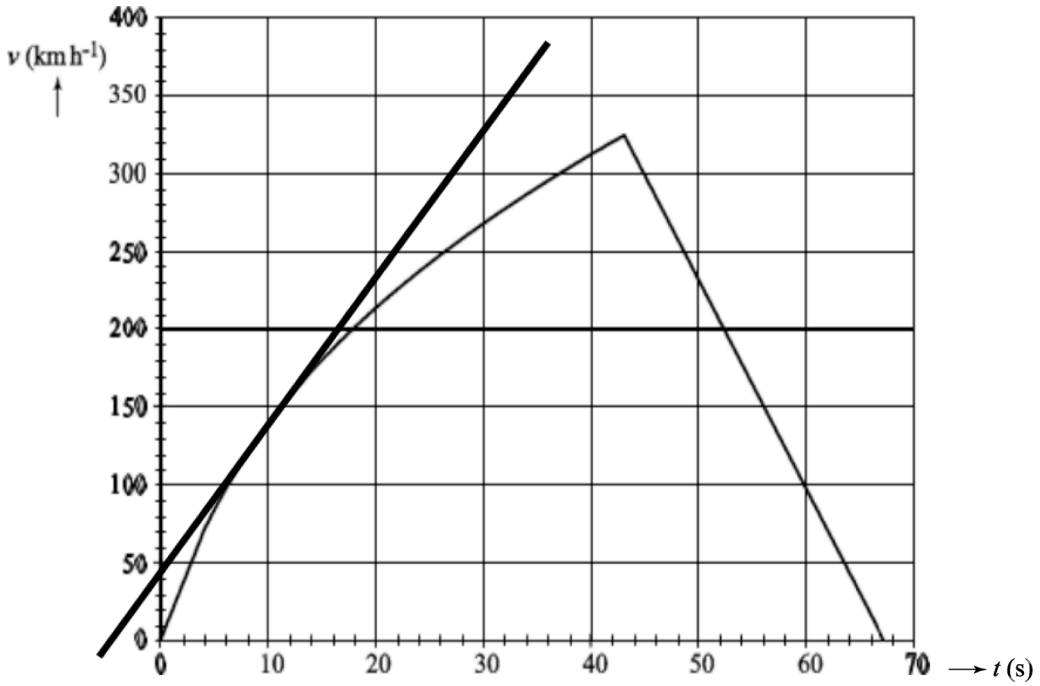


Fig. 2. The (v,t) diagram of an RTO test, from the Dutch central examination in senior secondary school (HAVO) physics in 2012), with a tangent line at t = 10 sec., as an example of a possible first descriptive action of a student's answer of the RTO assignment.

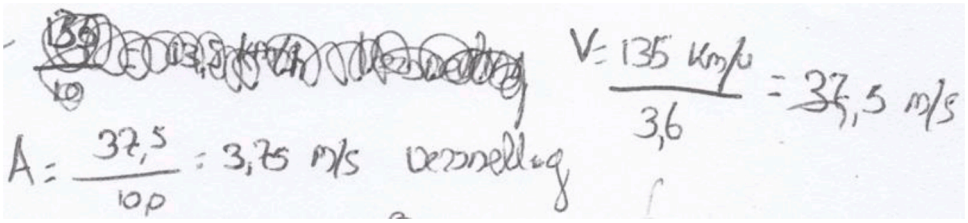


Fig. 3. Answer by Student I, who did not draw a tangent line as shown in Fig. 3.

4.4. Considerations concerning the accurate determination of students' level of understanding of a specific element of kinematics problems

In this section, we address the consequences of the scores assigned by teachers to the written answer of one student (student XVI, Table 5) on *two* assignments of the same kinematics element.

We are aware of the limitations to formulating general conclusions, but the example can provide guidelines to further research for pedagogical guidance. The results of determination of student XVI of the RTO assignment are showed in Table 5.

The second assignment is too a task over acceleration. The student has to answer a question in a diagram similar as of the RTO assignment. Here, the question is: 'Is the movement accelerated or decelerated between two seconds (t=0 sec. and t=30 sec.). *Explain* your answer.'

Students' results on the first assignment are recorded as levels of understanding: 2, 3, 2 (Table 5). The results of the second assignment are recorded as levels of understanding: 11, 5, 10. Comparing the results on both assignment concerning the same kinematics element, the teacher could conclude that student XVI can explain in rather good technical jargon the movement of acceleration, but has not the capability to solve the problem of acceleration in one moment of time. This exemplifies the discrimination capability of this instrument for identifying specific student weaknesses in solving physics problems.

5. Conclusion

We conducted an investigation in three phases to development of a formative assessment to determine students' need for corrective actions in physics: Defining the challenges of teachers, the design of a diagnostic instrument and the research on the measurement's qualities. Therefore, we analysed students' problem-solving approaches in the kinematics domain, according to the modelling cycle, which resulted in addressing of three episodes of students' cognitive activities. After that, we developed and implemented a cognitive

diagnostic instrument based on Fischer's (1980) and de Bordes et al. (2019) code tables that science teachers enable a useful and specific assessment of paper tests. This study shows that the diagnostic tool has acceptable inter-rater reliability to allow teachers to objectively assess students' progress in mastering the three cognitive levels of understanding and the 11 cognitive levels of the Competency Scale. The diagnostic instrument thus allows science teachers to identify essential gaps in students' knowledge and to provide appropriate feedback to address students' "Zone of Proximal Development". However, despite the acceptable reliability, our research also shows that there is still room for improvement regarding the instructions to teachers how to identify the exact nature of student failures in solving closely related problems and to determine the type of feedback that helps students improve their problem solving (see Limitations in the Discussion section).

The need for teachers' support is evident: in each of the three cognitive tiers, nearly 70% of students in the current study needed support in one of the nine kinematic elements. The value of the instrument is confirmed by our finding that students' levels of insight vary in almost every problem category (Table 4), and so —although the item of feedback is beyond the scope of this paper— do the types of feedback and help they need to improve their solution methods. And because of that, science teachers must be professionally strong in both subjects: physics and mathematics (Bati, 2022). Nonetheless, this study provides strong reasons to continue testing this approach to identify novel, additional ways to implement cognitive diagnostic and formative assessments in other areas of science education (i.e., knowledge transfer) to identify students' functional level of understanding.

CRediT authorship contribution statement

Frits F.B. Pals: Conceptualization, Validation, Formal analysis, Investigation, Writing – original draft, Visualization. **Jos L.J. Tolboom:** Conceptualization, Methodology, Validation, Data curation, Writing – review & editing, Supervision, Project administration. **Cor J.M. Suhre:** Conceptualization, Methodology, Validation, Formal analysis, Writing – review & editing, Supervision.

Data availability

Data will be made available on request.

References

Ashman, G., & Sweller, J. (2023). What every teacher should know about cognitive load theory and the importance of cognitive load to instruction. *In C. E. Overson, C. M. Hakala, L. L. Kordonowy, & V. A. Benassi (Eds.), In their own words: What scholars and teachers want you to know about why and how to apply the science of learning in your academic setting (pp. 185-195). Society for the Teaching of Psychology.* https://teachpsych.org/ebooks/itow, 2006.

Chapter 2 by Abend, M., Taber, K. S., & Brock, R. (2018). *Effective teaching and learning: perspectives, strategies and implementation.* New York: Nova Science Publishers, Inc.

Bati, K., & Rezaei, N. (2022). Education of integrated science: Discussions on importance and teaching approaches. eds. In *Integrated Education and Learning. Integrated Science, 13*. Cham: Springer. https://doi.org/10.1007/978-3-031-15963-3_19.

Britner, S. L., & Pajares, F. (2006). Sources of science self-efficacy beliefs of middle school students. *Journal of Research in Science Teaching, 43*(5). https://doi.org/10.1002/tea.20131

Chi, M. T. H., & Vosniadou, S. (2013). Two kinds and four sub-types of misconceived knowledge, ways to change it, and the learning outcomes. Ed.. *The international handbook of conceptual change* (2nd ed, pp. 49–70). New York: Routledge. https://doi.org/10.4324/9780203154472.ch3

Chi, M. T. H., Feltovic, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121–152. https://doi.org/10.1207/s15516709cog0502_2

Chinn, S. (2020). *More trouble with maths: A complete manual to identifying and diagnosing mathematical difficulties* (3rd ed.). Routledge. https://doi.org/10.4324/9781003017721

College voor Toetsen en Examens. (2012). *Examenopgave HAVO, HAVO natuurkunde, opgave 2*. tijdvak 2. Leiden: Stichting Studie Begeleiding.

de Bordes, Boom, J., Schot, W. D., van den Heuvel-Panhuizen, M., & Leseman, P. P. (2019). Modelling children's Gear task strategy use with the Dynamic Overlapping Waves Model. *Cognitive Development, 50*, 237–247.

Doorman, M. (2003). Inzicht in snelheid en afgelegde weg via grafieken. *Tijdschrift Voor Didactiek Der b Wetenschappen, 20*(1), 1–25.

Doorman, M., & Gravemeijer, K. (1999). Modelleren als organiserende activiteit in het wiskunde.-onderwijs. *Tijdschrift Voor Didactiek Der β Wetenschappen, 16*(1), 38–55.

Doyle, W., & Ponder, G. (1978). The practicality ethic in teacher decision making. *Interchange, 8*(3), 1–12.

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review, 28*, 717–741. https://doi.org/10.1007/s10648-015-9348-9

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87*(6), 477–531.

Fischer, K. W., Bidell, T. R., Damon, W., & Lerner, R. M. (2006). Dynamic development of action and thought. Eds.. *Theoretical models of human development. Handbook of child psychology* (6th ed, pp. 313–399). New York: Wiley

Freudenthal, H. (1978). *Vorrede zu einer wissenschaft vom mathematikunterricht.* München: Oldenbourg.

Greefrath, G., Vorhölter, K., & Kaiser, G. (2016). Teaching and learning mathematical modelling: Approaches and developments from German speaking countries. Ed.. *ICME-13 topical surveys* HamburgCham: University of Hamburg. https://doi.org/10.1007/978-3-319-45004-9_1. Springer

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77–89.

Hsu, L., Brewe, E., Foster, T. M., & Harper, K. A. (2004). Research in problem solving. *American Journal of Physics, 72*(9), 1147–11576. https://doi.org/10.1119/1.1763175

Ince, E. (2018). An overview of problem-solving studies in physics education. *Journal of Education and Learning, 7*(4), 191–200. https://doi.org/10.5539/jel.v7n4p191

Izquierdo-Acebes, E., & Taber, K. S. (2023). Secondary science teachers' instructional strategies for promoting the construction of scientific explanations. *Ski Education*. https://doi.org/10.1007/s11191-022-00412-5

Kisteman C.D., & Deutekom L.V. (2015). The relation between non-verbal behavior, the understanding that children express verbally and their performance in the knowledge domain of science and technology (Dutch: De relatie tussen non-verbaal gedrag, het begrip dat kinderen verbaal uiten en de prestatie in het kennisdomein wetenschap en techniek) (Master's thesis, Utrecht University).

Levering, B. (2010). Disappointment in teacher-student relationships. *Journal of Curriculum Studies, 32*(1), 65–74. https://doi.org/10.1080/002202700182853

Lijnse, P. (2002). Op weg naar een didactische structuur van de natuurkunde? de ontwikkeling van didactische structuren volgens een probleem stellende benadering. *Tijdschrift Voor Didactiek Der β Wetenschappen, 19*(1&2), 62–92.

Lingefjärd, T., & Farahani, D. (2018). The elusive slope. *International Journal of Science and Mathematics Education, 16*(6), 1187–1206. https://doi.org/10.1007/s10763-017-9811-9
Lovell, O., & Sherrington, T. (2020). *Sweller's cognitive load theory in action*. John Catt Educational, Limited. ProQuest Ebook Central. https://ebookcentral.proquest.com/lib/rug/detail.action?docID=6461837.
Maloney, D. P. (2011). An overview of physics education research on problem solving. *Getting started in physics education research*. Fort Wayne: Indiana University Purdue, University Fort Wayne.
Marsh, E., Eliseev, E., Dunlosky, J., & Rawson, K. (2019). Correcting student errors and misconceptions. Eds.. *The Cambridge handbook of cognition and education (Cambridge handbooks in psychology* (pp. 437–459) Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108235631.018
Mayer, R. E. (2014). *The Cambridge handbook of multimedia learning* (2nd. ed). New York: Cambridge University Press.
Molin, F., Cabus, S., Haelermans, C., et al. (2021). Toward reducing anxiety and increasing performance in physics education: evidence from a randomized experiment. *Research in Science Education, 51*(Suppl 1), 233–249. https://doi.org/10.1007/s11165-019-9845-9
Opposs, D., & Mapp, L. (2012). *International comparisons in senior secondary assessments*. Coventry, UK: Office of Qualifications and Examinations Regulation.
Ottevanger, W., Van Oorschot, F., Spek, W., Boerwinkel, D. J., Eijkelhof, H., De Vries, M., Van Der Hoeven, M., & Kuiper, W. (2014). *Kennisbasis natuurwetenschappen en technologie voor de onderbouw vo*. Enschede: SLO.
Paas, F., & van Merriënboer, J. J. (2020). Cognitive-load theory: Methods to manage working memory load in the learning of complex tasks. *Current Directions in Psychological Science, 29*(4), 394–398.
PISA 2012. (2013). *Assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris, France: OECD Publishing. doi.org/10.1787/9789264190511-en.
Powell, A. B., Borge, I. C., Fioriti, G. I., Kondratieva, M., Koublanova, E., & Sukthankar, N. (2009). Challenging tasks and mathematics learning. *Challenging mathematics in and beyond the classroom: The 16th ICMI study*, 133-170.
Pol, H. J., Harskamp, E. G., Suhre, C. J., & Goedhart, M. J. (2008). The effect of hints and model answers in a student-controlled problem-solving program for secondary physics education. *Journal of Science Education and Technology, 17*(4), 410–425. https://doi.org/10.1007/s10956-008-9110-x
Reinhard, A., Felleson, A., Turner, P. C., & Green, M. (2022). Assessing the impact of metacognitive postreflection exercises on problem-solving skillfulness. *Physical Review Physics Education Research, 18*(1), Article 010109.
Ryan, V., Fitzmaurice, O., & O'Donoghue, J. (2022). Student interest and engagement in mathematics after the first year of secondary education. *European Journal of Science and Mathematics Education, 10*(4), 436–454. https://doi.org/10.30935/scimath/12180
Schnotz, W. (2002). Towards an integrated view of learning from text and visual displays. *Educational Psychology Review, 14*(1), 101–120. https://doi.org/10.1023/A:1013136727916
Schnotz, W., & Mayer, R. E. (2014). Integrated model of text and picture comprehension. *Multimedia learning* (2nd. ed, pp. 72–103). Cambridge: Cambridge University Press.
Schoenfeld, A. H. (2016). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics (Reprint) *Journal of Education,, 196* (2), 1–38. Boston: Boston University Wheelock College of Education & Human Development. Sage Publishing. Reprinted with permission from Handbook of Research in Mathematics Teaching and Learning copyright 1992, by the National Council of Teachers of Mathematics.
Schutz, P. A., Hong, J., & Francis, D. C. (2020). *Teachers' goals, beliefs, emotions, and identity development: Investigating complexities in the profession*. Routledge.
Schwartz, M. S., Fischer, K. W., Demetriou, A., & Raftopoulos, A. (2005). Cognitive developmental change: Theories, models, and measurement. Eds.. *Building general knowledge and skill: Cognition and microdevelopment in science* (pp. 157–185) Cambridge; England: Cambridge University Press
Sherin, B. L. (2001). How students understand physics equations. *Cognition & Instruction, 19*(4), 479–541. https://doi.org/10.1207/S1532690XCI1904_3
Shuell, T. (1988). The role of the student in learning from instruction. *Contemporary Educational Psychology, 13*(3), 276–295.
Smith, L. B., & Thelen, E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences, 7*(8), 343–348. https://doi.org/10.1016/S1364-6613(03)0
Steenbeek, H., & van Geert, P. (2020). Education and development as complex dynamic agent systems: How theory informs methodology. *Handbook of integrative developmental science* (pp. 162–188). Routledge.
Sweller, J., Ayers, P., & Kalunga, S. (2011). *Cognitive load theory*. New York: Springer.
Taber, K. S., & Abend, M. (2018). Scaffolding learning: Principles for effective teaching and the design of class-room resources. Ed.. *Effective teaching and learning: Perspectives, strategies and implementation* (pp. 1–43) New York: Nova Science Publishers
Thiede, K., Oswalt, S., Brendefur, J., Carney, M., Osguthorpe, R, Dunlosky, J., & Rawson, K. (2019). Teachers' judgments of student learning of mathematics. Eds.. *The Cambridge handbook of cognition and education (Cambridge handbooks in psychology* (pp. 678–695) Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108235631.027
Tolboom, J. L. J. (2012). *The potential of a classroom network to support teacher feedback; A study in statistics education*. Groningen: University of Groningen.
Treffers, A. (1987). *Three dimensions, a model of goal and theory description in mathematics instruction - the Wiskobas Project*. Dordrecht: D. Reidel Publishing Company.
van Geert, P. L. C., & Butterworth, G. (1994). In the developing body and mind series. Ed.. *Dynamic systems of development. Change between complexity and chaos* (1st ed.). New York: Harvester & Wheatsheaf
van der Steen, S. (2014). *How does it work?": A longitudinal microgenetic study on the development of young children's understanding of scientific concepts*. Groningen: University of Groningen.
van der Zwaart, P. (2007). *Concretisering van de kerndoelen wiskunde*. Enschede: SLO.
van Vondel, S., Steenbeek, H., van Dijk, M., & van Geert, P. (2018). The effects of video feedback coaching for teachers on scientific knowledge of primary students. *Research in science education, 48*, 301–324.
Verkerk, G., Broens, J. B., Bouwens, R. E. A., Groot de, P. A. M., Kranendonk, W., Vogelezang, M. J., Westra, J. J., & Wevers-Prijs, I. M. (2004). *In NVON-commissie (Ed.), Binas havo/vwo* (5th ed.). Groningen: Wolter-Noordhoff.
Vygotsky, L. S. (1978). *Mind and society: The development of higher mental processes*. Cambridge: MA: Harvard University Press.
Whitelegg, E., & Parry, M. (1999). Real-life contexts for learning physics: Meanings, issues and practice. *Physics Education, 34*(2). https://doi.org/10.1088/0031-9120/34/2/014
Wittrock, M. C. (1989). Generative processes of comprehension. *Educational Psychologist, 24*, 34, 376.
Yan, Z., & Fischer, K. W. (2002). Always under construction dynamic, variations in adult cognitive micro development. *Human Development, 45*, 141–160.

Frits Pals works as a researcher in collaboration with the University of Groningen and the Netherlands Institute for Curriculum Development. In secondary education, he taught mathematics, chemistry and physics from 1978 until 2015. His research focuses on improvement of learning by developing learner-generated representations strategies, in memorization, reasoning and transfer of knowledge in secondary science education.

Jos Tolboom works as a mathematics and informatics curriculum developer for upper secondary education at the Netherlands Institute for Curriculum Development. His research interests focus on curricular and didactical aspects of STEM education, with an emphasis on the utilization of technology for mathematics and informatics education.

Cor Suhre is employed as a senior researcher in the teacher education department of the University of Groningen. His research interests include the development of powerful learning environment, the contribution of computer-assisted instruction to improve problem solving in Physics and Mathematics and the professional development of teachers.