Progress in Science Education
Year, Vol. 5, No. 1, 76-93
ISSN 2405-6057
DOI.10.25321/prise.2022.1330

Special Issue
Tasks in Science Education

Researched-based report of practice

Quality of task sets – An instrument for analysing science tasks with different functions along the learning process

Sebastian Stuppan[1,2], Markus Wilhelm[1,2], Katrin Bölsterli Bardy[1]

Received: October 2021 / Accepted: April 2022

Structured Abstract

Background: In competency-oriented education, tasks in science subjects have increased in importance in recent years. In addition, it is suggested that to develop a competency, a set of tasks with different functional embeddings is required. An option for the arrangement of tasks is explained in the Learning Process Model and starts with the so-called "confrontation" task at the beginning of a new topic, challenging learners with a new problem. Next, learners build up the required concepts and skills with the help of development tasks to be able to solve the initial problem, followed by exercises to train and expand the competency. To sum up, learners solve the initial problem in a synthesis task and may then need to apply the competency in a transfer task. Each task type (e.g., confrontation task) is described by the weighting of the following nine scales: 1) chart of competencies, 2) relationship to daily life, 3) learners' conceptions, 4) knowledge, 5) knowledge activities, 6) forms of representation, 7) task openness, 8) learning supports, and 9) learning paths. Each scale contains between one and four subscales that describe the task types by their weighting. The description in the form of scale values is taken up empirically.

Purpose: This study combines existing task scales from research with the different functions of tasks along the learning process model. An expert panel worked in a general-theoretical manner, and trained lecturers rated tasks with the instrument to analyse tasks.

Sample/setting, Design and Methods: In our study, we used the existing scales from the *Instrument to Analyse Tasks* (IAT). To develop the experts' proposed scale values, we consulted four experts. We calculated the AD_M index (average absolute deviation) as a quality-control measure for the experts' level of agreement. According to the tasks' scale values as rated by trained lecturers ($N = 2$), we selected 25 of the 146 science education tasks from the project *MINT*[a] *unterwegs* ("STEM on the move"). In the comparison, we calculated the score differences between the experts' scale values and the rated task.

Results: The results show that it is possible to describe different task types of a task set with the IAT when the scale values are obtained from expert proposals. Moreover, the IAT scale values obtained from the expert proposals are quite similar to those from task ratings by trained lecturers.

Conclusions: This study indicates that experts can distinguish and characterize the Learning Process Model's various task types by weighting IAT scale proposals. Furthermore, it has been shown that tasks can be analysed with the IAT. When the results of the task analysis by trained lecturers are compared to the experts' proposals, the tasks can be revised reasonably and optimized for the learning process.

Keywords: *competency, science, task sets, learning, model, instrument, analysis, task quality*

[a] "MINT" is the German acronym for "Mathematik, Informatik, Naturwissenschaften und Technik", the English version being "STEM" (science, technology, engineering and mathematics). Sample tasks from the project *MINT unterwegs* are available on the website: https://mint-erleben.lu.ch (retrieved 20 January 2022). The project has been running since 2016.

[1]Pädagogische Hochschule Luzern, Schweiz, [2]Pädagogische Hochschule Heidelberg, Deutschland
✉ sebastian.stuppan@phlu.ch

1 Introduction

Competency-enhancing education[b] leads to a new kind of learning. The main objective becomes the acquisition of competencies so that learners can solve specific problems. Therefore, the teaching structure changes, too. Lesson planning starts from the endpoint, the competency to be achieved, and then proceeds "backwards" (Lersch & Schreder, 2013). The questions asked during planning may be: What competency do I want learners to develop? What should learners consequently be able to do and know at the end? How can I subdivide the competency into sub-competencies so that I can gradually develop the competency to be achieved? What teaching materials can I use to develop the competency? How can I recognize and assess learners' performance? Answering these questions helps teachers to focus on acquisition of a sub-competency during long-term planning of a whole competency. Moreover, these questions serve as a guide for teachers, providing orientation and clarity in planning (Wilhelm et al., 2015).

In competency-enhancing education, a single task cannot incorporate several aspects of a whole competency to be built up. Instead, task sets come into focus for teaching and learning (Luthiger et al., 2018), as they can create learning environments and build up subject-specific, interdisciplinary competencies throughout a teaching unit. So, task sets play a central role in high-quality, competency-oriented teaching (Müller & Helmke, 2008).

In this article, task sets designed following the *Learning Process Model* are presented. Numerous Swiss textbooks, such as *NaTech* (Bölsterli Bardy et al., 2017), *Das WAH-Buch* (Wespi et al., 2019) and *Wer ist Landwirtschaft* (Schweizerisches Agrarmuseum Burgrain, 2020), and online learning platforms, such as *entdecke.lu.ch* (DVS, 2016) and *mint-erleben.lu.ch* (DVS, 2018) are already structured according to the Learning Process Model.

Although several research groups have developed and partially tested various category systems for assessing tasks, no empirically tested instrument is available to distinguish different task types (e.g., confrontation task) within a task set. Therefore, we have taken up the nominal settings of an instrument to distinguish tasks in a task set and from it developed an operationalised, empirically tested *Instrument to Analyse Tasks (IAT)* (Stuppan et al., in press). With the IAT, the quality of tasks is described with the following nine scales: 1) chart of competencies, 2) relationship to daily life, 3) learners' conceptions, 4) knowledge, 5) knowledge activities 6) forms of representation, 7) task openness, 8) learning supports; and 9) learning paths.

To use the IAT in the Learning Process Model, however, specific scale values for each subscale to characterize each task type must be obtained. The scale values indicate how strongly a subscale must be pronounced for a task to belong to a certain task type.

2 Research Background

For learners, the relevance of a given competency becomes obvious when it can be used in a particular situation. Consequently, a competency can be developed at school, starting with a situation or a problem that requires the competency to be achieved. At the beginning, learners do not need to solve the problem; however, they should formulate a hypothesis as to how the problem might be solved. Subsequently, the upcoming tasks are set up in such a way that the competency can be built up gradually. At the end, learners should be able to solve the initial problem. Therefore, competency orientation means more than simply learning contents and skills. It means that the learners acquire competencies by thinking and performing flexibly in real-life situations; foundations for this kind of flexibility are laid in school.

In the sense of Weinert (2002), competency-oriented education aims to make learners capable of solving real problems. If the instructional situations required are individually tailored to learners' learning processes, it is competency-enhancing education as an extension of competency-oriented education (Hammann, 2006).

Competency-enhancing education requires intensively working with tasks. Here, tasks can be assigned to multiple functions: they aim to build and expand subject-specific and generic competencies, they structure learning processes and they provide information about learners' competency development (Abraham & Müller, 2009; Luthiger, 2014). However, the current discussion about which tasks are competency-enhancing, and which are not, disregards the fact that the acquisition of competencies demands a complete learning process (2.1 Learning Process Model), takes time, and includes several stages. Consequently, a single task is mostly insufficient for building up a new competency. Instead, several tasks are needed, orchestrating competency acquisition in a logical way. In other words, not one task but a set of tasks forms the basic requirement of learning arrangements (Reusser, 2014, p. 77).

2.1 Learning Process Model

The Learning Process Model contains a set of tasks. The model emerged from the Creative Problem-Solving Model by de Haan (2009), the pedagogical model KAFKA by Reusser (2014) and the Lucerne Model for the Development of Competency-Enhancing Task Sets (LUKAS, Luzerner Modell zur Entwicklung Kompetenzfördernder Aufgabensets) by Luthiger et al. (2018). Models by Oser and Patry (1990) also influenced the creation of the Learning

[b] We use the term *competency* in the sense defined by the Swiss working group *HarmoS* (*Interkantonalen Vereinbarung über die Harmonisierung der obligatorischen Schule*) (EDK, 2011) and Blömeke et al. (2015)

Process Model. This shows that, in developing the Learning Process Model, aspects from general pedagogy and subject-specific education were taken into account.

In the Learning Process Model, a competency is built up stepwise, as recommended by learning psychology (Hattie & Yates, 2015; Helmke, 2015; Meyer, 2016; Wellenreuther, 2019). Lersch (2010) presents a two-dimensional model called the Competency Acquisition Scheme (*Kompetenzerwerbsschema*), in which the competency is divided into several subcompetencies, which are then acquired step-by-step. In contrast to the Competency Acquisition Scheme, the Learning Process Model assumes that the development of a competency is not necessarily linear, unlike as proposed by Lersch (2010), whose model was linear. It is, rather, a learning process that can take place in different directions, extending learners' prior knowledge in a complex manner. Previous empirical, cross-sectional studies in subject-specific education confirm this assumption, although corresponding longitudinal studies are still lacking (Bernholt et al., 2009).

Luthiger et al. (2018) assume that competencies are built up effectively when learners are confronted with the competency to be acquired in an initial situation or problem. Consequently, the Learning Process Model starts functionally, with the so-called *confrontation task* dealing with a new problem or everyday situation and inviting the learners to question their own concepts and ways of acting while they formulate a hypothesis about how the problem might be solved. In the subsequent learning process, the learners gradually acquire the necessary concepts and skills and deepen them to master (synthesize) the situation/problem and to transfer it to similar situations at the end of the teaching sequence. In summary, the Learning Process Model starts and ends with learners' daily life, for example, with an everyday problem, situation or context (right-hand arrows in Fig. 1). The Learning Process Model consists of the following tasks embedded in the learning process: *confrontation task*, *development tasks*, *exercises*, *synthesis task* and *transfer task* (Fig. 1).

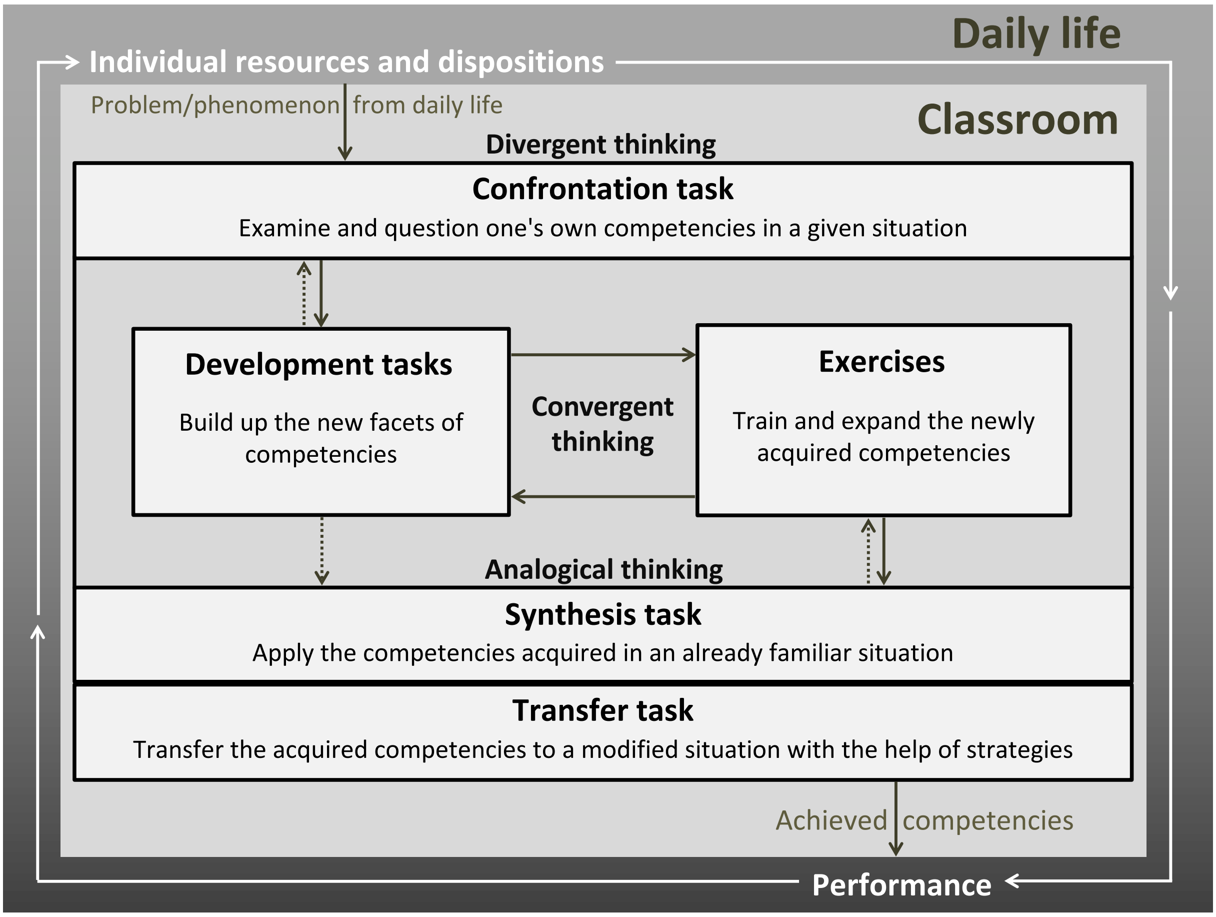
Daily life
Individual resources and dispositions
Classroom
Problem/phenomenon from daily life
Divergent thinking
Confrontation task
Examine and question one's own competencies in a given situation
Development tasks
Build up the new facets of competencies
Convergent thinking
Exercises
Train and expand the newly acquired competencies
Analogical thinking
Synthesis task
Apply the competencies acquired in an already familiar situation
Transfer task
Transfer the acquired competencies to a modified situation with the help of strategies
Achieved competencies
Performance

Fig. 1. Learning Process Model according to Luthiger et al. (2018). Classroom legend: solid arrows denote linear processes; dashed arrows indicate non-linear processes.

These tasks can be grouped according to three elements of thinking (de Haan, 2009): 1) *divergent thinking,* including ideational fluency or cognitive flexibility and accepting many ideas related to a problem; 2) *convergent thinking,* the request of inhibitory control to focus and mentally evaluate ideas*;* and 3) *analogical thinking*, the ability to understand a novel idea in terms of one that is already familiar.

The *confrontation task* allows and requires *divergent thinking* and welcomes learners' associations (Guilford, 1950). Consequently, the teacher does not correct the learners' hypotheses but leaves the "correct" answer open at this stage of the learning. The confrontation task establishes contact with the competency to be acquired and links the learners' daily life to the competency to be achieved at school (Luthiger et al., 2018). The confrontation task is based on real-life problems or authentic phenomena and should arouse curiosity, irritate learners and raise questions about the teaching unit's core idea and, so, increase the competency's relevance and leaners' readiness to acquire it (Luthiger et al., 2018).

The subsequent tasks in the Learning Process Model are the *development tasks,* followed by the *exercises*. Both demand *convergent thinking*. The development tasks should be cognitively activating and help learners to build up the new competency to be able to solve the initial confrontation task problem. Clear development task structure and immediate feedback make it possible to link the prior subjective concepts and behaviours with the competency to be achieved (de Haan, 2009; Hattie & Yates, 2015). Subsequently, different aspects of the competency are made more flexible and consolidated with the help of *exercises* (Luthiger et al., 2018). The Learning Process Model ends with the *synthesis task*, sometimes followed by the *transfer task*. Both task types primarily require *analogical thinking* (de Haan, 2009). The learners need to perform the acquired competencies to solve the initial confrontation task problem now reformulated as the synthesis task (Luthiger et al., 2018). In addition, transfer tasks aim to transfer and recombine the acquired competencies in modified situations with the help of strategies (Gysin & Brovelli, 2021).

2.2 Learning task quality analysis

As described above, tasks can be distinguished by their functional embedding in the learning process during education (chapter 2.1). Furthermore, tasks can be sorted according to various category systems. Several research groups in the field of general pedagogy and subject-specific education have developed and adapted category systems to analyse the quality of learning tasks. Wilhelm et al. (2014) have developed a category system to describe tasks in ten categories based on empirical (Blömeke et al., 2006; Jordan et al., 2006; Maier et al., 2010; Neubrand, 2002) and theoretical work (Büchter & Leuders, 2016). An overview of the selection of categories can be found in Luthiger et al. (2018). The work of Wilhelm et al. (2014) was the basis for the new instrument called *Instrument to Analyse Tasks (IAT)* (Stuppan et al., in press). The IAT is the instrument on which this study's results are based (chapter 3). In the following section, the theoretical background of the IAT's scales and their subscales are described in more detail.

2.2.1 Chart of competencies

The scale *chart of competencies* comprises the sub-competencies to be acquired within a task (Flechsig, 2008, p. 245; Lersch & Schreder, 2013, p. 50). According to the Swiss HarmoS (*Interkantonalen Vereinbarung über die Harmonisierung der obligatorischen Schule*) working group's report, a competency in science is composed of a skill linked to content/a concept (EDK, 2011). Moreover, a competency is divided into several sub-competencies (EDK, 2011). The chart of competencies scale consists of three subscales: singular (Cronbach's α of the subscale = .99, Number of items = 3), additive (α = .98, *n* = 3), and integrative (α = .98, *n* = 3). *Singular* tasks deal only with one aspect of a curricular sub-competency (Lersch & Schreder, 2013). Here, only one cognitive or personal skill is developed, enhanced, practised or applied. If a task involves several aspects of one or more curricular sub-competencies, it is rated as *additive*. In this case, several cognitive or personal skills are developed, enhanced, practised or applied. If a task is *integrative*, it involves as many aspects as possible of one or more of a competency's sub-competencies (Luthiger et al., 2018, p. 61). In this final case, many skills or abilities are developed, enhanced, practised, and applied in an interconnected way.

2.2.2 Relationship to daily life

The scale *relationship to daily life* refers to the relationship a task's plot has to learners' everyday experiences and environment (Maier et al., 2010, pp. 37–38; Stein et al., 1996, p. 486). This scale contains the following three subscales: constructed (α = .96, *n* = 3), authentic (α = .93, *n* = 3), and real (α = .96, *n* = 3). A task is called *constructed* when there is only an artificial link between the task's context and the learners' daily life (Maier et al., 2010, p. 38). An *authentic* task is one in which the relevance to the learners' daily life may be fictitious, but remains logical and authentic in the context (Blömeke et al., 2006; Luthiger et al., 2018). For instance, a task's context may be relevant to learners in future. A task is *real* when its context describes a situation that learners might have experienced themselves or that might have been experienced by another learners their age (Luthiger et al., 2018, p. 61). In a real-life context, there is no difference between the task and the learners' daily life (Neubrand, 2002, p. 113). Such a task must probably also be solved in learners' lives outside of school, too. Consequently, the task is personally relevant to the learners.

2.2.3 Learners' conceptions

Learners' conceptions determine learning in the sense that new input is perceived from the perspective of what is already known (Duit, 1995). Therefore, learners' prior knowledge and skills may be the starting point for tasks (Luthiger et al., 2018, p. 62). According to Wilhelm and Kalcsics (2017, p. 88), conceptions may be actively revised, altered or expanded during competency acquisition. According to Beerenwinkel et al. (2007), Luthiger et al. (2018, p. 62), and Wilhelm et al. (2014), the conceptions scale in tasks is divided into three subscales: raise (α = .91, *n* = 3), check (α = .90, *n* = 2), and reflect (α = .82, *n* = 3). *Raise* means learners share or write down their conceptions about the task. For example, assumptions about how to solve a problem are written down, exchanged orally among one another or communicated in another form (Mietzel, 2007, pp. 305–306). *Check* means learners' conceptions (e.g., an assumption about a solution or reasoning) are matched with the task's solution (Hattie & Yates, 2015, pp. 108–109). *Reflect* means learners reflect on their acquisition of knowledge and compare their preconceptions with their post-conceptions (Luthiger et al., 2018, p. 62). For instance, learners reflect on why their hypothesis matches the solution or why it does not.

2.2.4 Knowledge

*Knowledg*e refers to what kind of knowledge is to be acquired. The subscales of the knowledge scale are based on the work of Anderson and Krathwohl (2001, pp. 69–88). According to Anderson and Krathwohl, there are four subscales to be distinguished in the knowledge required by a task: factual knowledge (α = .91, *n* = 3), procedural knowledge (α = .96, *n* = 3), conceptual knowledge (α = .87, *n* = 3), and metacognitive knowledge (α = .89, *n* = 3). *Factual knowledge* includes knowledge about terminology, terms, specific details and elements (Anderson & Krathwohl, 2001; Gagné et al., 1992; Jordan et al., 2006). *Procedural knowledge* describes knowledge that is practically usable for action. This involves a continuum from basic types of behaviour to complex actions (e.g., solving unknown problems). Procedural knowledge can include cognitive aspects and/or courses of action (e.g., to perform an experiment) (Anderson & Krathwohl, 2001; Blömeke et al., 2006; Luthiger et al., 2018, p. 63). Learners apply *conceptual knowledge* in solving tasks when they explicitly link, classify, or structure their knowledge (Anderson & Krathwohl, 2001; Gagné et al., 1992; Hiebert & Lefevre, 1986, pp. 3–5; Jordan et al., 2006). Learners use *metacognitive knowledge* when they explicitly reflect on their knowledge or problem-solving strategies (Anderson & Krathwohl, 2001; Bremerich-Vos, 2008; Maier et al., 2010).

2.2.5 Knowledge activities

The *knowledge activities* scale refers to how prior knowledge must be transformed to solving a task (Luthiger et al., 2018, p. 63). All types of knowledge may be transformed: factual knowledge, procedural knowledge, conceptual knowledge, and metacognitive knowledge (Maier et al., 2014, p. 32). The knowledge activities scale consists of three subscales: reproduction (α = .88, *n* = 4), transfer (α = .94, *n* = 4), and creation (α = .96, *n* = 4). The subscale *reproduction* means learners need their memory to solve the task (Anderson & Krathwohl, 2001, p. 215). The task has a repetitive character for learners. Reproduction can refer to all four knowledge subscales and not only to recollecting factual knowledge. According to Ellis (1965), Maier et al. (2014, p. 32) and Dori and Sasson (2013), a *transfer* is needed when solving the task requires more than mere reproduction of prior knowledge. The task contains an unknown situation/problem/context to be solved or it is not immediately obvious to learners which type of knowledge they must apply. A task requires learners' *creation* when learners encounter a situation (often an unknown task) in which they need creativity to assemble or combine existing elements of prior knowledge to solve the task (Anderson & Krathwohl, 2001, p. 215).

2.2.6 Forms of representation

Based on existing category systems by Maier et al. (2014, pp. 39–40), Wilhelm et al. (2014) and Luthiger et al. (2018, p. 64), the *forms of representation* scale describes in its subscale a continuum from transform to not transform by comparing the form of representation in the task to the form of representation for processing and solving the task (α = .93, *n* = 4). Maier et al. (2010, p. 40) ask themselves two questions to determine the forms of representation in a task. First, in what form is the task presented to the learners? Second, in what form does the solution need to be processed and presented? *Transform* means that the form of representation (e.g., texts, figures, tables, etc.) in the task differs from the form of representation in the task progression and solution. This means learners must transform the learning content into another form of representation (Maier et al., 2010, pp. 39–40; Neubrand, 2002, p. 120; Stein et al., 1996). *Not transform* means the form of representation is the same in the task, the task progression, and the solution.

2.2.7 Task openness

The *task openness* scale describes the solution procedure (α = .95, *n* = 4), the result (α = .96, *n* = 4), and the editing (α = .92, *n* = 4). The *solution procedure* subscale is described by a continuum from undefined to defined. *Undefined* means the task allows variants during task procedure (Maier et al., 2010, p. 36). So, the task procedure becomes multi-layered and challenging. *Defined* means a clearly defined task procedure (e.g., algorithm). The *result* is described by a continuum from diverse to uniform. While *diverse* signifies that several goals or explanations are welcome as result of the task, *uniform* tasks expect a single solution (Stein et al., 1996). In this subscale, the correct solution (e.g., scientific knowledge or expertise) does not need to be defined (Maier et al., 2010, p. 36). The *editing* subscale defines the structuredness of the elements in a task and whether these structures correlate with the order in which the task must be solved (Luthiger et al., 2018, pp. 63–64; Maier et al., 2010, p. 36; Neubrand, 2002, pp. 122–129). The editing subscale comprises a continuum from structured to unstructured. *Structured* means the structure of the task is clearly defined and correlates with the order of solving the task. *Unstructured* means the sequence of structural elements in the task does not correspond to the task solution sequence. Consequently, learners must determine the structure while solving the task themselves (Luthiger et al., 2018, p. 64). If, for instance, a task is undefined, diverse, and unstructured, then learners can freely determine the task procedure and the solution is open. Moreover, the information given in the task does not correspond to the order learners must follow to process the information. How learners solve an undefined, diverse, unstructured task depends on, among others, learning prerequisites such as the (pre)conceptions, motivation, or task solution strategies (Ulrich, 2019, p. 79). The complexity of an undefined, diverse, unstructured task can still be controlled to some extent by the quality and structure of the information provided to the learners (Luthiger et al., 2018, p. 63). However, in general, it demands more competencies from learners than a defined, uniform, structured task.

2.2.8 Learning supports

The *learning supports* scale means any optional support offered to learners to enable them to work on tasks independently. According to Wilhelm et al. (2014) and Luthiger et al. (2018, pp. 65–66), the following three subscales are relevant in the learning supports scale: assistance (α = .97, *n* = 3), exchange opportunities (α = .88, *n* = 3), and teacher feedback (α = .98, *n* = 3). The *assistance* subscale means learners may get hints or prompts as supports to solve the task. The support is often provided in steps, to reduce cognitive load and take into account the limited capacity of working memory (Sweller et al., 1998). The *exchange opportunities* subscale means learners are given the chance to interact with one another during task processing. The interaction stimulates cognitive processing and supports memorisation and understanding. Interestingly, learners benefit not only when explaining to a classmate, but also when receiving explanations from another learner (Mietzel, 2007, p. 381). The *teacher feedback* subscale means teachers give feedback immediately after task processing or later (e.g., during the next lesson). So, task performance is evaluated and compared to a standard, options on how to improve performance are described, or incomplete or false solutions are corrected. Furthermore, teacher feedback provides structural support and contains motivational messages and the like (Mietzel, 2007, p. 381).

2.2.9 Learning paths

In the *learning paths* scale, the following three subscales are distinguished according to Niggli (2013, p. 35), Wilhelm et al. (2014), Büchter and Leuders (2016, p. 111) and Luthiger et al. (2018, p. 66): compensating (α = .93, *n* = 3), profiling (α = .92, *n* = 3), and self-differentiating (α = .92, *n* = 3). The *compensating* subscale means tasks allow individual learning and task processing at different performance levels (guided by compensation principles). When different learning statuses are compensated for, individual learners or learning groups come into focus instead of the entire class, which lets compensation of individual learning status become more likely (Niggli, 2013, p. 36). Consequently, a compensatory task enables learners to reach a defined competence level and compensate for any learning gaps. Such tasks offer the opportunity to supplement missing knowledge and clarify ambiguities. The *profiling* subscale means a task offers a choice, according to individual interests, to experience a competency expansion in one's profile. The *self-differentiating* subscale means a task can be solved at various levels of abstraction, ways of task processing and problem solving, or possibilities for deepening the topic (Büchter & Leuders, 2016, p. 111).

2.3 Research questions

The purpose of this study is to combine existing task scales from research with the different functions of tasks along the learning process model. The Learning Process Model and the scales introduced are already being used in various development projects (chapter 1). In addition, existing category systems are used in research on tasks for the nominal description of tasks. The results are often presented as text modules with explanations. Exemplary studies are those by Gloe and Miller (2017), Kleinknecht et al. (2013), Maier et al. (2010) and Reinfried (2016). But how do the described scales and their subscales relate to the task types (e.g., confrontation tasks) in the Learning Process Model?
So, the following research questions can be derived:

1) What scale values of the IAT do experts propose to characterize the various Learning Process Model task types?
2) How do the scale values gained from analysing confrontation tasks differ from the experts' scale values?

To answer the research questions, a panel of experts was consulted (chapter 3.2) and data from the development study of the Instrument to Analyse Tasks (IAT) by (Stuppan et al., in press) were re-analysed (chapter 3.3).

3 Methods

3.1 Instrument

Wilhelm et al. (2014) and Luthiger et al. (2018) have proposed a synoptic view of the Lucerne Model for the Development of Competency-Enhancing Task Sets (LUKAS) model and created a theory-based category system to study the potential quality of learning tasks. From this category system, the *Instrument to Analyse Tasks (IAT)* was developed. The IAT does not contain new scales of analysis, nor does the setting-up of the instrument include creative effort by the authors. Rather, the creation of the tool was a matter of selecting and compiling categories from established systems (chapter 2.2) to be combined and operationalised in this new instrument. The item construction accounted for the development of the IAT follows the methods described by Bühner (2006), Busker (2014), Döring and Bortz (2016), and Mummendey and Grau (2014). Thanks to its multi-stage development and complex validation process, the IAT is an empirically tested instrument. To validate the IAT, 146 tasks were used. The tasks originated from the project *MINT unterwegs*, a collaborative project between the University of Teacher Education Lucerne and the Lucerne Office of Education (DVS, Dienststelle Volksschulbildung Luzern). A paper describing the detailed, methodical approach used to create the IAT is in press (Stuppan et al., in press). The instrument consists of the following nine scales, each with one to four subscales: 1) chart of competencies, 2) relationship to daily life, 3) learners' conceptions, 4) knowledge, 5) knowledge activities 6) forms of representation, 7) task openness, 8) learning supports, and 9) learning paths. Examples of the subscales, Cronbach alpha coefficient and number of items are described in chapter 2.2. Each scale uses a five-level rating scale (0 = do not agree, 1 = agree slightly, 2 = agree moderately, 3 = agree considerably and 4 = agree strongly).

3.2 Experts' proposed scale values

A panel of experts (*N* = 4, 50% of them female) proposed IAT agreement scale values for each Learning Process Model task type. The panel consisted of the four authors of the LUKAS model (Luthiger et al., 2018). All members had either a general pedagogy or subject-specific education background. The proposed scale allows for discrete integer values ranging from 0 to 4 (0 = do not agree, and 4 = agree strongly).

First, the nominal descriptions from the LUKAS model (Luthiger et al., 2018; Wilhelm et al., 2014) were transferred and expanded to generate provisional scale values for the IAT. Next, these provisional scale values of the IAT were proposed in a panel meeting by the LUKAS model experts. Subsequently, the four experts each generated a personal proposal of the provisional scale values for each task type. From these scale values, the mean (rounded to two decimal places) for each task type was calculated to obtain the final scale values (Tab. 1). The median could also have been used here, as the use of the mean can lead to a shift in the average values. We have retained the mean values because, in this study, they lead to the same conclusions as using the median would have. Since the expert panel did not rate multiple tasks, but proposed scale values for each Learning Process Model task type, we did not calculate an intraclass correlation coefficient (ICC). Instead, we calculated the AD_M index (average absolute deviation) for each scale value as an interrater agreement. The standard for essential agreement was given by Burke et al. (1999) as A/6, where A is the number of response categories. For a five-level scale, an AD_M below 5/6 = .83 is an essential agreement. For A/6, Burke et al. (1999) followed the acceptable match threshold of .70 proposed for the r_{WG} (assessment of within-group interrater agreement) (James et al., 1984). If this value is exceeded, it is marked with ([a]) in Table 1. As described in the Delphi method (Döring & Bortz, 2016, pp. 420–421), the group result should be more meaningful than the individual experts' judgments.

3.3 Rated task scale values

The authors involved in the project *MINT unterwegs* created teaching units with the help of task sets. The authors were asked to label the developed tasks during *MINT unterwegs* task creation according to the different task types. We adopted this attribution of the task types and were able to index 25 confrontation tasks. These 25 confrontation tasks were empirically rated by trained lecturers of STEM subjects (*N* = 2) with the rating manual of the IAT. In total, 24% of the tasks were double-rated. The agreement on the scale values was tested by using the ICC and resulted in a mean ICC(A,1) value of .69. Then, the mean (rounded to two decimal places) of the scale values obtained from the 25 tasks was calculated (Fig. 2). The confrontation tasks are used as examples because they are the first task type in the Learning Process Model. Further analyses of other task types are in this article's supplementary materials. These additional analyses are not included in the Discussion and Conclusions section.

3.4 Scale value comparison

The comparison of the experts' scale values and the rated STEM confrontation tasks was calculated (Fig. 2). The integer point difference of the two scale values can reach the values 0 (perfect match), 1/-1 (substantial match), 2/-2 (moderate match), 3/-3 (slight match) and 4/-4 (poor match). The closer a point difference is to 0, the more similar the scale values of the experts' proposals and the rated tasks are to each other.

4 Results

In the following section, the results of the scale values of the experts' proposals are presented and explained (Tab. 1). Figure 2 shows the comparison between the scale values of the experts' proposals and the scale values of the rated STEM confrontation task.

4.1 Experts' proposed scale values

The description of tasks shown in this study does not aim to establish a hierarchical taxonomy of proficiencies, as is the case with Bloom's (1976) taxonomy. Instead, the different task types in the *Learning Process Model* (Fig. 1) are characterized by scale values of each subscale in each scale as proposed by experts. The result of the scale values is presented in Table 1. For the description of the scales and subscales, see 2.2 Learning task quality analysis.

In each task type, all scales are present. However, the scale values of each subscale within each scale differ. Consequently, the scale values of the subscales characterize the different task types.

Tab. 1. Experts' proposed scale values by mean for each task type

		Task type and proposed scale values									
Scale	Subscale	Confrontation		Development		Exercises		Synthesis		Transfer	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
chart of competencies	singular	0.00	0.00	2.00	0.00	3.00	0.00	0.50	0.50	0.00	0.00
	additive	1.25	0.43	3.00	0.00	2.00	0.00	1.00[a]	0.87	2.00	0.00
	integrative	3.50	0.50	2.00	0.00	0.00	0.00	3.50	0.50	3.50	0.50
relationship to daily life	constructed	2.75	0.43	3.75	0.43	2.25	0.43	2.75	0.43	2.75	0.43
	authentic	3.00	0.00	2.00	0.00	1.75	0.43	3.00	0.00	3.00	0.00
	real	1.75	0.43	1.00	0.00	1.00	0.00	1.25	0.43	2.00	0.00
learners' conceptions	raise	4.00	0.00	3.50[a]	0.87	2.50	0.50	4.00	0.00	3.00	0.00
	check	0.00	0.00	4.00	0.00	3.00	0.71	2.25	0.43	2.00	0.00
	reflect	0.50[a]	0.87	3.00	0.00	1.75	0.43	1.50[a]	1.22	1.00[a]	0.87
knowledge	factual knowledge	1.00	0.00	3.25	0.43	3.25	0.43	2.25	0.43	1.00	0.00
	procedural knowledge	2.00	0.00	3.00	0.00	3.00	0.00	2.00	0.00	2.00	0.00
	conceptual knowledge	4.00	0.00	4.00	0.00	2.75	0.43	2.50[a]	0.87	3.00	0.00
	metacognitive knowledge	3.00	0.00	3.00	0.00	2.00	0.00	3.00	0.00	4.00	0.00
knowledge activities	reproduction	1.00	0.00	2.00	0.00	4.00	0.00	3.00	0.00	1.75	0.43
	transfer	1.00[a]	0.87	3.00	0.00	2.50	0.50	3.00	0.00	4.00	0.00
	creation	4.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	2.25	0.43
forms of representation	transform	2.25	0.43	1.25	0.43	2.00	0.00	3.00	0.00	3.75	0.43
task openness	solution procedure undefined	3.00	0.00	1.00	0.00	1.00	0.00	3.00	0.00	3.00	0.00
	result diverse	4.00	0.00	1.00	0.00	0.00	0.00	2.25	0.43	3.00	0.71
	editing structured	3.50	0.50	3.25	0.43	2.25	0.43	3.00	0.00	2.00	0.00
learning supports	assistance	0.25	0.43	2.25	0.43	2.25	0.43	2.50[a]	0.83	2.25	0.43
	exchange options	2.50	0.50	3.00	0.00	1.75	0.43	2.00	0.00	2.25	0.43
	teacher feedback	0.00	0.00	4.00	0.00	3.50[a]	0.83	3.00	0.00	2.75	0.43
learning paths	compensating	1.75	0.43	3.75	0.43	3.25	0.43	1.75	0.43	1.75	0.43
	profiling	1.75	0.43	1.00	0.00	1.75	0.43	3.00	0.00	3.00	0.00
	self-differentiating	4.00	0.00	2.00	0.00	1.00	0.00	3.00	0.00	3.75	0.43

5-level rating scale (from 0 = do not agree to 4 = agree strongly); [a] According to Burke et al. (1999), the AD_M index (average absolute deviation) shows no essential agreement.

To explain the results by example, the first column (confrontation task) is described in more detail (Tab. 1). The detailed description is mainly focused on scale values that show considerably (3) to strongly (4) agreement. According to the scale values proposed by the experts in the scale *chart of competencies*, confrontation tasks are integrative (3.50). This implies that the confrontation task includes as many aspects as possible of one or more sub-competencies defined by the curriculum. As per the scale *relationship to daily life,* the confrontation task is either constructed (2.74) or authentic

(3.00). This means the confrontation task has, according to the expert proposal, either a context that may be relevant to learners in future, or a constructed one. Looking at the scale *learners' conceptions*, raise (4.00) has the highest scale value. This means it is central that learners come up with their own solution ideas and communicate them in a confrontation task. According to the scale values in the scale *knowledge*, conceptual knowledge (4.00) and, to a smaller extent, metacognitive knowledge (3.00) apply in the confrontation task. As per the scale *knowledge activities*, the highest scale value reaches the subscale creation (4.00). Consequently, a confrontation task, according to the experts consulted, enables and fosters creativity in learners. The scale *forms of representation* shows, with its scale value 2.25, that a confrontation task is moderately transformed. According to the scale *task openness,* a confrontation task has a rather undefined solution procedure (3.00), and the solution is diverse (4.00). However, the editing is structured (3.50). The scale *learning supports* displays an expectation of exchange opportunities for confrontation tasks (3.50), but no assistance nor teacher feedback. Finally, the scale values of the scale *learning paths* are proposed for confrontation tasks to be self-differentiating (4.00). This means a confrontation task allows various levels of abstraction and can be solved with a range of prior knowledge.

4.2 Scale value comparison

Figure 2 presents, by way of example, an assessment of confrontation tasks. In the first column, the experts' mean scale value proposals are shown. The second column displays the mean scale values of all the rated confrontation tasks (*n* = 25) in the *MINT unterwegs* project.

		Experts' proposed scale values		Rated tasks scale values STEM confrontation tasks - mean (*n* = 25)		Point difference of the scale values	
Scale	**Subscale**	*M*	*SD*	*M*	*SD*	Point diff.	Colour scale
chart of competencies	singular	0.00	0.00	0.63	0.96	-0.63	🟩
	additive	1.25	0.43	2.47	1.48	-1.22	🟩
	integrative	3.50	0.50	1.08	1.13	2.42	🟧
relationship to daily life	constructed	2.75	0.43	1.73	1.17	1.02	🟩
	authentic	3.00	0.00	2.69	1.04	0.31	⬜
	real	1.75	0.43	1.53	1.11	0.22	⬜
learners' conceptions	raise	4.00	0.00	3.52	0.45	0.48	⬜
	check	0.00	0.00	0.96	0.88	-0.96	🟩
	reflect	0.50 [a]	0.87	0.93	0.76	-0.43	⬜
knowledge	factual knowledge	1.00	0.00	2.43	0.83	-1.43	🟩
	procedural knowledge	2.00	0.00	1.77	1.31	0.23	⬜
	conceptual knowledge	4.00	0.00	2.95	0.44	1.05	🟩
	metacognitive knowledge	3.00	0.00	0.57	0.72	2.43	🟧
knowledge activities	reproduction	1.00	0.00	1.11	0.97	-0.11	⬜
	transfer	1.00 [a]	0.87	1.97	1.11	-0.97	🟩
	creation	4.00	0.00	2.16	1.08	1.84	🟧
forms of representation	transform	2.25	0.43	2.02	0.98	0.23	⬜
task openness	solution procedure undefined	3.00	0.00	2.92	0.69	0.08	⬜
	result diverse	4.00	0.00	2.58	1.13	1.42	🟩
	editing structured	3.50	0.50	2.78	1.01	0.72	🟩
learning supports	assistance	0.25	0.43	0.84	0.67	-0.59	🟩
	exchange options	2.50	0.50	3.19	0.88	-0.69	🟩
	teacher feedback	0.00	0.00	1.23	0.89	-1.23	🟩
learning paths	compensating	1.75	0.43	1.21	1.15	0.54	🟩
	profiling	1.75	0.43	1.29	1.12	0.46	⬜
	self-differentiating	4.00	0.00	2.60	0.87	1.40	🟩

Comment:
[a] According to Burke et al. (1999), the AD_M index (average absolute deviation) shows no essential agreement.

5-level rating scale	
4	agree strongly
3	agree considerably
2	agree moderately
1	agree slightly
0	do not agree

Colour scale regarding to integer point difference	
0	⬜
1/-1	🟩
2/-2	🟧
3/-3	🟥
4/-4	🟫

Fig. 2. Scale value comparison: assessment of confrontation tasks by the expert proposals and rated confrontation tasks (*n* = 25)

In the third column, the experts' proposed scale values are compared to the scale values of the rated STEM tasks. The assessment shows substantial-to-perfect matches (the integer point difference is between 0 and -1 or 1) for all subscales, except for the subscales: *chart of competencies - integrative* (point difference = 2.42), *metacognitive knowledge* (point difference = 2.43) and *knowledge activities - creation* (point difference = 1.84).

The scale values with a substantial-to-perfect match between the experts' proposals and the task rating deviate by an integer point difference of -1 or 1. The scale values matching moderately have scale values with an integer point difference of -2 or 2. No slight matches (integer point differences = 3/-3) nor poor matches (integer point differences = 4/-4) are present. To clearly and visually display the matches identified (integer point difference), we have introduced a "traffic light" colour-coding system (Fig. 2 Colour scale).

5 Discussion and Conclusions

In this study, a panel of experts was consulted to characterize each task type in the Learning Process Model by scale expressions from the *Instrument to Analyse Tasks* (IAT) (Tab. 1). This study's results make it evident that experts can propose different task types and that they agree on most of the scale expressions. In the second part, confrontation tasks from the *MINT unterwegs* project were analysed. For this purpose, the potential of task quality was assessed with the IAT. Subsequently, the rated tasks were compared with the experts' proposals (Fig. 2). The results show that the tasks rated correspond to a large extent with the experts' proposals. The scale value differences are produced in the form of an integer colour scale (0 to 4). With this "traffic light" system, a quick overview of the differences of the rated tasks from the expert proposal becomes visible. These variations can be analysed qualitatively to revise the observed tasks based on criteria and optimise them for the learning process.

Interpretation of the findings

To answer the first research question*: What scale values of the IAT do experts propose to characterize the various Learning Process Model task types?,* the mean of the expert proposals was calculated for each task type's subscales (Tab. 1). Except for a few scale values (e.g., *learners' conceptions - reflect*) (Tab. 1), the different experts proposed the scale values quite congruently. The different scale values across the task types indicate that the task types of the Learning Process Model can be distinguished from one another (Fig. 1). This means that a distinction between different task types is possible, as described by Luthiger et al. (2018). Furthermore, this is also accentuated by Abraham and Müller (2009) and by Büchter and Leuders (2006). It is noted that the experts do not have a homogeneous scale values proposal in some task types. Therefore, a further survey of the experts' proposals would be welcome (Diaz-Bone & Weischer, 2015, p. 83).

The answer to the second question: *How do the scale values gained from analysing confrontation tasks differ from the experts' scale values?* is displayed in Figure 2. The presented results compare the scale values obtained in the two approaches. There are many substantial-to-perfect matches between the experts' characterisations of the confrontation task and the rated tasks, which indicates a high agreement between the two approaches. A perfect match (integer point difference of 0) is visible in the subscales:

- *relationship to daily life - authentic*
- *relationship to daily life - real*
- *learners' conceptions - raise*
- *learners' conceptions - reflect*
- *knowledge - procedural knowledge*
- *knowledge activities - reproduction*
- *forms of representation - transform*
- *task openness - solution procedure undefined*
- *learning paths - profiling*

Some subscales, however, demonstrate differences from the *experts' proposal.* These differences can be used to improve the analysed tasks according to the experts' characterization of a confrontation task. So, this analysis and the subsequent revision of the tasks certainly serves as an important step for planning general pedagogy and subject-specific education (Reinfried, 2016, p. 12).

In cases where the match is only "moderate" (integer point difference of 2), one can think about the reasons for this discrepancy between the expert's proposal and the tasks rating. For instance, why does the *integrative* subscale of the *chart of competencies* scale with a point difference of 2.42 display such a high deviation between the experts' proposal and the task rating? One reason may be that the two subscales *integrative* and *additive* are quite close to one another, as there is only a small difference in content between the wording "several aspects of one or more sub-competencies" and "as many aspects of one or more sub-competencies as possible". In addition, experts propose a confrontation task to be *integrative* (M = 3.50); the empirical finding, however, shows most confrontation tasks to be *additive.* This discrepancy may also reflect the fact that it is very demanding to formulate a confrontation task involving as many aspects on one or more sub-competencies as possible, leading to several aspects of one or more sub-competencies and consequently being rated as *additive* (M = 2.47). People developing learning tasks should be advised to keep this in mind and to choose confrontation tasks with as many aspects of one or more sub-competencies as possible. The *metacognitive knowledge* subscale within the *knowledge* scale matches moderately well, with a point difference of 2.43. Several reasons may explain this difference. First, it may be that, in the experts' proposal, both the *conceptual knowledge* (point difference of 1.05) and the *metacognitive knowledge* (point difference of 2.43) subscales are rated higher than the analysed tasks, whereas the rating of the *procedural knowledge* (point difference of 0.23) is similar in both methods and the *factual knowledge* (point difference of -1.43) is higher in the actual tasks analysed than in the experts' proposal. The discrepancies between the experts' proposals and the analysed tasks may be explained, therefore, by the fact that it is far easier to create tasks dealing with factual knowledge and procedural knowledge compared to conceptual knowledge or metacognitive knowledge. The high expert weightings may be explained by the fact that the experts are familiar with research results showing the importance of conceptual knowledge and metacognitive knowledge (Brüchner, 2007; Kaiser & Kaiser, 2018). Second, metacognitive knowledge may be very hard to implement in tasks and therefore has a lower scale value in the task analysis. In both cases, it may help to point out to people developing learning tasks the advantage of involving metacognitive knowledge very consciously in confrontation tasks.

Limitations of the study and implications for research

In this study, four experts were consulted to characterize the various Learning Process Model task types. To generalize the characterization of the task types even more, a survey with further experts could be considered.

The confrontation tasks analysed were constructed and labelled as confrontation tasks by the *MINT unterwegs* group of authors. In this study, the mean of all assigned confrontation tasks was compared with the characterisation of a confrontation task made by experts. In addition, single confrontation tasks could be put in focus. Then, it may happen that an assigned confrontation task might not correspond to a confrontation task according to the experts' task characterization, but matches better with, for example, an exercises task. Whether such discrepancies exist and, if so, to what extent, might be the topic of further research using cluster analyses.

Future research might focus on individual task types in task sets along with analyses of any science tasks (e.g., in textbooks) or tasks from other subjects. Furthermore, it would be interesting to analyse and compare task sets from other educational systems (e.g., outside Switzerland), while being aware that these tasks were not developed according to the LUKAS model. So far, the evaluation of the confrontation tasks has been done descriptively. Further conclusions and statements might be formulated using inferential statistics.

With the theoretical task setting in Table 1, existing category systems were operationalised, and individual subscales were assigned to a task type in the Learning Process Model (Bölsterli Bardy & Wilhelm, 2018; Luthiger et al., 2018). This normative posting is based on scientific work (Blömeke et al., 2006; de Haan, 2009; Jordan et al., 2006; Krabbe et al., 2015, 2015; Lersch, 2010; Maier et al., 2014; Neubrand, 2002; Reusser, 2014). A meta-analysis of the individual scales and subscales could further validate the IAT.

Additionally, only the theoretical potential of the tasks is determined in this project. So far, no studies about the learning progression or the learners' curiosity or motivation using analysed tasks have been carried out. The IAT can be used to measure the potential of science tasks in nine quality dimensions according to Stuppan et al. (in press). Indeed, there may be other scales and subscales also contributing to the characterization of task types. For example, Maier et al. (2010) analysed the dimension: *sprachlogische Komplexität* ("linguistic complexity") in their category system, while Blömeke et al. (2006) examined *Chance auf Bewältigung* ("likelihood of coping") in their work.

In addition, teachers embed tasks in their lessons (Blömeke et al., 2006). Consequently, we expect some differences when analysing tasks by themselves (Fig. 2) versus embedded in education. Therefore, research is also foreseen to investigate the effect of tasks in class.

Potential of the IAT in the future

The IAT and the experts' proposals can be used in future development projects dealing with tasks (by publishers of teaching materials, textbook authors, etc.). An iterative development process in the field, as pursued by the design-based research approach in research design (Baumgartner et al., 2003; Reinmann, 2005), and multiple additional uses of the IAT, contribute to revising the quality of the tasks according to the intended task type.

The IAT can be used to analyse single tasks and determine the deviation between the experts' proposal and the single task or the deviation between the rating of the tasks compared to a single task. So, for each task, an assessment could be created and evaluated accordingly.

The assessment of task quality is a central element of teachers' diagnostic competence. With the IAT and the experts' proposal, this competence can be trained. For example, the quality of tasks from didactic materials can be analysed with teachers in training, whether in degree programmes or continuing professional development settings, and suggestions for their further development and optimisation can be discussed and explored (Maier et al., 2010, pp. 73–74; Reinfried, 2016, p. 13).

We consider the modularity of the IAT a promising approach to help sharpen the function of different task types in the learning process. Hopefully, the presented work will enrich the dialogue around the central function of tasks in education.

Acknowledgements

We would like to thank the anonymous reviewers for their detailed feedback and valuable, critical and constructive suggestions. Furthermore, we would like to thank the Institute for Education in Science and Social Studies and the Research Fund at the University of Teacher Education Lucerne for their active and financial support. Great thanks go to Roland Künzle for his help in rating the Learning Process Model tasks and to Alex Brown for editing the text and providing linguistic feedback. Finally, special thanks go to the authors of the LUKAS model: their expertise was very helpful and had a great impact on this research.

References

Abraham, U., & Müller, A. (2009). Aus Leistungsaufgaben lernen. *Praxis Deutsch*, *36*(214), 4–12.

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (Complete ed). Longman.

Baumgartner, E., Bell, P., Brophy, S., Hoadley, C., Hsi, S., Joseph, D., Orrill, C., Puntambekar, S., Sandoval, W., & Tabak, I. (2003). Design-Based Research: An Emerging Paradigm for Educational Inquiry. *Educational Researcher*, *32*, 5–8, 35. https://doi.org/10.3102/0013189X032001005

Beerenwinkel, A., Parchmann, I., & Gräsel, C. (2007). *Chemieschulbücher in der Unterrichtsplanung – Welche Bedeutung haben Schülervorstellungen? Chemkon, 14*(1), 7–14.

Bernholt, S., Parchmann, I., & Commons, M. L. (2009). Kompetenzmodellierung zwischen Forschung und Unterrichtspraxis. *Zeitschrift für Didaktik der Naturwissenschaften*, 15, 219–245.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. (2015). Beyond Dichotomies Competence Viewed as a Continuum. *Zeitschrift für Psychologie*, *223*, 3–13. https://doi.org/10.1027/2151-2604/a000194

Blömeke, S., Risse, J., Müller, C., Eichler, D., & Schulz, W. (2006). Analyse der Qualität von Aufgaben aus didaktischer und fachlicher Sicht. Ein allgemeines Modell und seine exemplarische Umsetzung im Unterrichtsfach Mathematik. *Unterrichtswissenschaft*, *34*(4), 330–357.

Bloom, B. S., & Engelhart, M. D. (Eds.). (1976). *Taxonomie von Lernzielen im kognitiven Bereich* (5. Aufl.). Beltz.

Bölsterli Bardy, K., Brugger, P., Brückmann, M., von Fischer, E., Flory, T., Jakober, M., Metzger, S., Möschler, L., Müller, N., & Tempelmann, S. (2017). *NaTech 1-6*. Schulverlag Plus, Lehrmittelverlag Zürich.

Bölsterli Bardy, K., & Wilhelm, M. (2018). Von kompetenzorientierten zu kompetenzfördernden Aufgaben im Schulbuch. *Erziehung & Unterricht*, *168*(1/2), 121–129.

Bremerich-Vos, A. (2008). Benjamin S. Bloom (und andere) revisited. In A. Bremerich-Vos, D. Granzer, & O. Köller (Eds.), *Lernstandsbestimmung im Fach Deutsch. Gute Aufgaben für den Unterricht.* (pp. 29–49). Beltz.

Brüchner, K. (2007). *Metakognition und Lernen in Chemie :Studien zur Domänengeneralität versus Domänenspezifitätund Förderung der Metakognition beim Lernen in Chemie* [PhD Thesis]. https://macau.uni-kiel.de/receive/diss_mods_00002965

Büchter, A., & Leuders, T. (2006). Ein Aufgabenmodell für die Praxis. Einschätzung, Auswahl und Entwicklung von Mathematikaufgaben. *Praxis der Naturwissenschaften - Chemie in der Schule*, *55*, 16–20.

Büchter, A., & Leuders, T. (2016). *Mathematikaufgaben selbst entwickeln: Lernen fördern - Leistung überprüfen* (7., überarbeitete Neuauflage). Cornelsen.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion* (2., aktualisierte und erw. Aufl., [Nachdr.]). Pearson Studium.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On Average Deviation Indices for Estimating Interrater Agreement. *Organizational Research Methods*, *2*(1), 49–68. https://doi.org/10.1177/109442819921004

Busker, M. (2014). Entwicklung eines Fragebogens zur Untersuchung des Fachinteresses. In D. Krüger, I. Parchmann, & H. Schecker (Eds.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (pp. 269–281). Springer Spektrum.

de Haan, R. L. (2009). Teaching creativity and inventive problem solving in science. *CBE Life Sciences Education*, *8*(3), 172–181. https://doi.org/10.1187/cbe.08-12-0081

Diaz-Bone, R., & Weischer, C. (Eds.). (2015). *Methoden-Lexikon für die Sozialwissenschaften*. Springer VS.

Dori, Y., & Sasson, I. (2013). A three-attribute transfer skills framework-part I: Establishing the model and its relation to chemical education. *Chem. Educ. Res. Pract.*, *14*. https://doi.org/10.1039/C3RP20093K

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. vollständig überarbeitete, aktualisierte und erweiterte Auflage). Springer.

Duit, R. (1995). Vorstellungen und Lernen von Physik und Chemie. Zu den Ursachen vieler Lernschwierigkeiten. *Plus Lucis*, *2*, 11–18.

DVS. (2016). *Entdecke.lu.ch*. Dienststelle Volksschulbildung Luzern. entdecke.lu.ch

DVS. (2018). *Mint-erleben.lu.ch*. Dienststelle Volksschulbildung Luzern. mint-erleben.lu.ch

EDK. (2011). *Grundkompetenzen für die Naturwissenschaften. Nationale Bildungsstands.* Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren.

Flechsig, K.-H. (2008). Komplexe Aufgaben in der beruflichen Aus- und Weiterbildung. In J. Thonhauser (Ed.), *Aufgaben als Katalysatoren von Lernprozessen: Eine zentrale Komponente organisierten Lehrens und Lernens aus der Sicht von Lernforschung, Allgemeiner Didaktik und Fachdidaktik*. Waxmann Verlag.

Gagné, R. M., Briggs, L. J., & Wager, W. W. (1992). *Principles of instructional design* (4th ed). Harcourt Brace Jovanovich

College Publishers.

Gloe, M., & Miller, J. (2017). *Aufgaben im Politikunterricht—Analyse von Lernaufgaben in baden-württembergischen Schulbüchern für das Gymnasium. 43*(7), 10–14.

Guilford, J. P. (1950). Creativity. *American Psychologist*, *5*(9), 444–454. https://doi.org/10.1037/h0063487

Gysin, D., & Brovelli, D. (2021). Use of knowledge pieces and context features during the transfer process in physics tasks. *International Journal of Science Education*, *43*(13), 2108–2126. https://doi.org/10.1080/09500693.2021.1952334

Hammann, M. (2006). Kompetenzförderung und Aufgabenentwicklung. *Der mathematische und naturwissenschaftliche Unterricht, 59*(2), pp. 85–95).

Hattie, J., & Yates, G. C. R. (2015). *Lernen sichtbar machen aus psychologischer Perspektive* (W. Beywl & K. Zierer, Eds.; 1. Auflage, überarbeitete deutschsprachige Ausgabe von "Visible learning and the science of how we learn"). Schneider Verlag Hohengehren GmbH.

Helmke, A. (2015). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts: Franz Emanuel Weinert gewidmet* (6. Auflage). Klett/Kallmeyer.

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In *Conceptual and procedural knowledge: The case of mathematics* (pp. 1–27). Lawrence Erlbaum Associates, Inc.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, *69*(1), 85–98. https://doi.org/10.1037/0021-9010.69.1.85

Jordan, A., Ross, N., Krauss, S., Baumert, J., Blum, W., Neubrand, M., Löwen, K., Brunner, M., & Kunter, M. (2006). *Klassifikationsschema für Mathematikaufgaben. Dokumentation der Aufgabenkategorisierung im COACTIV-Projekt.* Max-Planck-Institut für Bildungsforschung.

Kaiser, A., & Kaiser, R. (2018). Die Neue Didaktik. In A. Kaiser, A. Lambert, R. Kaiser, & K. Hohenstein (Eds.), *Metakognition: Die Neue Didaktik: Metakognitiv fundiertes Lehren und Lernen ist Grundbildung* (pp. 21–29). Vandenhoeck & Ruprecht.

Kleinknecht, M., Bohl, T., Maier, U., & Metz, K. (Eds.). (2013). *Lern- und Leistungsaufgaben im Unterricht: Fächerübergreifende Kriterien zur Auswahl und Analyse.* Verlag Julius Klinkhardt.

Krabbe, H., Zander, S., & Fischer, H. E. (2015). *Lernprozessorientierte Gestaltung von Physikunterricht: Materialien zur Lehrerfortbildung.* Waxmann.

Lersch, R. (2010). Didaktik und Praxis kompetenzfördernden Unterrichts. In K. Faulstich-Christ & K. Moegling (Eds.), *Kompetenzorientierung in Theorie, Forschung und Praxis: Sekundarstufen I und II* (pp. 31–60). Prolog-Verlag

Lersch, R., & Schreder, G. (2013). *Grundlagen kompetenzorientierten Unterrichtens: Von den Bildungsstandards zum Schulcurriculum.* Budrich.

Luthiger, H. (2014). *Differenz von Lern- und Leistungssituationen. Eine explorative Studie zu ihrer theoretischen Grundlegung und empirischen Überprüfung.* Waxmann.

Luthiger, H., Wilhelm, M., Wespi, C., & Wildhirt, S. (Eds.). (2018). *Kompetenzförderung mit Aufgabensets: Theorie - Konzept - Praxis* (1. Auflage). hep.

Maier, U., Bohl, T., Kleinknecht, M., & Metz, K. (2014). Allgemeine Didaktik und ein Kategoriensystem der überfachlichen Aufgabenanalyse. In P. Blumschein (Ed.), *Lernaufgaben – Didaktische Forschungsperspektiven* (pp. 35–51). Verlag Julius Klinkhardt.

Maier, U., Kleinknecht, M., Metz, K., Schymala, M., & Bohl, T. (2010). *Entwicklung und Erprobung eines Kategoriensystems für die fächerübergreifende Aufgabenanalyse. Schulpädagogische Untersuchungen Nürnberg.* Forschungsbericht Nr. 38, Friedrich-Alexander-Universität.

Meyer, H. (2016). *Was ist guter Unterricht?* (11. Auflage). Cornelsen.

Mietzel, G. (2007). *Pädagogische Psychologie des Lernens und Lehrens.* Hogrefe.

Müller, A., & Helmke, A. (2008). Qualität von Aufgaben als Merkmale der Unterrichtsqualität – verdeutlicht am Fach Physik. In J. Thonhauser (Ed.), *Aufgaben als Katalysatoren von Lernprozessen: Eine zentrale Komponente organisierten Lehrens und Lernens aus der Sicht von Lernforschung, allgemeiner Didaktik und Fachdidaktik* (pp. 31–46). Waxmann.

Mummendey, H. D., & Grau, I. (2014). *Die Fragebogen-Methode: Grundlagen und Anwendung in Persönlichkeits-, Einstellungs- und Selbstkonzeptforschung* (6., korrigierte Auflage). Hogrefe.

Neubrand, J. (2002). *Eine Klassifikation mathematischer Aufgaben zur Analyse von Unterrichtssituationen. Selbsttätiges Arbeiten in Schülerarbeitsphasen in den Stunden der TIMSS-Video-Studie.* Franzbecker.

Niggli, A. (2013). *Didaktische Inszenierung binnendifferenzierter Lernumgebungen: Theorie - Empirie - Konzepte - Praxis.* Klinkhardt.

Oser, F., & Patry, J.-L. (1990). *Choreographien unterrichtlichen Lernens: Basismodelle des Unterrichts.* Freiburg Schw. Pädagogisches Institut der Universität Freiburg.

Reinfried, S. (2016). *Kompetenzorientierte Lernaufgaben – mehr als alter Wein in neuen Schläuchen?* Geographie aktuell & Schule, 38/223, 4-14.

Reinmann, G. (2005). Innovation ohne Forschung? Ein Plädoyer für den Design-Based Research-Ansatz in der Lehr-Lernforschung. *Unterrichtswissenschaft*, *33*(1), 52–69.

Reusser, K. (2014). *Aufgaben – Träger von Lerngelegenheiten und Lernprozesse im kompetenzorientierten Unterricht.* Seminar, 4/2014, 77-101.

Schweizerisches Agrarmuseum Burgrain. (2020). *Wer ist Landwirtschaft.* museumburgrain.ch/lehren-und-lernen

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building Student Capacity for Mathematical Thinking and Reasoning: An Analysis of Mathematical Tasks Used in Reform Classrooms. *American Educational Research Journal*, *33*(2), 455–488. JSTOR. https://doi.org/10.2307/1163292

Stuppan, S., Wilhelm, M., Bölsterli Bardy, K., & Künzle, R. (in press). *Messinstrument zur Analyse und Kategorisierung von MINT-Aufgaben – Konstruktion und Validierung*. Proceedings Tagung Fachdidaktiken 2022, Lugano.

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review*, *10*(3), 251–296. https://doi.org/10.1023/A:1022193728205

Ulrich, N. (2019). *Interaktive Lernaufgaben in dem digitalen Schulbuch eChemBook Einfluss des Interaktivitätsgrads der Lernaufgaben und des Vorwissens der Lernenden auf den Lernerfolg*. Universität Hannover.

Wellenreuther, M. (2019). *Lehren und Lernen - aber wie? Ein Studienbuch für das Lehramtsstudium* (10., unveränderte Auflage). Schneider Verlag Hohengehren GmbH.

Wespi, C., Senn, C., & Schelbert, Z. (2019). *Das WAH - Themenbuch* (1. Auflage). Schulverlag plus.

Wilhelm, M., & Kalcsics, K. (2017). *Lernwelten Natur – Mensch – Gesellschaft: Fachdidaktische Grundlagen. Ausbildung. Studienbuch 3. Zyklus* (1. Auflage). Schulverlag.

Wilhelm, M., Luthiger, H., & Wespi, C. (2014). *Kategoriensystem für ein kompetenzorientiertes Aufgabenset*. Luzern: Entwicklungsschwerpunkt Kompetenzorientierter Unterricht, Pädagogische Hochschule Luzern.

Wilhelm, M., Wespi, C., Luthiger, H., & Rehm, M. (2015). Mit Aufgaben Kompetenzen und Vorstellungen erfassen. Ein Kategoriensystem und ein Prozessmodell als Hilfe zur Planung von Aufgaben. *Naturwissenschaften im Unterricht. Chemie*, *26*(149), 9–15.

Supplementary materials

In the following, we compare the task types from the *MINT unterwegs* project to the expert proposals, in the same way as the confrontation task (Fig. 2): development task, exercises task, synthesis task and transfer task.

		Experts' proposed scale values		Rated tasks scale values STEM development tasks - mean (n = 52)		Point difference of the scale values	
Scale	**Subscale**	*M*	*SD*	*M*	*SD*	Point diff.	Colour scale
chart of competencies	singular	2.00	0.00	0.90	1.31	1.10	🟩
	additive	3.00	0.00	2.63	1.50	0.37	⬜
	integrative	2.00	0.00	0.78	1.07	1.22	🟩
relationship to daily life	constructed	3.75	0.43	2.06	1.16	1.69	🟧
	authentic	2.00	0.00	2.43	0.99	-0.43	⬜
	real	1.00	0.00	1.40	1.08	-0.40	⬜
learners' conceptions	raise	3.50 [a]	0.87	3.28	0.73	0.22	⬜
	check	4.00	0.00	2.03	1.17	1.97	🟧
	reflect	3.00	0.00	1.29	0.83	1.71	🟧
knowledge	factual knowledge	3.25	0.43	2.40	0.95	0.85	🟩
	procedural knowledge	3.00	0.00	2.77	1.04	0.23	⬜
	conceptual knowledge	4.00	0.00	2.71	0.77	1.29	🟩
	metacognitive knowledge	3.00	0.00	0.40	0.54	2.60	🟥
knowledge activities	reproduction	2.00	0.00	1.21	0.79	0.79	🟩
	transfer	3.00	0.00	2.02	0.96	0.98	🟩
	creation	0.00	0.00	1.25	0.92	-1.25	🟩
forms of representation	transform	1.25	0.43	2.63	1.00	-1.38	🟩
task openness	solution procedure undefined	1.00	0.00	1.75	1.22	-0.75	🟩
	result diverse	1.00	0.00	1.49	1.26	-0.49	⬜
	editing structured	3.25	0.43	3.04	0.94	0.21	⬜
learning supports	assistance	2.25	0.43	1.46	1.18	0.79	🟩
	exchange options	3.00	0.00	2.55	1.15	0.45	⬜
	teacher feedback	4.00	0.00	1.31	1.21	2.69	🟥
learning paths	compensating	3.75	0.43	2.72	0.87	1.03	🟩
	profiling	1.00	0.00	1.37	1.11	-0.37	⬜
	self-differentiating	2.00	0.00	2.38	0.91	-0.38	⬜

Comment:

[a] According to Burke et al. (1999), the AD_M index (average absolute deviation) shows no essential agreement.

5-level rating scale	
4	agree strongly
3	agree considerably
2	agree moderately
1	agree slightly
0	do not agree

Colour scale regarding to integer point difference	
0	⬜
1/-1	🟩
2/-2	🟧
3/-3	🟥
4/-4	🟫

Fig. 3. Scale value comparison: assessment of development tasks by the experts' proposal and rated development tasks (*n* = 52)

		Experts' proposed scale values		Rated tasks scale values STEM exercises tasks - mean (n = 33)		Point difference of the scale values	
Scale	**Subscale**	*M*	*SD*	*M*	*SD*	Point diff.	Colour scale
chart of competencies	singular	3.00	0.00	0.74	1.04	2.26	🟧
	additive	2.00	0.00	2.96	1.22	-0.96	🟩
	integrative	0.00	0.00	0.99	1.05	-0.99	🟩
relationship to daily life	constructed	2.25	0.43	2.26	1.11	-0.01	⬜
	authentic	1.75	0.43	2.10	0.88	-0.35	⬜
	real	1.00	0.00	0.99	0.98	0.01	⬜
learners' conceptions	raise	2.50	0.50	3.29	0.51	-0.79	🟩
	check	3.00	0.71	2.15	1.03	0.85	🟩
	reflect	1.75	0.43	1.25	0.74	0.50	🟩
knowledge	factual knowledge	3.25	0.43	2.75	0.55	0.50	🟩
	procedural knowledge	3.00	0.00	2.14	1.13	0.86	🟩
	conceptual knowledge	2.75	0.43	2.88	0.64	-0.13	⬜
	metacognitive knowledge	2.00	0.00	0.58	0.76	1.42	🟩
knowledge activities	reproduction	4.00	0.00	1.88	0.91	2.12	🟧
	transfer	2.50	0.50	2.47	0.82	0.03	⬜
	creation	0.00	0.00	1.28	0.98	-1.28	🟩
forms of representation	transform	2.00	0.00	2.36	0.95	-0.36	⬜
task openness	solution procedure undefined	1.00	0.00	2.13	1.15	-1.13	🟩
	result diverse	0.00	0.00	1.39	1.11	-1.39	🟩
	editing structured	2.25	0.43	2.77	0.86	-0.52	🟩
learning supports	assistance	2.25	0.43	1.39	1.05	0.86	🟩
	exchange options	1.75	0.43	3.02	1.08	-1.27	🟩
	teacher feedback	3.50 [a]	0.83	1.40	1.16	2.10	🟧
learning paths	compensating	3.25	0.43	2.58	0.61	0.67	🟩
	profiling	1.75	0.43	1.38	1.15	0.37	⬜
	self-differentiating	1.00	0.00	2.10	1.10	-1.10	🟩

Comment:

[a] According to Burke et al. (1999), the AD_M index (average absolute deviation) shows no essential agreement.

5-level rating scale	
4	agree strongly
3	agree considerably
2	agree moderately
1	agree slightly
0	do not agree

Colour scale regarding to integer point difference	
0	⬜
1/-1	🟩
2/-2	🟧
3/-3	🟠
4/-4	🟫

Fig. 4. Scale value comparison: assessment of exercises task by the experts' proposal and rated exercises tasks (*n* = 33)

		Experts' proposed scale values		Rated tasks scale values STEM synthesis tasks - mean (*n* = 13)		Point difference of the scale values	
Scale	**Subscale**	*M*	*SD*	*M*	*SD*	Point diff.	Colour scale
chart of competencies	singular	0.50	0.00	0.87	1.33	-0.37	⬜
	additive	1.00 [a]	0.87	2.08	1.52	-1.08	🟩
	integrative	3.50	0.50	1.49	1.48	2.01	🟧
relationship to daily life	constructed	2.75	0.43	2.03	1.14	0.72	🟩
	authentic	3.00	0.00	2.56	0.67	0.44	⬜
	real	1.25	0.43	1.05	1.16	0.20	⬜
learners' conceptions	raise	4.00	0.00	3.62	0.51	0.38	⬜
	check	2.25	0.43	1.77	0.95	0.48	⬜
	reflect	1.50 [a]	1.22	1.67	0.75	-0.17	⬜
knowledge	factual knowledge	2.25	0.43	2.90	0.39	-0.65	🟩
	procedural knowledge	2.00	0.00	1.77	1.24	0.23	⬜
	conceptual knowledge	2.50 [a]	0.87	3.23	0.48	-0.73	🟩
	metacognitive knowledge	3.00	0.00	0.79	0.94	2.21	🟧
knowledge activities	reproduction	3.00	0.00	1.90	0.76	1.10	🟩
	transfer	3.00	0.00	2.44	1.19	0.56	🟩
	creation	2.00	0.00	0.96	0.80	1.04	🟩
forms of representation	transform	3.00	0.00	2.29	0.98	0.71	🟩
task openness	solution procedure undefined	3.00	0.00	2.08	1.18	0.92	🟩
	result diverse	2.25	0.43	1.63	1.12	0.62	🟩
	editing structured	3.00	0.00	2.96	0.73	0.04	⬜
learning supports	assistance	2.50 [a]	0.83	1.49	0.97	1.01	🟩
	exchange options	2.00	0.00	2.92	1.20	-0.92	🟩
	teacher feedback	3.00	0.00	1.62	1.20	1.38	🟩
learning paths	compensating	1.75	0.43	2.62	0.73	-0.87	🟩
	profiling	3.00	0.00	1.59	1.21	1.41	🟩
	self-differentiating	3.00	0.00	2.67	1.13	0.33	⬜

Comment:

[a] According to Burke et al. (1999), the AD_M index (average absolute deviation) shows no essential agreement.

5-level rating scale	
4	agree strongly
3	agree considerably
2	agree moderately
1	agree slightly
0	do not agree

Colour scale regarding to integer point difference	
0	⬜
1/-1	🟩
2/-2	🟧
3/-3	🔶
4/-4	🟫

Fig. 5. Scale value comparison of the scale values: assessment of synthesis task by the experts' proposal and rated synthesis tasks (*n* = 13)

		Experts' proposed scale values		Rated tasks scale values STEM transfer tasks - mean (n = 14)		Point difference of the scale values	
Scale	**Subscale**	*M*	*SD*	*M*	*SD*	Point diff.	Colour scale
chart of competencies	singular	0.00	0.00	0.74	1.13	-0.74	🟩
	additive	2.00	0.00	2.62	1.48	-0.62	🟩
	integrative	3.50	0.50	1.57	1.60	1.93	🟧
relationship to daily life	constructed	2.75	0.43	2.07	1.26	0.68	🟩
	authentic	3.00	0.00	2.21	1.08	0.79	🟩
	real	2.00	0.00	1.21	1.08	0.79	🟩
learners' conceptions	raise	3.00	0.00	3.33	0.45	-0.33	⬜
	check	2.00	0.00	2.14	0.79	-0.14	⬜
	reflect	1.00 [a]	0.87	1.79	0.66	-0.79	🟩
knowledge	factual knowledge	1.00	0.00	2.98	0.44	-1.98	🟧
	procedural knowledge	2.00	0.00	2.76	1.08	-0.76	🟩
	conceptual knowledge	3.00	0.00	3.26	0.53	-0.26	⬜
	metacognitive knowledge	4.00	0.00	0.79	0.92	3.21	🟥
knowledge activities	reproduction	1.75	0.43	1.93	0.90	-0.18	⬜
	transfer	4.00	0.00	2.80	1.10	1.20	🟩
	creation	2.25	0.43	1.55	0.88	0.70	🟩
forms of representation	transform	3.75	0.43	2.30	1.24	1.45	🟩
task openness	solution procedure undefined	3.00	0.00	2.45	1.30	0.55	🟩
	result diverse	3.00	0.71	1.80	1.28	1.20	🟩
	editing structured	2.00	0.00	2.59	1.02	-0.59	🟩
learning supports	assistance	2.25	0.43	1.57	1.25	0.68	🟩
	exchange options	2.25	0.43	3.19	0.74	-0.94	🟩
	teacher feedback	2.75	0.43	1.57	1.36	1.18	🟩
learning paths	compensating	1.75	0.43	2.81	0.45	-1.06	🟩
	profiling	3.00	0.00	1.62	1.25	1.38	🟩
	self-differentiating	3.75	0.43	2.38	0.85	1.37	🟩

Comment:

[a] According to Burke et al. (1999), the AD$_M$ index (average absolute deviation) shows no essential agreement.

5-level rating scale	
4	agree strongly
3	agree considerably
2	agree moderately
1	agree slightly
0	do not agree

Colour scale regarding to integer point difference	
0	⬜
1/-1	🟩
2/-2	🟧
3/-3	🟥
4/-4	🟫

Fig. 6. Scale value comparison: assessment of transfer task by the experts' proposal and rated transfer tasks (n = 14)