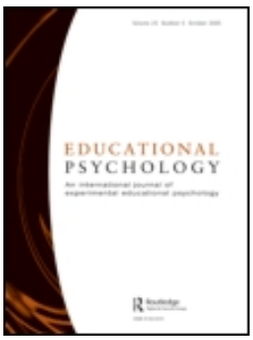
The differential effects of task complexity on domain-specific and peer assessment skills



Routledge
Taylor & Francis Group

Educational Psychology

An International Journal of Experimental Educational Psychology

ISSN: 0144-3410 (Print) 1469-5820 (Online) Journal homepage: https://www.tandfonline.com/loi/cedp20

The differential effects of task complexity on domain-specific and peer assessment skills

Marjo J. van Zundert , Dominique M. A. Sluijsmans , Karen D. Könings & Jeroen J.G. van Merriënboer

To cite this article: Marjo J. van Zundert , Dominique M. A. Sluijsmans , Karen D. Könings & Jeroen J.G. van Merriënboer (2012) The differential effects of task complexity on domain-specific and peer assessment skills, Educational Psychology, 32:1, 127-145, DOI: 10.1080/01443410.2011.626122

To link to this article: https://doi.org/10.1080/01443410.2011.626122

Educational Psychology
Vol. 32, No. 1, January 2012, 127–145

Routledge
Taylor & Francis Group

The differential effects of task complexity on domain-specific and peer assessment skills

Marjo J. van Zundert[a]*, Dominique M.A. Sluijsmans[b], Karen D. Könings[a] and Jeroen J.G. van Merriënboer[a]

[a]*Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands;* [b]*Faculty of Education, HAN University, Nijmegen, The Netherlands*

(*Received 3 January 2011; final version received 20 September 2011*)

In this study the relationship between domain-specific skills and peer assessment skills as a function of task complexity is investigated. We hypothesised that peer assessment skills were superposed on domain-specific skills and will therefore suffer more when higher cognitive load is induced by increased task complexity. In a mixed factorial design with the between-subjects factor task complexity (simple, $n = 51$; complex, $n = 59$) and within-subjects factor task type (domain-specific, peer assessment), secondary school students studied four integrated study tasks, requiring them to learn a domain-specific skill (i.e. identifying the six steps of scientific research) and to learn how to assess a fictitious peer performing the same skill. Additionally, the students performed two domain-specific test tasks and two peer assessment test tasks. The interaction effect found on test performance supports our hypothesis. Implications for the teaching and learning of peer assessment skills are discussed.

Keywords: peer assessment; task complexity; learning hierarchy; cognitive load

1. Introduction

In modern society, employees are expected to perform increasingly complex professional tasks in increasingly complex workplaces (van Merriënboer & Kirschner, 2007). Because education has to prepare students for tomorrow's work environments, it has to respond to changing societal demands. This has indeed triggered shifts in education. Whereas traditional education emphasised domain-specific knowledge and skills, modern education also addresses so-called twenty-first century skills (Scardamalia, 2001): more flexible, higher order skills, including skills for problem solving, critical thinking and lifelong learning.

Peer assessment as an instructional strategy is in line with these new developments. It is considered an important professional skill for employees. Physicians, for example, have to be able to evaluate professional behaviour of their co-workers, and researchers are expected to review papers written by their fellow researchers. Because of its alignment with educational goals, peer assessment is often applied in schools nowadays: students are involved in evaluating the quality of the school work of their fellow students.

*Corresponding author. Email: m.vanzundert@maastrichtuniversity.nl

ISSN 0144-3410 print/ISSN 1469-5820 online

http://dx.doi.org/10.1080/01443410.2011.626122
http://www.tandfonline.com

In spite of its alignment with educational developments, it is yet unknown whether peer assessment is readily applicable in educational contexts where the focus is on learning increasingly complex tasks. This paper investigates whether peer assessment performance is hindered if tasks have complex domain-specific content. It is hypothesised that when tasks become more complex, the peer assessment performance will suffer first. If this claim proves to be correct, this might have implications for the application of peer assessment in case of complex tasks.

Before providing theoretical support for this claim, a short overview of the state-of-the-art of peer assessment research will be given. There are a number of gaps in existing peer assessment research, that gave cause to performing the current study. These gaps include the context in which peer assessment research is conducted, the distinction between giving and receiving peer feedback, and the instruction of the peer assessment skill.

1.1. Gaps in existing peer assessment research

1.1.1. Research context of peer assessment

First, the growing popularity of peer assessment in education has triggered a vast body of research, especially in higher education. Although peer assessment is often applied in schools, peer assessment research in secondary education is relatively scarce (van Zundert, Sluijsmans, & Van Merriënboer, 2010). Among the few studies that have been conducted in secondary education, a specific focus on peer assessment and the use of experimental designs are scarce. Peterson and Irving (2008) as well as Brown, Irving, Peterson, and Hirschfeld (2009), studied secondary education students' conceptions about assessment in general, and also reported some specific findings about peer assessment. The former study used focus groups, and the latter an extensive questionnaire to measure the conceptions; however, in a non-experimental design. In fact, there appears to be a shortage of true and quasi-experimental studies in all types of education (van Zundert et al., 2010): the great majority of published studies on peer assessment are case studies or studies using a pre-experimental design. The current study is experimental and conducted in secondary education.

1.1.2. Assessor vs. assessee

Many different aspects of peer assessment have been covered in research so far. Researchers have focused among others on the variety in peer assessment practices (e.g. Topping, 1998), the properties of peer feedback (e.g. Gielen, Peeters, Dochy, Onghena, & Struyven, 2010), and peer tutoring or peer-assisted learning (e.g. Spörer & Brunstein, 2009; Topping, 2005; Topping & Ehly, 2001; Topping, Peter, Stephen, & Whale, 2004). A second gap, however, is that experimental studies reporting on peer assessment seldomly differentiate explicitly between giving and receiving peer feedback (van Zundert et al., 2010). Studies acknowledging this difference, often seem to choose the assessee, or the receiver of the peer feedback, as their most important perspective by examining how feedback from the assessor can influence the learning of the assessee (e.g. Strijbos, Narciss, & Dünnebier, 2010). While recognising the importance of effects of peer feedback for assessees, this paper aims to concentrate on the peer assessor, or the feedback giver.

1.1.3. Instruction of peer assessment skill

Third, little is known about the teaching and learning of skills for giving peer feedback. So far, researchers have paid little attention to the special relationship between the learning of domain-specific skills and learning how to provide peer feedback on those skills, nor have they examined how this might be connected with task complexity. Available cognitive theories of instruction, such as cognitive load theory (Sweller, 1988, 2010), cognitive theory of multimedia learning (Mayer, 1997, 2001) and four-component instructional design (4C/ID: Van Merriënboer, 1997; Van Merriënboer, Clark, & De Crook, 2002), focus mainly on domain-specific skills.

Studies founded in cognitive load theory, for example, suggest several approaches to effectively teach complex materials. One approach is to start with relatively simple tasks, and to gradually increase complexity as learning progresses (simple-to-complex sequencing or part-whole approach, e.g. Van Merriënboer & Kirschner, 2007). In line with this approach, Pollock, Chandler, and Sweller (2002) studied an instructional method for teaching students in higher education complex electrical engineering materials. Their instructional method entailed first presenting students with material containing only a few information elements, and in a later stage presenting them with all information elements required for full understanding. Another approach includes starting the instruction by presenting the learners with all information elements (i.e. full complexity), but specifically directing the learners' attention to only a few information elements (whole-part approach; for an overview of sequencing approaches, see Van Merriënboer, Kester, & Paas, 2006).

However, it remains a challenge how to integrate the teaching of complex domain-specific skills with higher order skills like peer assessment (cf. Van Merriënboer & Sluijsmans, 2009). Similarly, the literature review by van Zundert and colleagues (2010) revealed no studies investigating the particular relation between domain-specific skills and peer assessment skills as a function of task complexity. Although peer assessment is based on educational paradigms like social constructivism, rooted in cultural–historical theory (e.g. Vygotsky, 1978) and developmental psychology (e.g. Piaget, 1971), to our knowledge, a research-based theory of peer assessment does not yet exist.

1.2. Main theoretical foundations

The current study focuses on the instruction of complex domain-specific tasks in combination with peer assessment tasks. Additionally, it incorporates the main principles of learning hierarchies and cognitive load theory.

1.2.1. Learning hierarchies

Gagné (1968) stipulated the notion of learning hierarchies, and in particular the hierarchy of intellectual skills. Although this is a classical theory, it has had a profound impact on instructional theory (Smith & Ragan, 2000). Learning hierarchies refer to the prerequisite relationships between several types of cognitive skills. The acquisition of skills higher in the hierarchy is conditional on the possession of skills positioned lower in the hierarchy. For instance, the ability to recognise stimuli is a prerequisite for the ability to generate a response, and problem solving presupposes the ability to apply rules (Gagné, 1985). In a similar vein, Anderson and Krathwohl

(2001) conducted an extensive revision of Bloom's taxonomy of educational objectives (Bloom & Krathwohl, 1956). By means of this revision, Anderson and Krathwohl advocate the widespread usefulness of the taxonomy and simultaneously incorporate recent knowledge into it. In the revised taxonomy, knowledge is an underlying dimension for the dimension of cognitive processes. The dimension of cognitive processes ranges from remembering and understanding to more complex processes including applying, analysing, evaluating and, at the highest level, creating.

These ideas feed one of our assumptions: it could be argued that the position of peer assessment skills (i.e. knowledge and understanding of peer assessment) in the learning hierarchy is above that of the domain-specific skills (i.e. knowledge and understanding of the domain). Peer assessment skills can be considered higher order skills (i.e. skills higher up in the learning hierarchy), for which domain-specific skills might be a prerequisite. For example, it would be very difficult, if not impossible, for a researcher to assess a fellow researcher's research design if he or she possessed no skills related to the research domain in question. To our knowledge, the assumption of learning hierarchies has not yet been empirically proven. This makes it even more interesting to examine whether such a hierarchical relationship between domain-specific skills and higher order skills, like peer assessment skills, actually exists.

A discussion on the conditional relation between domain-specific skills and higher order skills can also be found in literature about problem-based learning (PBL, e.g. Schmidt, 1995). In PBL, small groups of students are presented with problems to be discussed, explained and studied. Much debate has been going on (e.g. Dolmans et al., 2002) about whether or not the tutor who guides the group discussions should have content expertise of the problems (i.e. domain-specific skills), in order to be able to guide the group process successfully (i.e. higher order skills). According to Barrows (1988), one of the founding fathers of PBL, tutors should have both content expertise and expertise in the guiding of the group process.

This line of thought, supported by the idea of learning hierarchies, may very well give rise to assuming that in order to be able to provide good feedback, it is conditional to have domain-specific skills. Even without addressing instructional theories, the principle of necessary domain-specific knowledge for peer assessment skills may appear to be rather obvious. One could logically assume that a student needs to know something about a particular domain before being able to peer assess that domain. In their research report about the training of peer assessment skills in peer assessment education, Sluijsmans, Brand-Gruwel, and Van Merriënboer (2002, p. 452) indeed claim that: 'Peer assessing is a complex skill that cannot be demonstrated outside a particular domain. It can be hypothesized that students who are novices in a certain domain are also less capable of assessing'. Although reasonable, to our knowledge this hypothesis has never been empirically tested. If we truly want to build on instructional theories for peer assessment, we need to know more about the teaching and learning of peer assessment skills.

1.2.2. Cognitive load theory

Next to the structural explanation of learning hierarchies, our hypothesis can be clarified by a cognitive resources explanation. People have virtually unlimited

long-term memory capacity, but only limited capacity of working memory. In other words, the resources needed to store and process new information are limited. Cognitive load theory (Sweller, 1988, 2010) assumes these capacity limitations. Task complexity is an important factor in cognitive load theory. As opposed to simple tasks, complex tasks require multiple practice sessions if they are to be taught and learned effectively. In addition, complex tasks are often ecologically valid, have more than one possible solution, and impose a high cognitive load on learners (Van Merriënboer et al., 2006). Complex tasks consist of many interacting elements that learners need to process simultaneously, which imposes high demands on working memory (Paas, Renkl, & Sweller, 2003; Paas & Van Gog, 2006). As human working memory capacity is severely limited (i.e. not many elements can be processed simultaneously), high element interactivity makes complex tasks difficult to understand. Task complexity induced by high element interactivity is referred to as intrinsic cognitive load, because the cognitive load is *intrinsic* to the nature of the task. Instructional procedures may also increase cognitive load by sub-optimal designs, known as extraneous cognitive load which does not contribute to learning. Germane cognitive load, finally, refers to the working memory capacity actually allocated to the intrinsic task aspects, hence being most important for learning (ibid.; Van Merriënboer & Sweller, 2010).

Next to task complexity, another important factor in cognitive load theory is the learner's level of expertise. The amount of cognitive load induced by the number and interactivity of elements, depends on the individual learner (e.g. Van Merriënboer & Sluijsmans, 2009; Van Merriënboer et al., 2006). Novice learners have not yet acquired mental models with chunks of related elements (i.e. cognitive Schemata), so they perceive all the elements of a task as separate units. As their expertise level increases, learners are able to link elements and form chunks. Consequently, given that each chunk is considered as one element, the number of interacting elements decreases and thereby cognitive load is reduced.

1.3. Hypotheses and aims

Combining these perspectives of learning hierarchies and cognitive load theory, the following assumptions are made: (1) peer assessment skills are superposed on domain-specific skills; and (2) complex tasks use up much cognitive resources. The combination of a complex domain-specific task and a peer assessment task induces a higher cognitive load. As human working memory capacity is limited, a cognitive overload may occur in case of a combination of these tasks. In other words, insufficient cognitive resources are available for the combination of a complex domain-specific task and a peer assessment task. As peer assessment skills are superposed on domain-specific skills, the domain-specific task will be addressed first, using up most of the available cognitive resources. Few cognitive resources will be left for the peer assessment task then. Therefore, it can be assumed that an increase in task complexity will cause the peer assessment skill to suffer first.

This assumption is depicted in Figure 1. Imagine a novice student, presented with a simple peer assessment task (see left part of Figure 1). In order to perform this task, the student first needs to call on his or her domain-specific skills, because they are supposed to be a prerequisite for performing the superposed peer assessment skills. The task being simple, the student's cognitive load will be relatively low, and he or she will have a surplus of resources for the domain-specific part of

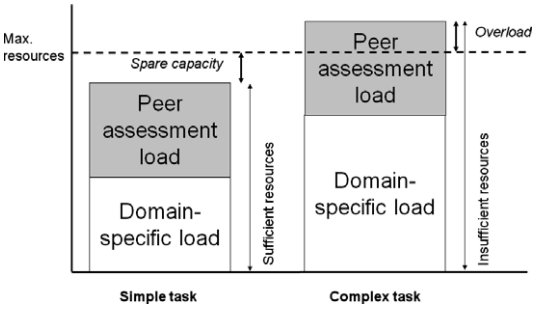


Figure 1. Hypothesised effects of task complexity on cognitive load for the combination of domain-specific skills and superposed peer assessment skills.

the task. As this leaves sufficient working memory capacity unused, there will be no problems in performing the peer assessment task. With complex tasks the situation is different, however (see the right part of Figure 1). Faced with such a task, the novice student again has to use the prerequisite domain-specific skills first. However, the task being a complex one, the cognitive load will be high and most of the student's working memory capacity will be used to process the complex domain information, leaving little or no working memory capacity for the peer assessment task. So, if task complexity is high, peer assessment suffers first and cognitive overload occurs.

If our hypothesis about the superposition of peer assessment skills would be incorrect, complex tasks will cause both domain-specific and peer assessment performance to deteriorate. In this case, if a novice student is faced with a complex task, the student will use the domain-specific skills and peer assessment skills simultaneously. As the task is complex, the cognitive load will be high and the student's working memory capacity will be used to process both the domain-specific information and the peer assessment information. In case of overload, resources are insufficient for both of these tasks, causing both the domain-specific performance and the peer assessment performance to suffer. This alternative hypothesis is depicted in Figure 2.

Taken together, this study aims to clarify the relationship between domain-specific skills and peer assessment skills as a function of task complexity. Because

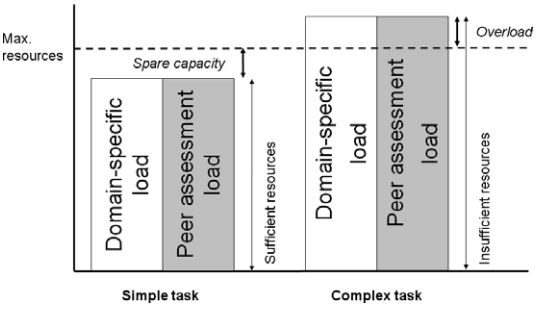


Figure 2. Possible effects of task complexity on cognitive load for the combination of domain-specific skills and peer assessment skills in case of no superposition.

peer assessment skills are hypothesised to be superposed on domain-specific skills (Figure 1), the following interaction effect will be tested: as task complexity increases, the learning of peer assessment skills will suffer more than the learning of domain-specific skills.

2. Method

2.1. Participants

A total of 110 tenth grade students from a school for secondary education in the Netherlands participated in this study (52 students attended senior general secondary education and 58 attended pre-university education). The students were randomly assigned to a simple task condition ($N = 51$) and a complex task condition ($N = 59$). All students worked on two types of tasks (i.e. a within-subjects factor task type): domain-specific tasks, covering the six steps of scientific research and peer assessment tasks. The mean age of the students was 15.43 years (SD = .60) and 47.3% were male. The students were novices in the domain of the tasks. Prior to taking the test, students filled out a questionnaire to reveal their attitudes towards peer assessment. No significant attitude differences were observed between the simple task group and the complex task group. Hence, it can be assumed that the students held comparable attitudes towards peer assessment and there was no need to correct for any differences. A more detailed description of the attitude questionnaire is provided in Section 2.2.5.

2.2. Materials

2.2.1. Integrated study tasks

For both the simple and the complex task conditions, four integrated study tasks were designed to teach the students both the domain-specific skills and the peer assessment skills required for the test tasks. The domain-specific skills that students were required to learn comprised knowledge and understanding of the six main steps of scientific research (i.e. 1 – observation, 2 – problem statement, 3 – hypothesis, 4 – experimental stage, 5 – results and 6 – conclusions). The peer assessment skills that students were required to learn encompassed evaluating a peer's knowledge and understanding of these six research steps.

The *domain-specific instruction* in each integrated study task included a description of a scientific experiment related to a topic like operant conditioning, appetite or the effects of taking vitamin C. The description comprised a random sequence of unidentified events representing the six main steps of scientific research. The instruction aimed to help students identify the six research steps in such a description of an experiment. For this purpose, students were taught by means of worked examples. In these examples, the identification of a research step was shown, along with the argumentation for this identification. The argumentation consisted of an explanation of the research step (e.g. 'This is Step 6, the conclusion. In the conclusion the researcher compares the results of the experiment with the hypothesis that was formed'.), and a link of this explanation to the description of the experiment (e.g. 'That is indeed the case here, because John finds out that taking vitamin C is beneficial: The results are compared with the hypothesis'.). Hence, the domain-specific instruction consisted of two parts: teaching students to identify a research

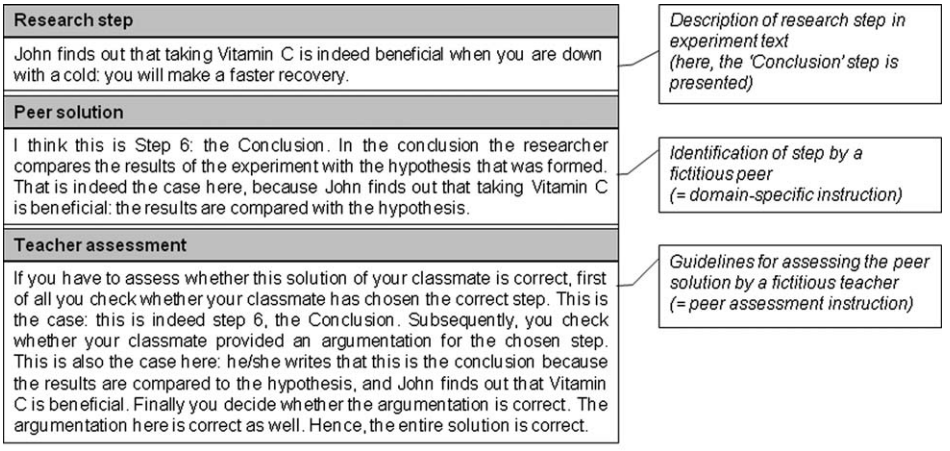
step, and teaching them how to provide a sound argumentation for the identification of each research step. An overview of the six research steps with their corresponding explanations is presented in Table 1.

The *peer assessment instruction* in each integrated study task focused on helping students to evaluate the domain-specific task solution of a fictitious peer (i.e. the peer's identification of a research step and is accompanying argumentation). For this purpose, students were taught by means of the same worked examples as used for teaching the domain-specific skills. A judgement of accuracy by a fictitious teacher for the peer's solution was shown, along with an argumentation for this judgement. The argumentation consisted of an explanation of the judgement (e.g. 'If you have to assess whether the solution of your classmate is correct, first of all you check whether your classmate has chosen the correct step. Subsequently, you check whether he or she provided an argumentation for the chosen step. Finally, you decide whether the argumentation is correct'.), and a link of this judgement to the peer's solution of the task (e.g. 'Your classmate has chosen the correct step: This is indeed step 6, the conclusion. Your classmate also provided an argumentation: He or she writes that this is the conclusion because John finds out that taking vitamin C is beneficial: The results are compared with the hypothesis. This argumentation is correct as well'.). Hence, the peer assessment instruction consisted of two parts: teaching students to judge the accuracy of a given judgement, and teaching them how to provide a sound argumentation for their judgement. A fragment of a simple integrated study task is shown in Figure 3.

The four complex study tasks were created by editing the simple tasks in accordance with three principles of cognitive load theory. First, tasks were made more complex by adding more elements (variables). For example, in the simple versions of a task on appetite, it was described that food was prepared using varying levels of sunflower oil. In the complex versions of these tasks, it was described that food was prepared with varying levels of oil as well as varying levels of salt. Second, irrelevant distracters (redundant information) were added. For example, in a task on the operant conditioning of a dog, irrelevant information about the occurrence of cardiomyopathy in dogs was added. Third, ambiguous relations between the elements (higher interactivity) were provided. For instance, in a task on gender differences and classroom behaviour, in the simple task versions it was described that both boys and girls scored similarly in IQ tests. In the complex versions, interaction effects of gender and communication patterns were presented, related to the IQ test scores. A total of 10 of these alterations were made to create a complex version of each task.

Table 1. Explanations of the research steps as presented in the integrated study tasks.

Research step	Explanation
Observation	A certain phenomenon that is considered for further research is perceived
Problem statement	The researcher experiences the observation as a problem and formulates a problem statement
Hypothesis	An attempt is made to provide a logical explanation for the problem
Experimental stage	It is tested whether the hypothesis is correct or incorrect. In order to do this, the researcher conducts an experiment and collects data
Results	The collected data are displayed synoptically, for example in graphs, diagrams or tables
Conclusion	The results are compared with the hypothesis



Note. In this integrated study task, entitled 'Vitamin C', John's classmates had colds after returning from a school trip to an amusement park. John wants to find out whether taking vitamin C will shorten the duration of their illness, and conducts an experiment. In this example, the peer's solution is correct. However, in all the study tasks both correct and incorrect peer solutions were presented.

Figure 3. Fragment of a simple integrated study task.

2.2.2. Domain-specific test tasks

Two domain-specific test tasks were developed to assess students' domain-specific skills, that is, their *own* knowledge and understanding of the six research steps. Similar to the integrated study tasks, each test task included a description of a scientific experiment. The description comprised a random sequence of the six unidentified steps of scientific research. The test tasks concerned other topics but were of similar complexity as the integrated study tasks. Students were asked to identify each research step by selecting one of the six steps from a scroll menu. They were also asked to provide a rationale for each of their choices by entering arguments supporting their answer in a text box.

For each accurately identified research step, a score of 1 was awarded. If the research step was not identified accurately, a score of 0 was assigned. As students had to identify six research steps in every test task, scores per task varied between 0 and 6. Students' performances on the domain-specific tasks were calculated by adding the scores on the two tasks. As a result, total performance scores varied between 0 and 12.

A coding scheme was developed to score the quality of the argumentation for the test tasks (see Appendix). The minimum score per argumentation was zero and the maximum score was four. Cohen's Kappa (κ) was calculated for the interrater reliability of the argumentation scores. Two researchers independently scored 50 student argumentations. The number of agreements between the two researchers was 42. The calculated interrater reliability is $\kappa = .73$. As this indicates that the interrater reliability of the coding scheme was satisfactory, one rater scored all argumentations of all participants. The scores assigned by this rater were used for the analyses.

Students provided six argumentations per test task. As there were two test tasks, and scores per argumentation varied between 0 and 4, the total quality of argumentation scores varied between 0 and 48.

2.2.3. Peer assessment test tasks

Two peer assessment tasks were developed to measure students' ability to assess *a peer's* knowledge and understanding of the research steps. The description of the scientific experiment in each peer assessment test task differed in topic but was of equal complexity as the integrated study tasks and the domain-specific test tasks. Students were provided with solutions of a fictitious peer performing the domain-specific tasks. For each peer solution to the test tasks, the students were asked to judge whether or not the step was accurately identified. The students also had to type in the rationale for their assessment by entering supporting arguments. The scoring system of performance and quality of argumentation of the peer assessment test tasks were similar to the scoring system used for the domain-specific test tasks.

2.2.4. Cognitive load measure

In order to measure students' perceived cognitive load in relation to the domain-specific tasks and the peer assessment tasks, a subjective 9-point rating scale developed by Paas (1992) was used. This rating scale measures perceived mental effort, which is considered a valid indicator of cognitive load. As cognitive load refers to the demands a task imposes on a person's cognitive resources, mental effort refers to the amount of resources the person actually invests in the task. This scale and similar subjective rating scales are often used in cognitive load research (Paas, Tuovinen, Tabbers, & Van Gerven, 2003). Students were posed the following question: 'How much effort did studying/doing this task cost you?'. They provided their answer on a scale ranging from one very, very low effort to nine very, very high effort. In the original article by Paas (1992), Cronbachs alpha for the scale was .90.

2.2.5. Attitudes towards peer assessment

To investigate potential differences in student attitudes towards peer assessment in the experimental conditions, a questionnaire consisting of 41 items was developed. Van Zundert and colleagues (2010) distinguished four outcome categories of peer assessment studies. We used these outcome categories as constructs for our questionnaire. The items of each construct were partly derived from Sluijsmans (2002). Cronbach's alpha for internal consistency was satisfactory for the four constructs: (a) measurement issues (10 items; $\alpha = .78$), e.g. 'If I assess an assignment of a classmate, my assessment will be in agreement with that given by a teacher'; (b) domain-specific skills (10 items; $\alpha = .83$), e.g. 'If I assess the assignment of a classmate, it will help me to better understand the subject matter of the assignment'; (c) peer assessment skill (11 items; $\alpha = .78$), e.g. 'I am capable of assessing a classmates assignment'; and (d) valence (10 items; $\alpha = .69$), e.g. 'I feel positive about assessing assignments of classmates'. Cronbach's alpha for the overall questionnaire was satisfactory as well ($\alpha = .77$). All the items were rated on a 4-point Likert scale, ranging from one (totally disagree) to four (totally agree). The questionnaire also contained some demographic items and two items on students previous experience with peer assessment.

2.3. Procedure

After a short explanation about the procedure, all students logged onto a computer and entered the electronic learning environment. The students in the simple task condition were requested to study four simple integrated study tasks, perform two simple domain-specific test tasks and, finally, perform two simple peer assessment tasks. The students in the complex task condition were requested to do the same, however they received complex versions of these tasks. The sequence of the tasks in the different conditions is depicted graphically in Figure 4. The students were asked to rate their perceived cognitive load on four occasions: after each of the two domain-specific tasks and after each of the two peer assessment tasks. Then, all students filled out the attitude questionnaire. The entire procedure took approximately an hour. Directly after the experiment, all the students received a 10 euro gift voucher for their participation.

3. Results

Table 2 presents the means and standard deviations for test performance, quality of argumentation and cognitive load. A significance level of .05 was used for all analyses, and partial eta-squared and Cohen's d are provided as an estimate of effect size. As for partial eta-squared, .01 corresponds to a small effect, .06 to a medium effect and .14 to a large effect (Tabachnick & Fidell, 2007). For Cohen's $d < .15$ indicates a small developmental effect; .20–.40 indicates a medium teacher effect and >.40 indicates a large and desirable effect (Hattie, 2009).

3.1. Test performance

As expected, analysis of variance revealed a significant interaction effect of task type and task complexity on test performance, $F(1, 92) = 4.17$, mean square error (MSE) = 2.13, $\eta_p^2 = .043$, $p < .05$. For the simple tasks, there was little difference in

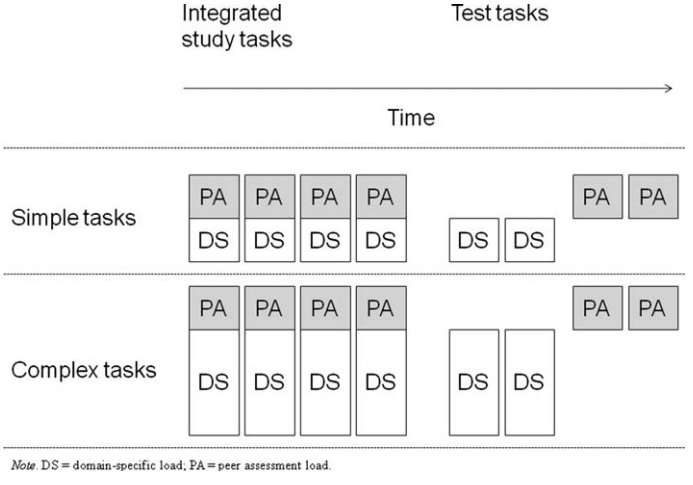


Figure 4. Sequence of presented simple and complex tasks and their assumed cognitive load.

Table 2. Means and standard deviations for test performance, quality of argumentation and cognitive load.

	Simple tasks ($n = 51$)				Complex tasks ($n = 59$)			
	Domain-specific		Peer assessment		Domain-specific		Peer assessment	
	M	SD	*M*	SD	*M*	SD	*M*	SD
Test performance[a]	9.80	2.18	9.98	1.84	9.68	2.22	8.86	1.73
Quality of argumentation[b]	28.47	4.80	22.43	6.60	26.85	5.94	20.90	5.44
Cognitive load[c]	4.43	1.35	4.79	1.20	4.47	1.48	4.79	1.37

[a]Scale = 0–12.

[b]Scale = 0–48.

[c]Scale = 0–9.

performance between domain-specific tasks ($M = 9.80$, SD = 2.18) and peer assessment tasks ($M = 9.98$, SD = 1.84), but for the complex tasks, domain-specific tasks ($M = 9.68$, SD = 2.22) suffered less from the increase in complexity than did the peer assessment tasks ($M = 8.86$, SD = 1.73; see Figure 5). The effect size of .043 indicates that this is a small to medium effect.

A paired samples *T*-test showed a significant medium-sized difference between domain-specific task performance and peer assessment task performance for the complex tasks ($t(58) = 3.07$, $p < .05$, $d = .63$), but not for the simple tasks. The effect size of $d = .63$ indicates that this is a large effect. There were no main effects of task type or task complexity on test performance.

3.2. Quality of argumentation

There was a significant main effect of task type on quality of argumentation ($F(1, 108) = 119.98$, MSE = 16.38, $\eta_p^2 = .53$, $p < .05$), with lower quality of argumentation for peer assessment tasks ($M = 21.61$, SD = 6.03) than for domain-specific tasks ($M = 27.60$, SD = 5.48). The effect size of .53 indicates that this is a large effect. There was neither a main effect of task complexity nor an interaction effect of task type and task complexity on quality of argumentation.

Differences in quality of argumentation between domain-specific tasks and peer assessment tasks can be illustrated by some representative quotes. The higher quality of argumentation in domain-specific tasks shows in the following examples. A student explained her (correct) identification of the 'Conclusion' step as follows: 'This is a conclusion. No data are shown here, so it is not Results. However, it is inferred from the data (concluded) that it [the sports drink] helps to run faster but not longer'. The student here shows knowledge about the step in question and links it appropriately to the description of the experiment. Another student provided a rationale for correctly choosing 'Problem Statement': 'Here, he [the researcher] wonders about something and formulates a question. He wonders whether the sports drink will actually cause runners to run faster and tire later'. Although making no specific reference to the description of the experiment, one Student displayed good knowledge of the 'Experimental stage': 'In this piece of text it is described how he

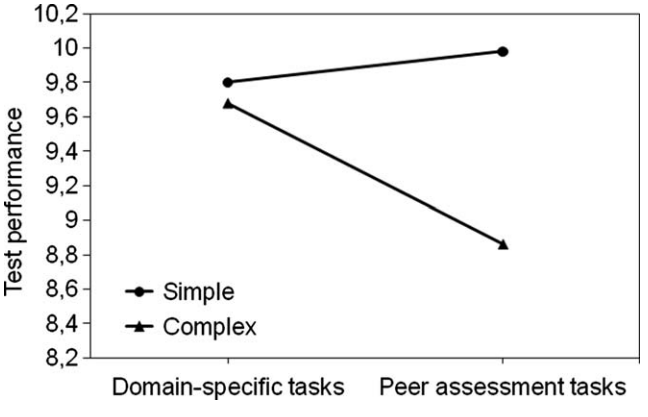


Figure 5. Interaction effect of task complexity and task type on task performance.

[the researcher] conducted his experiment. Several steps of the procedure are described, after which it is possible to report results and form a conclusion'.

Argumentation relating to the peer assessment tasks were of poorer quality. Commenting on a fictitious peer's identification of 'Experimental stage', one student wrote: 'Correct, because it is an experiment' without further explaining 'Experimental stage' or referring to the description of the experiment. Another student accounted for why he thought the fictitious peer had erroneously identified 'Results' by writing: 'No, I think it should be the conclusion'. This student too failed to provide an explanation or refer to the description of the experiment. Some students merely repeated the experiment text that was provided in the task. Explaining why 'Observation' was the correct choice for an item about a task on conditioning, a student wrote: 'The dog does not listen'.

3.3. Perceived cognitive load

There was a significant main effect of task type in relation to perceived cognitive load, $F(1, 108) = 9.90$, MSE = 0.63, $\eta_p^2 = .084$, $p < .05$. Students' perceived significantly higher cognitive load for peer assessment tasks ($M = 4.79$, SD = 1.29) than for domain-specific tasks ($M = 4.45$, SD = 1.41). This is a medium to large effect. There was neither a main effect of task complexity nor an interaction effect of task type and task complexity on cognitive load.

3.4. Student attitudes

No significant effects were found for student attitudes towards peer assessment. The groups did not differ with respect to the subscales measurement issues, domain skill, peer assessment skill and valence. Thus, the experimental groups had comparable views on peer assessment and there was no reason to assume the different conditions induced discernable attitudes. The whole group scored a mean of 2.90 (SD = .43) on measurement issues; a mean of 2.76 (SD = 2.56) on domain skill; a mean of 2.97 (SD = .41) on peer assessment skill, and a mean of 2.66 (SD = .52) on valence. All scores are close to the Likert-score of three, corresponding to 'agree' and indicating moderately positive views on all four constructs. Of all students, 71.8% indicated that they had had some previous experience with peer assessment; there was no difference in experience between the groups. The attitude questionnaires administered prior to the test taking yielded no different results than the questionnaires administered at the end of the test taking.

4. Conclusions and discussion

In this study we investigated the relationship between domain-specific skills and peer assessment skills as a function of task complexity. We hypothesised that increased task complexity would have a stronger negative effect on peer assessment skills than on subordinate, domain-specific skills. Our results on test performance clearly support this hypothesis. Furthermore, although both the domain-specific and peer assessment skills were new for the students, the observed lower quality of argumentation and higher cognitive load indicate that, overall, students' experience more difficulties with peer assessment skills than with domain-specific skills.

Our findings are of relevance for the way peer assessment has to be included in modern learning environments, which often use relatively complex tasks. Apparently, when domain-specific skills and peer assessment skills are learned simultaneously, performance may be sub-optimal when tasks are complex and learners are novices in the domain in question.

Future research is needed to provide further insights in how to design most effective instruction of peer assessment skills for complex tasks, especially when tasks require integration of first order and higher order skills. One suggestion is to provide part-whole instruction, that is, first teach the domain-specific skills separately and, subsequently, teach domain-specific skills and peer assessment skills simultaneously. This way, learners could construct cognitive schemata of the domain-specific information first. Building schemata indicates that fewer elements have to be kept in mind but with more information per element. Fewer elements imply lower cognitive load, and hence more cognitive resources to deal with superposed peer assessment skills at a later stage (cf. Pollock et al., 2002).

Other guidelines that have been shown to be effective in decreasing cognitive load in instruction in complex domain-specific skills may also be applicable to instruction in complex peer assessment skills. Simple-to-complex sequencing, for example, entails starting with simple tasks and gradually progressing towards more complex tasks. This way, the demand on working memory that would occur when complex tasks are presented right from the start, is diminished (Van Merriënboer & Sluijsmans, 2009). Another suggestion is scaffolding through fading-guidance strategies, with students first receiving ample guidance and support, which is then withdrawn gradually as students' expertise grows. For example, in the beginning a teacher might direct students' attention to the key aspects of a task, thereby reducing cognitive load, since students can disregard the less relevant task aspects (ibid.).

An unexpected finding in our study is the absence of a significant effect of task complexity on perceived cognitive load. It could have been expected that students in the complex task condition indicated a higher perceived cognitive load on the 9-point mental effort rating scale than students in the simple task conditions, but this was not found in the current study. A possible explanation for this is that the response was an artefact of the research design. Students' answers to the mental effort questions tended towards the mean. Students were assigned to either a condition with simple tasks or a condition with complex tasks. This way, they could not compare the two task complexities. If students had received both simple and complex tasks, they might have indicated greater differentiation in cognitive load on the rating scale.

Several limitations to this study should be taken into account when interpreting the results and they should be addressed in future studies. First, the cognitive load associated with reading the study tasks was not measured in our study. Students did not fill in the mental effort rating scale after each study task, so we cannot compare students' level of cognitive load on the simple and complex study tasks. Nevertheless, this might have provided valuable indications as to whether the cognitive load, induced by the combined domain-specific and peer assessment information in the integrated study tasks, is in fact acceptable with simple tasks, but too high with very complex tasks.

Another limitation concerns generalisability of the findings. Although the complex tasks were modified according to cognitive load theory principles by adding more elements, irrelevant distracters, and by providing ambiguous relations between

the elements, the nature of the task may have prevented full element interactivity. The six steps of scientific research are related to each other, but they can be considered as isolated entities to a certain extent. So, it is unknown to what extent our findings are generalisable to other settings. Moreover, the peer assessment tasks were quite basic. They consisted of two elementary components: judging whether a particular piece of information was incorrect or correct and providing arguments to support that judgement. These are the basic components of every assessment. In reality, however, peer assessment tasks are often less basic, such as when students have to write extensive assessment reports on the presentation skills of their peers. Also, no attention was paid to social aspects of peer assessment, which may play an important role when students provide face-to-face feedback on their peer's professional behaviour. Therefore, in order to determine the generalisibility of our findings, future studies should try to replicate these findings with different peer assessment tasks.

Although our sample size was deemed sufficient for the current study, it should be noted that our results apply to students in this particular educational context. An interesting point for future research may be to examine whether similar results can be found when using samples of different student populations in different settings. This will also provide more insights into the generalisability of our findings.

Additionally, a notable point for future research is the distinction between knowledge and skills for peer assessment. Theories on learning hierarchies focus on the prerequisite relationship between skills. Indeed, in this study, we reasoned that the possession of domain-specific *skills* is a prerequisite for conducting a peer assessment. It is an interesting question however, whether it is absolutely necessary to possess skills in the domain, or whether domain-specific *knowledge* alone is sufficient for conducting a peer assessment.

The current study complements existing peer assessment research in the following ways. First, it provides support for the hierarchical relationship between the learning of domain-specific skills and peer assessment skills. Until now, to our knowledge there was no empirical evidence available concerning the relation between domain-specific and higher order, peer assessment skills. Our notions are in line with expectations by Falchikov and Goldfinch (2000). In their meta-analysis of peer assessment studies they hypothesised to find effects of domain expertise (course level) on peer assessment validity. However, in the 200 peer assessment studies they analysed, they did not find convincing support for course level effects. In more advanced courses, peer assessments correlated more with staff assessments only when a study reporting extremely high correlations was included in their analyses. These high correlations resulted possibly from an artefact of the research method, making inclusion of this study in the analyses debatable. The support we found for the hierarchical relation between domain-specific and peer assessment skills, linked to cognitive load theory, may provide a foundation for developing instructional theories for peer assessment.

Second, in contrast to many earlier peer assessment studies, it used a true experimental design. This enables the drawing of inferences in terms of cause and effect, which may also be of use in developing peer assessment instructional theories. Third, the research was conducted in secondary education, an area which has been relatively neglected in peer assessment research (van Zundert et al., 2010). As peer assessment gains more popularity in this type of education, more research here may contribute to better peer assessment practices in secondary education.

To conclude, it would be good when educational practitioners and peer assessment theorists are aware that for peer assessment instruction, a sub-optimal performance on peer assessment can be expected when new, complex domain-specific skills are taught together with peer assessment. If tasks are relatively complex, as they often are in modern education, students may not have sufficient cognitive resources at their disposal to integrally conduct both the domain-specific part and the peer assessment part of a task. This inevitably leads to sub-optimal performance on peer assessment tasks. Future research should provide more insight into how this problem can be resolved as well as guidelines for the teaching of complex peer assessment skills.

Acknowledgements

This research project is funded by the Netherlands Organisation for Scientific Research (NWO) under project number 411-05-110.

References

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

Barrows, H.S. (1988). *The tutorial process*. Springfield, Illinois: Southern Illinois University School of Medicine.

Bloom, B.S., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals, by a committee of college and university examiners. Handbook 1: Cognitive domain*. New York, NY: Longman.

Brown, G.T.L., Irving, S.E., Peterson, E.R., & Hirschfeld, G.H.F. (2009). Use of interactive–informal assessment practices: New Zealand secondary students' conceptions of assessment. *Learning and Instruction, 19*, 97–111.

Dolmans, D.H.J.M., Gijselaars, W.H., Moust, J.H.C., De Grave, W.S., Wolfhagen, I.H.A.P., & Van der Vleuten, C.P.M. (2002). Trends in research on the tutor in problem-based learning: Conclusions and implications for educational practice and research. *Medical Teacher, 24*, 173–180.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 3*, 287–322.

Gagné, R.M. (1968). Learning hierarchies. *Educational Psychologist, 6*, 1–9.

Gagné, R.M. (1985). *The conditions of learning* (4th ed.). New York, NY: Holt, Rinehart & Winston.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304–315.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Mayer, R.E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist, 32*, 1–19.

Mayer, R.E. (2001). *Multimedia learning*. New York, NY: Cambridge University Press.

Paas, F.G.W.C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*, 429–434.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*, 1–4.

Paas, F., Tuovinen, J.E., Tabbers, H., & Van Gerven, P.W.M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*, 63–71.

Paas, F., & Van Gog, T. (2006). Optimising worked example instruction: Different ways to increase germane cognitive load. *Learning and Instruction, 16*, 87–91.

Peterson, E.R., & Irving, S.E. (2008). Secondary school students' conceptions of assessment and feedback. *Learning and Instruction, 18*, 238–250.
Piaget, J. (1971). *Science of education and the psychology of the child*. London: Longman.
Pollock, E., Chandler, P., & Sweller, J. (2002). Assimilating complex information. *Learning and Instruction, 12*, 61–86.
Scardamalia, M. (2001). Big change questions: Will educational institutions, within their present structures, be able to adapt sufficiently to meet the needs of the information age? *Journal of Educational Change, 2*, 171–176.
Schmidt, H.G. (1995). Problem-based learning: An introduction. *Instructional Science, 22*, 247–250.
Sluijsmans, D.M.A. (2002). *Student involvement in assessment: The training of peer assessment skills* (Unpublished doctoral dissertation). Open University of The Netherlands, Heerlen.
Sluijsmans, D.M.A., Brand-Gruwel, S., & Van Merriënboer, J.J.G. (2002). Peer assessment training in teacher education: Effects on performance and perceptions. *Assessment & Evaluation in Higher Education, 27*, 443–454.
Smith, P.L., & Ragan, T.J. (2000). The impact of R.M. Gagnés work on instructional theory. In R.C. Richey (Ed.), *The legacy of Robert M. Gagné* (pp. 147–181). Syracuse, NY: ERIC Clearinghouse on Information & Technology.
Spörer, N., & Brunstein, J.C. (2009). Fostering the reading comprehension of secondary school students through peer-assisted learning: Effects on strategy knowledge, strategy use, and task performance. *Contemporary Educational Psychology, 34*, 289–297.
Strijbos, J.W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction, 20*, 291–303.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257–285.
Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*, 123–138.
Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson Education.
Topping, K.J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*, 249–276.
Topping, K.J., & Ehly, S.W. (2001). Peer assisted learning: A framework for consultation. *Journal of Educational and Psychological Consultation, 12*, 113–132.
Topping, K.J., Peter, C., Stephen, P., & Whale, M. (2004). Cross-age peer tutoring of science in the primary school: Influence on scientific language and thinking. *Educational Psychology, 24*, 57–75.
Topping, K.J. (2005). Trends in peer learning. *Educational Psychology, 25*, 631–645.
Van Merriënboer, J.J.G. (1997). *Training complex cognitive skills*. Englewood Cliffs, NJ: Educational Technology.
Van Merriënboer, J.J.G., Clark, R.E., & De Crook, M.B.M. (2002). Blueprints for complex learning: The 4C/ID-model. *Educational Technology, Research and Development, 50*, 39–64.
Van Merriënboer, J.J.G., Kester, L., & Paas, F. (2006). Teaching complex rather than simple tasks: Balancing intrinsic and germane load to enhance transfer of learning. *Applied Cognitive Psychology, 20*, 343–352.
Van Merriënboer, J.J.G., & Kirschner, P.A. (2007). *Ten steps to complex learning: A systematic approach to four-component instructional design*. Mahwah, NJ: Lawrence Erlbaum.
Van Merriënboer, J.J.G., & Sluijsmans, D.M.A. (2009). Toward a synthesis of cognitive load theory, four-component instructional design, and self-directed learning. *Educational Psychology Review, 21*, 55–66.
Van Merriënboer, J.J.G., & Sweller, J. (2010). Cognitive load theory in health professions education: Design principles and strategies. *Medical Education, 44*, 85–93.
Van Zundert, M.J., Sluijsmans, D.M.A., & Van Merriënboer, J.J.G. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction, 20*, 270–279.
Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes*. London: Harvard University Press.

Appendix

Coding scheme: quality of argumentation for content-related and peer assessment tasks

Code	Score	Explication	Example
A: Not meaningful or wrong argumentation	0	The argumentation the student provides is incorrect or not meaningful	I guessed this one
B: Repetition of research step	1	The argumentation the student provides is merely a reiteration of the selected step	This is the problem Statement
C: Reference to text	2	The argumentation refers to the description of the research provided in the task	They want to know if it is true whether the teammates run faster and gain more energy after drinking the sports energy drink
D: Description of research step	3	The student provides a definition of the chosen step	He tries to find out whether his hypothesis is correct by running a test
E: Combining B and/or C and/ or D	4	The student shows insight by explaining the step and linking the explanation to the research description in the task	He thinks that the consequences of the sports energy drink are that everybody gets more energy. Hence, he makes a certain assumption of which he is not sure whether it is correct or not